Beyond Predictive Algorithms in Child Welfare

Erina Seh-Young Moon* University of Toronto
Devansh Saxena† Carnegie Mellon University
Tegan Maharaj‡ University of Toronto
Shion Guha§ University of Toronto

ABSTRACT

Caseworkers in the child welfare (CW) sector use predictive decision-making algorithms built on risk assessment (RA) data to guide and support CW decisions. Researchers have highlighted that RAs can contain biased signals which flatten CW case complexities and that the algorithms may benefit from incorporating contextually rich case narratives, i.e. - the **casenotes** written by caseworkers. To investigate this hypothesized improvement, we quantitatively deconstructed two commonly used RAs from a United States CW agency. We trained classifier models to compare the predictive validity of RAs with and without casenote narratives and applied computational text analysis on casenotes to highlight topics uncovered in the casenotes. Our study finds that common risk metrics used to assess families and build CWS predictive risk models (PRMs) are unable to predict discharge outcomes for children who are not reunified with their birth parent(s). We also find that although casenotes cannot predict discharge outcomes, they contain contextual case signals. Given the lack of predictive validity of RA scores and casenotes, we propose moving beyond quantitative risk assessments for public sector algorithms and towards using contextual sources of information such as narratives to study public sociotechnical systems.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Empirical studies in HCI; Human-centered computing—Applied computing—Computing in government

1 INTRODUCTION

Predictive risk models (PRMs) have increasingly infiltrated our everyday lives, including within the criminal justice system [4, 32], homelessness [74], employment services [71, 112], the child welfare system [92, 109], and education sector [72], to study and predict the future given the presently available information and generate decisions that carry significant ramifications for individuals. While initially developed with the expectation that such systems can reduce costs and generate objective decision-making [19], recent research has become increasingly critical of their objectivity and impact.

The US child welfare (CW) sector is one such sector that has extensively adopted PRMs to predict child maltreatment risk. Given the high stakes for the lives of children and families, such algorithmic systems have been introduced to alleviate caseworkers' high workloads, minimize costs, and arrive at objective, evidence-based decisions with a particular focus on empirical risk measurement and prediction [15, 91, 92, 97, 100], i.e., proactively recognizing high-risk cases. However, recent research that examines stakeholder perceptions towards these systems has found that they use potentially biased quantitative inputs related to factors like poverty and unjustly surveil and stigmatize low-income and minority communities when they enter the CW system [44, 50, 86, 104]. There is also increasing awareness in the general scientific community on the relevance of negative results [47, 94, 114] to evaluate and audit the impact of algorithmic systems through multiple methodological lenses.

*e-mail: erina.moon@mail.utoronto.ca
†e-mail: devanshsaxena@cmu.edu
‡e-mail: tegan.maharaj@utoronto.ca
§e-mail: shion.guha@utoronto.ca

Model (best-performing in overall accuracy)	True Negative Rate
Random Forest using RA scores (AAPI scores)	0.167
Random Forest using Topic Model Probabilities	0.357
Constant Predictor	0.5
Random Predictor	0.2

Table 1: Classification model metrics for the top performing models: We show how often the top performing models could identify when children were not reunified with their bio-parent(s) given a balanced dataset. The Constant Predictor predicts reunified ('R') and not reunified ('NR') outcomes 50% of the time each. The random predictor predicts 'R' outcomes 80% of the time and 'NR' 20% of the time.

In recent years, scholars have increasingly explored how CW narrative data can be used to infer contextual insights around the CW system and be incorporated into PRMs. Saxena et al. applied topic modeling on caseworker casenotes to illuminate invisible patterns of frontline caseworker labor [95] and uncovered that while risk in CW is primarily measured as a function of *risk introduced by clients*, this does not fully account for a clients' holistic circumstances and the dynamic multiplicity/temporality of risk factors underpinning CW cases [94]. Field et al. [46] also found training computational models using text data **with** quantitative, structural administrative data (e.g., demographic information, past public welfare receipt history, criminal history) may mitigate racial disparities in out-of-home placements predictions. With these works in mind, we critically examined the predictive validity of the *conceptualization of risk* in predictive risk models prevalent in the child welfare sector (CWS). We quantitatively compared signals from RA data and CW narratives to ask **RQ1:** How does risk measured in CW risk assessments affect case outcomes? **RQ2:** How do case narratives inform child welfare caseworker decisions and case outcomes? and **RQ3:** How can researchers better support child welfare caseworkers?

We partnered with a child welfare agency in a US Midwestern state and examined casenotes written by caseworkers and CW risk assessment (RA) scores assessed for 277 families discharged from the agency between November 2018, to March 2022. We trained 14 classifier models using support vector machines (SVM) and random forest models, incorporating narrative data from CW caseworker casenotes using topic modeling and RA scores from the Adult Adolescent Parenting Inventor-2.5 (AAPI) and North Carolina Family Assessment Score (NCFAS) (two commonly used RAs which aims to comprehensively measure case risk factors for families [35, 67]). We compared the predictive validity of our trained classifier models to examine how well casenote data and RA scores can predict reunification (discharge) outcomes. We also qualitatively examined topics in our topic model solution to better understand the types of case signals our topic model solution provided to the trained classification models. In summary, our study makes the following contributions.

- We quantitatively deconstruct AAPI and NCFAS scores, which measure child maltreatment risk across 100 metrics, to examine their ability to correctly predict discharge outcomes. We find quantitative support that RAs are largely irrelevant in informing discharge outcomes. We find that while the best performing classifier built using *only* RA scores shows a high

accuracy rate of 86.4%, the model can correctly identify cases that end in non-reunification ('NR') between child and bio-parent(s) only 16.7% of the time (seen in Table 1). This is worse than a constant predictor or random predictor.

- Through computational text analysis, we show that case narratives provide different signals around cases compared to the RA scores. Case narratives can highlight contextual details between individuals involved in a case that can affect the trajectory and outcome of cases. We find that a random forest model built using *only* topic model solution data can correctly identify cases that end in 'NR' 35.7% of the time (see Table 1). This is better than a random predictor but worse than a constant predictor.
- We show how casenotes inherently contain limitations in prediction tasks in CW as they present the writer's perspectives. We also show narratives in aggregation can provide the contextual rationale behind CW decisions.
- We computationally critique CW algorithms that use RA scores to predict and affect case outcomes. We find support to shift away from predictive algorithms in CW and towards using contextual information such as narratives to study public sociotechnical systems.

2 Ethical Considerations

We had Institutional Review Board (IRB) approval from our university to use CW casenotes and discharge records for this research. The data we used for this research was bound by the IRB and a research agreement with the CW agency, which only permits us to use the data for research purposes. We are also aware that beyond the IRB guidelines and data-sharing research agreement, the nature of our research is highly sensitive, and the process in which CW data is collected is a morally complicated matter related to privacy, oppression, and surveillance. Considering the pervasiveness of predictive risk models (PRMs) in CW and their biased impact on families, we hope our study can illuminate the limitations of using RA scores to build and guide CW decisions. In this study, to minimize the risk of unintended harm, we considered the principles of beneficence, respect for law, and public interest when determining how and to what extent we should utilize the information from our datasets for our analysis. We anonymized names and numbers from the data to ensure private information does not leak into our predictive models and do not make any raw data public.

3 Related Work

3.1 Algorithmic "Risk" Assessment Tools in the Public Sector

Public sector agencies in several countries are experiencing increasing pressures from policymakers to use cross-sector public administrative data to improve decision-making in managing and delivering public services by employing algorithmic systems [2, 56, 91, 110]. Risk assessment (RA) algorithms are now being used to make high-stakes decisions about human lives by several public agencies such as the child welfare system, criminal justice system, public education, housing authority, and public benefits programs to decide which neighborhoods require more surveillance, predict the risk of recidivism [111], tackle school segregation and assign students to school zones [87, 88], determine student performance [113], and to assess the risk of long-term unemployment for a person [2, 56].

As Redden et al. [83] note, one of the critical drivers in this push for algorithmic systems for governance has been the preemptive recognition and mitigation of "risk", i.e. – identifying clients in the riskiest circumstances and extending services to them. However, many of these algorithms have inadvertently exacerbated racial biases where minority communities are being over-surveilled [111], schools are becoming more segregated [106], and minority families are being over-investigated for allegations of child maltreatment [54]. On the other hand, due to an overemphasis on clients "most in need", resources from public benefits programs are being directed away from citizens who only require temporary assistance to regain stability in their lives, i.e. – clients who require significantly fewer government resources but are labeled "low risk" and are consequently, screened out of consideration [44]. As highlighted by prior work, four critical factors make the task of risk prediction especially hard in the public sector – 1) outcomes in the public sector such as the risk of recidivism and risk of child maltreatment are poorly and inconsistently defined [9, 44, 53, 92], and 2) it is not possible to directly observe and assess an outcome such as "risk of future child maltreatment", therefore, algorithms are trained to predict proxy outcomes such as the likelihood of re-referral or placement in foster care within the next two years [25, 60]. Furthermore, 3) a person's life can stabilize and destabilize several times which makes it hard to assess what may constitute a successful outcome or intervention [56, 91], and 4) administrative data in the public sector is poorly and inconsistently collected which leads to several algorithmic biases in outcomes that seek to predict client's behavior and may lead to unfairness in decision-making processes [34, 51, 94]. Therefore, there is a need to critically examine how predictive risk models are formulated in the public sector and whether they are achieving better outcomes for citizens.

3.2 Predictive Risk Models (PRMs) in Child Welfare

Predictive risk models (PRMs) are deeply embedded in various stages of the CW system [91]. Initially introduced to mitigate concerns around subjective decision-making by individuals and to arrive at data-driven, evidence-based, consistent decision-making for CW cases [92, 97, 100], PRMs help determine which families should be screened in for further investigation when an allegation of child maltreatment is made, deciding whether to remove a child from their biological parents and predicting future maltreatment risk [44,51,91]. For a more in-depth analysis of child welfare algorithms, we direct the readers to Saxena et al. [92] which provides a comprehensive review of predictive risk models in CWS.

Recent scholarship studying CW algorithms has largely used a participatory approach to evaluate the downstream effects of implementing these systems on different racial or income groups, envision solutions to mitigate algorithmic harms [25, 52, 104], and qualitatively understand stakeholder perceptions and usages of these computational systems [57, 61, 92]. Notably, Stapleton et al. [104], and Brown et al. [18] highlighted the harms caused by these models finding that CW stakeholders (e.g., caseworkers, families involved with CWS) have concerns about how the models focus on case deficits (i.e., risk) rather than considering case strengths which can inadvertently perpetuate stigmatization and racial biases. Moreover, researchers have found that subjectivity and caseworker discretion are absorbed into PRMs despite their touted objectivity. Inputs into these models often involve factors that require inherent interpretation such as parent cooperativeness, parenting attitudes, or home visits [15, 91] and systematic factors where due to high staff turnover, inexperienced caseworkers measure risk based on impressions [15, 36, 96]. In field studies, Cheng et al. [25], Kawakami et al. [61], and Saxena et al. [91] found that caseworkers actively exercise discretion to mitigate racially biased outcomes from PRMs, gamed the models, or tried to guess how features affected risk scores to better serve their clients. Saxena et al. also [94] further problematized empirical CW risk prediction. The authors found that PRMs often treat risk as a static construct when in fact, risk in CW cases is dynamic and caused by multiple factors, including the child welfare system itself which faces a shortage of good foster homes and experienced caseworkers who can better support clients. In sum, several qualitative [61, 91], quantitative [25, 96], and mixed-methods

studies [94] conducted on PRMs in child welfare have questioned the validity of empirical risk prediction itself.

3.3 Assessing the validity of Child Welfare PRMs

Coston et al. [38] and Raji et al. [82] argue it is important to examine the validity of decision-making algorithms to evaluate their functionality. In a related vein, Saxena et al. [94,96] recently raised construct validity concerns around PRMs in child welfare, finding marked divergences between how risk is quantified as a static construct in risk ratings and how they are depicted as a temporally dynamic construct in CW casenotes. These studies raised the need to better understand and account for contextual systematic and procedural factors to measure 'risk' in CW. And yet, interestingly, although PRMs mostly focus on predicting child maltreatment 'risk', to the best of our knowledge, few studies quantitatively examine their validity to question how 'risk' in these models is conceptualized, measured, and utilized to guide decision-making. To date, Vaithianathan et al. [108] have examined the predictive validity of a PRM by measuring the association between risk scores and future child hospitalization/emergency department visits in the follow-up period. However, because their study focused on probabilistic associations between risk scores and future maltreatment events, their findings could be subject to confounding factors as CWS would likely intervene once a family is deemed high risk.

In sum, we find that while many child welfare PRMs focus on measuring and predicting child maltreatment risk, there is a gap in the literature around our understanding of how risk is conceptualized in these models from a validity perspective. In this study, we specifically examined the predictive validity of risk assessments (RAs) used to build PRMs by examining how RAs and casenotes inform discharge outcomes. In CW, discharge outcomes signal whether the courts and caseworkers determine a child is safe to be reunified with their parents. Using machine learning techniques, our study asked how risk measured in RAs affects CW discharge outcomes and how casenotes inform CW decisions and outcomes. Through our research questions, we seek to better understand how researchers can better support CW caseworkers.

4 Research Context

We partnered with a CW agency in a Midwestern US state. The agency is contracted by the state's Department of Children and Families (DCF) to support foster homes and offers case management services in compliance with all DCF directives. When allegations of child maltreatment are reported and hereafter substantiated by a DCF worker, the case is referred to the CW agency. Following DCF standards, caseworkers conduct initial assessments to determine the family's needs and provide recommendations to the court. Then, after discussions with birth parent representatives, the judge, and the district attorney's office, CW teams at the agency work together to offer services to families. Within the CW agency, various teams provide different services to families, including teams that focus on offering in-home services, preventing sex trafficking, and helping foster youth transition out of foster care.

For this study, we specifically obtained casenotes and a discharge dataset with RA scores from the Family Preservation Services (FPS) team at the CW agency. When a case is assigned to the FPS team, FPS caseworkers assist bio-parents achieve reunification with their children by providing crisis support, parenting classes, and helping improve family functioning [26]. Following FPS intervention, FPS caseworkers provide evidence to the DA's office and judge to recommend a discharge outcome for the family, such as reunification with their bio-parent, placement into foster care, kinship care, or adoption. Discharge outcomes on families meaningfully signal whether caseworkers and the courts deem a child safe to be reunited with their bio-parent(s) or if they should be placed into other placement forms such as foster care or adoption [29].

An integral part of FPS caseworker duties include recording casenotes following stringent documentation standards guided by social work theory [22]. These casenotes serve multiple critical purposes, including, providing evidence for the DA's office to recommend reunification, ensuring accountability to caseworkers and the agency, and helping guide the child welfare process by providing a summary of services offered to families [22,102]. DCF also requires FPS caseworkers to complete quantitative RAs to assess family risk and safety per the Child Abuse Prevention and Treatment Act [29]. In the CW agency's Midwestern state, caseworkers are required to conduct RAs at intake and discharge (when families enter and before leaving CW services) to consistently collect information on factors that threaten child safety. Based on the RA scores and casenotes recorded, caseworkers then use casenotes and RAs as evidence to justify case discharge recommendations.

5 Methods

5.1 Dataset

For our study, we combined two different datasets provided by the Family Preservation Services (FPS) team at our partner child welfare agency, the discharge dataset and the casenote dataset.

Casenote dataset The casenotes dataset comprised 12,576 casenote records for 471 families that received services from the FPS team from May 1, 2019, to March 31, 2022. Families were not necessarily involved with the agency for this entire period but engaged with the agency within this timeframe. The dataset included all details pertaining to any caseworker interactions and observations of persons involved in a case (e.g., biological parents, children, foster parents) and internal communications between caseworkers. Interactions could be in the form of phone calls, email messages, and in-person or virtual meetings. Each casenote record included a family ID and narrative casenotes.

Discharge dataset The discharge dataset comprised 371 discharge records for 339 families discharged from the FPS team from November 2, 2018, to March 24, 2022. Each discharge record included a Family ID, date of admission and discharge from the program, and RA scores for the AAPI [11], and NCFAS [64]. These two assessments are two commonly used RAs in CW that fulfill state DCF requirements and are similar to other CW assessments that categorize risk into various levels[1]. The AAPI asks parents to rate their parenting attitudes related to child maltreatment from 1 to 10 where scores can be binned into low (8-10), moderate (4-7), and high risk (1-3) [43]. In contrast, the NCFAS is completed by caseworkers and asks caseworkers to rate multiple domains of family functioning related to a family's environment, parental capabilities, family interactions, family safety, child well-being, and readiness for reunification [63, 64]. NCFAS scores range from -3 to +2 and can also include not applicable ('NA') or unknown ('UK') [77]. Negative scores indicate challenges, 0 signifies the baseline/adequate level, meaning no interventions are needed, and positive scores mean strengths [64]. By design, the AAPI and NCFAS scores are taken twice as a structured means for caseworkers to compare intake and discharge scores and assess how safety and risk factors in cases have been addressed following CW intervention. However, it is not always possible to collect all RA data. Our dataset showed that some intake and discharge RA scores were missing because families refused to take the assessments or caseworkers could not contact the bio-parents as they were hospitalized, incarcerated, missing, or had moved to a different state. Other times, the caseworkers determined the case was a low-needs case and an assessment was unnecessary or if the family had recently been re-referred to the agency and their scores were already on file.

[1]AAPI (Adult Adolescent Parenting Inventory) has been administrated in over 2 million administrations [45], and NCFAS has been used in over 1,000 US agencies and in 20 countries [73].

Dataset Name	N	Features Included	Target Variable
TM	277	19-topic model solution probabilities (19TM)	Discharge outcome
AAPI	145	Intake & discharge AAPI scores	Discharge outcome
NCFAS	221	Intake & discharge NCFAS scores	Discharge outcome
AAPI + NCFAS	133	Intake & discharge AAPI and NCFAS scores	Discharge outcome
AAPI + TM	145	Intake & discharge AAPI scores and 19TM	Discharge outcome
NCFAS + TM	221	Intake & discharge NCFAS scores and 19TM	Discharge outcome
AAPI + NCFAS + TM	133	Intake & discharge AAPI and NCFAS scores and 19TM	Discharge outcome

Table 2: Descriptions of the six datasets (N here indicates the number of observations in the dataset)

5.2 Preprocessing the datasets

To track each family's entire narrative casenote history, we collated all casenote records in the *Casenote dataset* (N=12,576) by each family ID ordered chronologically (oldest to latest). We cleaned and anonymized the casenotes by removing stopwords, punctuation, numbers, and names. After preprocessing the casenotes, we noted that many of the casenotes were lengthy, with collated casenotes averaging 3,299 words, with 53,303 unique words across the casenotes, and there being 38,748 words in the longest casenote.

In the *Discharge dataset*, each record detailed the discharge outcome of a case which could be either adoption, guardianship/independent living, reunification, and concurrent permanency plan[2]. As we wanted to understand how RAs can meaningfully inform whether it is deemed safe for a child to be reunified with their bio-parent, we grouped the discharge outcomes into two: reunified ('R') and not reunified ('NR') [3]. After dichotomizing the discharge outcomes, we noticed more than 80% of cases ended in 'R.'

Combined dataset Finally, we combined the two preprocessed datasets by using Family ID as the unique identifier so that we could have one set of casenotes and one discharge outcome for each family ID. Due to how data was recorded at the agency and because some families were continuing to receive CW services, some family IDs that had an associated casenote did not have a discharge outcome and vice versa. In this case, we excluded those family IDs, so that we ended up with a collection of 277 discharge records with associated 277 collated casenotes for 277 families.

5.3 Topic modeling on narrative casenotes

We adapted the steps taken by Saxena et al. [95] and applied Mallet's implementation of LDA topic modeling [14] on our collated and preprocessed casenotes for 277 families (from section 5.2). LDA can generate topic distributions over words for topics, produce distributions of topic probabilities for each text document, and scale to dense texts [5, 13, 76, 81]. Using LDA, we ran a 19-topic model solution, where we computed 19 topic probabilities for each of the 277 collated casenotes. Prior work by Baumer et al. [10] has argued that conducting grounded thematic methods can help illuminate thematic patterns in text from topic model solutions. Accordingly, to contextually understand topics generated from our topic model solution, the first author of this paper then used an open-coding process to manually inspect casenotes with the highest probabilities associated with each of the 19 topics and keywords associated with the topics to understand and label each topic [16]. After labeling each topic, we grouped topics together if they carried similar thematic content to better gauge the high-level themes that emerge from the casenotes (as seen in topic groupings by theme in Fig 1).

5.4 Dataset Creation

To compare and contrast how different combinations of RA (risk assessment) scores and narratives can inform discharge outcomes, we created six different datasets from the combined dataset outlined in section 5.2. Each dataset contained a combination of intake and discharge RA scores or topic model solution probabilities as features and the discharge outcome for the respective family ID. Table 2 shows the names and data included in the datasets.

5.5 Classification Models

We trained classification models using the six different datasets presented in Table 2 to understand how RA metrics can inform CW discharge decisions. We one-hot encoded all intake and discharge AAPI and NCFAS scores in the datasets and then created train and test sets for our datasets using a 7:3 ratio. Training and evaluating risk assessment classification models often carry challenges due to class imbalances common in risk assessment contexts. Class imbalances can cause models to fail to identify patterns from the minority class and instead favor the majority class [48, 59, 69]. Because our datasets were relatively small and imbalanced (more than 80% of cases in the datasets ended in 'R'), we thus resampled and balanced the training datasets to include 1000 rows of training data where 500 observations ended in 'R' and 500 observations ended in 'NR' using synthetic minority oversampling techniques (SMOTE) [69]. SMOTE is useful because its variants can be applied to both continuous and categorical data [69], and for each minority class observation, the technique can quickly generate new examples by randomly finding closest neighbors in the feature space [24]. Next, we reduced the dimensionality of our datasets by applying principal component analysis (PCA) to denoise the datasets [79]. Finally, using the resampled and dimension-reduced datasets, we trained and compared the performance metrics of random forest models setting a threshold at 0.5 [17] and support vector machines (SVM) using a radial basis function kernel [37] for each of the datasets. We trained classifiers using random forests and SVMs because they can work with high-dimensional data efficiently and on problems that are non-linear [20, 68, 80]. While studies often adjust decision thresholds and validate risk assessment classifiers by inspecting the AUC of the classifier across different thresholds, Lobo et al. [70] discussed by Gerchick et al. [50], and Kwegyir-Aggrey [66] find (1) model comparisons with AUCs obfuscate different error types from different thresholds, (2) determining thresholds based on AUC ignores the inherently normative task of assigning costs to incorrect classifications, and (3) AUCs provide model performance summaries over the receiver operating characteristic (ROC) curve that may not be of relevance. In light of these findings, we evaluated the performance of the classifiers by inspecting the accuracy, false positive rate, false negative rate, and specificity rate of each of the models.

[2] This is when caseworkers are working towards reunifying children with birth parents while simultaneously identifying and working towards other placement options for the child such as guardianship with a relative or adoption in case placement with birth parents is unsuccessful [28]

[3] We categorized adoption, concurrent permanency plan, and guardianship/independent living outcomes as 'NR' and reunification outcomes as 'R'. We categorized concurrent permanency plan as 'NR' as this outcome signaled caseworkers anticipated a possibility that children would not be reunified with their biological parent(s).

6 RESULTS

In this section, we present our findings from the trained classification models and organized by our research questions.

6.1 Risk assessments do not inform discharge outcomes (RQ1)

As seen in the orange rows in Table 3, we found that while classifiers built using only RA scores had high accuracy rates, the models were predominantly predicting children would be reunified with their bio-parent(s) and could not distinguish between cases where children were not reunified with their bio-parent(s). A false positive rate (FPR) in the predictive models indicates how often the models are predicting that a case would end in 'R' (reunification) when actually, the child was not reunified ('NR') with their bio-parent(s) in real life[4]. As seen in Table 3, SVM models trained using the AAPI or NCFAS or both AAPI+NCFAS dataset all had a FPR equalling 1. This meant that all cases in these models were miscategorized where all 'NR' cases (actual outcome) were predicted as 'R' cases. The random forest models trained using different combinations of the RA scores, similarly showed a high FPR, suggesting that RA data was not helping predict reunification outcomes regardless of the machine learning algorithm used. A false negative rate (FNR) in the predictive models indicates how often the models are predicting that children will not be reunified with their bio-parent(s) when in fact, they were reunified [5]. We noted that models built on RA data made almost no errors or any errors when predicting 'R' cases (FNR equaled 0 or nearly 0), meaning that the models were unable to distinguish cases that ended in 'R' and 'NR' through the RA scores. **Collectively, we noted classifier models that were trained using either AAPI (parent perspective) or NCFAS scores (caseworker perspectives), or both AAPI and NCFAS scores, did not improve the models' ability to correctly predict discharge outcomes. These results suggested RA scores were not identifying risk factors that determine case outcomes and were providing misleading signals regarding case risk factors.**

6.2 Case narratives can provide some signals towards discharge outcomes (RQ2)

We also trained classifier models using the probabilities from our trained topic model solution as features to examine whether narrative data can predict discharge outcomes. The blue rows in Table 3 show the model metrics that use only topic model data. A specificity rate in the models indicates how often the models are correctly predicting children are not reunified with their bio-parent(s) given the fact they were actually not reunified[6]. Table 3 shows that classifier models that used topic model data had a lower accuracy rate, lower FPR, and higher FNR compared to models built solely on risk assessment data. Compared to models built using RA data, the lower FPR of 64.3% (random forest model) and 71.4% (SVM model) indicated that the models were miscategorizing proportionally fewer actual 'NR' cases as 'R', and the comparatively higher specificity rates showed that the models were doing a better job at predicting 'NR' cases. However, we also noted that models built using TM data had a FNR greater than 0, indicating the models were making several 'NR' predictions for actual 'R' cases. **These results suggested that topic model solution probabilities may be providing some indicators that can predict discharge outcomes (particularly in identifying and predicting cases that do not end in reunification) compared to using only risk assessment scores.** Our results resonate with Saxena et al. [96] who found differences between RA ratings and risks noted in caseworker casenotes, noting caseworkers were recording different facets of the same case in casenotes compared to RAs.

As shown in the gray rows in Table 3, we found that predictive models built using topic model data and RA scores exhibited similar performance metrics to models built using only RA scores. **Adding risk assessment data to topic model data did not improve the classifiers' ability to correctly predict 'NR' outcomes, further suggesting that while our topic model solution probabilities may be providing potentially more useful signals towards case outcomes, these signals are weak.** All in all, our results suggested topic model data may not be appropriate to predict case outcomes but could provide a unique lens into child welfare cases compared to risk assessment scores.

6.3 Case narratives provide a holistic lens of inquiry towards discharge outcomes (RQ2)

To further understand the signals provided by the topic probabilities used in the classifier models, we manually examined original casenotes that had the highest probabilities associated with the topics in our topic model solution. **Manual inspection of casenotes showed that all casenotes were primarily written in a neutral tone that described facts and observations of case details (in fact, caseworkers in the US are trained to write casenotes in a neutral tone [22]). At the same time, we noted these casenotes revealed specific case details through the *lens of caseworkers* as they conduct a diverse range of street-level discretionary work.** These findings added support to findings from Section 6.2 that case narratives were providing contextual information on cases. As the main purpose of this study was to understand how RA scores and topic model solutions can inform case discharge outcomes, we present our topics by high-level themes to highlight how certain topics shared common themes in a manner consistent with prior literature that have performed topic modeling on CW casenotes [94, 95]. The following paragraphs further explain these themes. We also provide paraphrased exemplar sentences related to these themes.

Theme 1 was about case details that arise as caseworkers support birth parents to achieve reunification with their children by improving and strengthening relationships with their children, helping them understand their responsibilities, and fostering teamwork between different parties (i.e., foster parents, friends and family of bio-parents). Specifically, this theme revealed case details that emerged as caseworkers helped biological parents and foster parents establish boundaries and responsibilities (topic 5) [12], mediated interactions between children and birth parents when emotions ran high (topic 12), and observed and aided interactions between bio-parents and children and between siblings to improve family relations and placement stability (topic 6 and 9).

Topic 5 paraphrased sentence: [Bio-parent] has not been able to hold down a job for more than a month. [Caseworker] spoke to parent about how important it is to have a stable job and what can happen if she does not.

Topic 6 paraphrased sentence: [Baby] was making noises and bouncing on the jumperoo and [bio-parent] sat down in front of the baby. [Caseworker] encouraged [bio-parent] to talk to the baby and explained that babies love receiving attention and love from their caregivers.

Theme 2 highlighted information that arises when caseworkers communicate regularly with bio-parents and with other caseworkers within the agency to make collaborative decisions on cases. Case information emerged as caseworkers regularly attempted to contact bio-parents regarding case updates, schedule parenting classes and supervised visits, and ensure families get the support they need (topic 1) [12]. We found caseworkers also regularly internally shared details on cases with other caseworkers at the agency to make collective case decisions and coordinate services for families (topic 2).

Topic 1 paraphrased sentence: [Casewoker] tried to get in touch with [bio-parent] but the bio-parent did not answer the phone so caseworker left a voicemail.

Topic 2 paraphrased sentence from an email: Hello team. Thanks for providing the update. I am going to try to reach the [bio-parent] for their five-day visit and let the team know when we are scheduling the visit.

[4] FPR =FP/(FP+TN)

[5] FNR= FN/(FN+TP)

[6] Specificity is calculated by TN/(TN+FP)

Dataset	Model	Accuracy	FPR	FNR	Specficity
AAPI	SVM	0.864	1	0	0
AAPI	Random Forest	0.864	0.833	0.026	0.167
NCFAS	SVM	0.836	1	0	0
NCFAS	Random Forest	0.806	0.909	0.054	0.091
AAPI + NCFAS	SVM	0.825	1	0	0
AAPI + NCFAS	Random Forest	0.825	1	0	0
TM	Random Forest	0.702	0.643	0.229	0.357
TM	SVM	0.643	0.714	0.286	0.286
AAPI + TM	SVM	0.841	1	0.026	0
AAPI + TM	Random Forest	0.841	1	0.026	0
NCFAS + TM	SVM	0.806	0.909	0.054	0.091
NCFAS + TM	Random Forest	0.806	1	0.036	0
AAPI + NCFAS + TM	SVM	0.825	1	0	0
AAPI + NCFAS + TM	Random Forest	0.800	1	0.030	0

Table 3: Performance metrics for trained classifiers using for the seven datasets. The table shows the datasets used to train and test the model (1st column); the type of machine learning algorithm used (2nd column); and the accuracy, false positive, false negative, and specificity rate of the models (last 4 columns). Note that all datasets are class-balanced prior to training.

Theme 3 showed how caseworkers manage and arrange for catered services to be delivered to their clients. Topics under this theme detailed the paperwork caseworkers helped clients fill out and admin consent forms caseworkers needed to collect from bio-parents so that the children could receive required services such as medical services (topic 3) [105] and coordinate travel arrangements for children to supervised visits (topic 11). This theme also included details on how caseworkers addressed scheduling, or familial conflicts between biological parents or caregivers to ensure visits with children were taking place (topic 10) [103].

Topic 10 paraphrased sentence: [Caseworker] talked to [bio-parent] regarding [child] and visitation. [Bio-parent] said that bio-parent tried to get in touch with [foster parents] about her kids but was not able to get in touch with them.

Topic 11 paraphrased sentence: [Foster parents] called [caseworker] to determine if there was a visit today. [Caseworker] confirmed there is one today and had discussed this with [bio-parent] about the visit last week regarding visit details.

Theme 4 highlighted the different types of services CW staff were delivering to different families in addition to transportation and legal support. Casenotes revealed how caseworkers helped families obtain economic or material resources by connecting them to employment training centers, finding kitchen appliances and toys (topic 13), and checking in with biological parents or caregivers to ensure they were keeping up with a child's medical schedule (topic 17). Depending on family needs, this theme highlighted how caseworkers also drove children to supervised visits (topic 14), facilitated virtual interactions due to COVID-19 (topic 19), and advocated for families in in-person or virtual court sessions (topics 8 and 19).

Topic 13 paraphrased sentence: [Caseworker] will help [bio-parent] secure a weight scale. [Caseworker] left a voicemail to [furniture shop] about the delivery of bio-parent's stove and fridge.

Topic 17 paraphrased sentence: In today's visit, the caseworker noted no safety concerns. [Bio-parent] is up to date with all of [child]'s medical appointments.

Theme 5 was about caseworkers noting potential risks and safety concerns in cases and mediating them with protective factors. Caseworkers observed and recorded how children were dressed and acted before, after, and during transporting children to visits (topics 4 and 16), how biological parents responded to course content and questions during parenting classes (topic 7), and how children and birth parents interacted with each other (topic 15) [27]. Caseworkers also recorded the safety and suitability of a home for a child by observing who lived in the house and the presence of appliances, furniture, toys, and food at home (topic 18).

Topic 7 paraphrased sentence: [Caseworker] met with [bio-parent] online and completed a chapter of the Active Parenting Curriculum. [Bio-parent] was engaged in the meeting and it was clear bio-parent had read the material before the meeting.

Topic 15 paraphrased sentence: [Child] begins to rock in the chair and hitting the chair on the wall. [Bio-parent] tells child to stop and child hits the wall harder. [Bio-parent] moves the chair away and the child begins rocking in the chair slowly.

7 Discussion

Abebe et al. [1]'s "Roles for Computing in Social Change" outline four ways computational work can address social issues. Per the researchers, computing as a *diagnostic* can help us understand social problems that manifest in sociotechnical systems and as a *formalizer* to clarify how rule-based models define social issues via the inputs and outputs of the models. Computing can also offer a *rebuttal* to illuminate technical limitations and a *synecdoche* to highlight existing social problems by bringing renewed focus and attention to the issues. Our study takes on these roles and provides a quantitative commentary on existing literature on CW algorithms to affect change in how researchers study and support work in the child welfare domain and more broadly in the public sector.

7.1 A quantitative lens into the gaps in risk assessment data in predicting case risk factors (RQ1)

The poor predictive ability of our trained classifiers from section 6.1 acts as a *diagnostic* [1] of child welfare decision-making algorithms by raising questions about using RA data to build such models [46, 91, 92]. The high false positive rate (FPR) in Table 3 quantitatively showed that RA data (i.e., using only AAPI, only NCFAS, or AAPI + NCFAS scores) could not help predict discharge outcomes for cases. Most times, the models predicted 'R' (reunified) for most cases and nearly all cases that actually ended in 'NR' (not reunified) were incorrectly predicted to end in 'R.' The inability of the models to correctly identify 'NR' cases suggested that the RA scores were not signaling higher risk for these cases. These results were particularly notable because discharge AAPI and NCFAS scores were included in our datasets to train the classifier models, indicating that even discharge scores collected just before a family is discharged from the program cannot provide meaningful signals to predict case outcomes. These findings suggest that the pre-defined risk domains measured in the RAs flatten a case's complexity [15] and cannot fully capture pertinent risk factors in cases, raising questions on whether we should be using RA data to inform child welfare outcomes.

The preponderance of the trained classifier models (built with RA data) in predicting 'R' outcomes for cases (shown by the FPR which equals or nearly equals 1 in section 6.1) find support for studies highlighting how different caseworkers can give very different risk ratings for the same case depending on an individual's age, experience, stress level, and personal biases [15, 84, 91, 96]. It

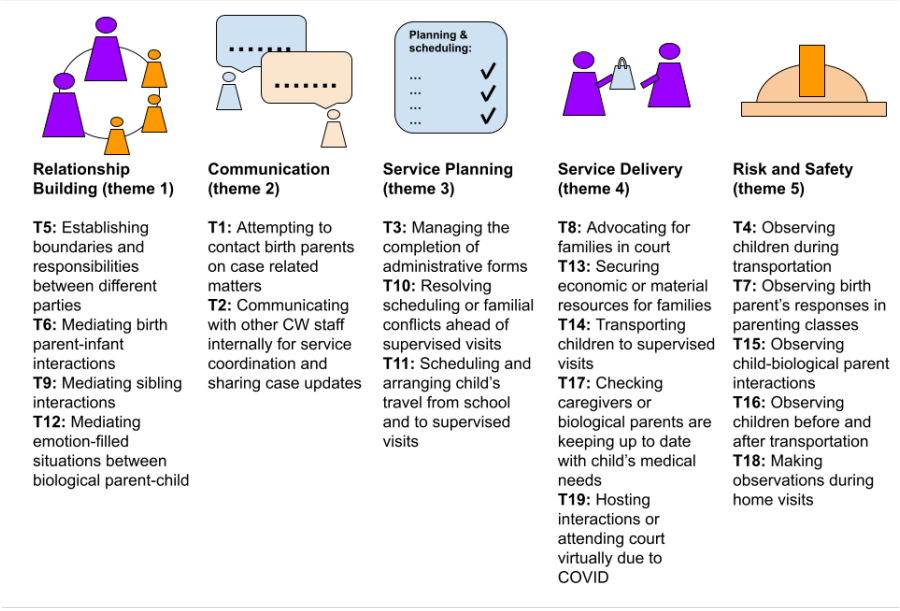

Figure 1: 19 topic model solution grouped by 5 high-level themes ('T' stands for topic).

is possible that the motivations of the biological parents and caseworkers who completed the RAs may have influenced them to give scores that favorably align with their objectives. Most times, biological parents want to be reunified with their children, and the caseworkers are part of the Family Preservation Services (FPS) team, whose main objective is to support families achieve reunification as well [95]. Previous research on caseworker interactions with CW decision-making algorithms has shown that caseworkers will often adjust algorithmic inputs to serve their needs. For example, due to limited good foster homes and to dissuade foster parents from ending placements to children, caseworkers often adjusted algorithm scores so foster parents can receive higher compensation for their services [91]. Furthermore, as AAPI and NCFAS were designed to measure theoretically derived risk-related domains from child welfare literature, caseworkers are likely working on helping families make changes to familial issues, which are constructs measured in the RAs [35, 64]. In this case, RA scores may be merely measuring the biological parent's responses to caseworker interventions [96]. Jacobs and Wallach [58] argue harms related to fairness can arise in computational systems where there exists a mismatch in the intended construct to be measured and its operationalization. Consistent with this argument, our findings question the construct validity of risk prediction models such as the Allegheny Family Screening Tool (AFST) which seeks to predict child maltreatment using proxies such as child welfare records, demographic information, or recent incarceration records [30, 107], records which in themselves are influenced by potentially biased human discretionary decisions [44].

Prior research examining perceptions surrounding predictive risk models used in child welfare through interviews with affected stakeholders (e.g., family members, caseworkers, and specialists) [18, 91, 93, 104] have found that stakeholders expressed doubt in the utility of CW algorithmic decision-making systems as they primarily focus on the deficit of cases without accounting for system-level risks. In our study, we find quantitative support for these concerns by showing how RAs can present a reductive representation of complex cases and produce biased scores that are largely irrelevant to discharge outcomes while ignoring structural causes of social issues [62, 83]. A *diagnostic* [1] can serve as a starting point for social change by raising awareness of a problem. Through this study, we add quantitative support to existing qualitative concerns surrounding the use of RAs in algorithmic decision-making tools to affect child welfare case outcomes.

7.2 Shifting away from predictive child welfare predictive analytics by investigating case narratives (RQ2)

Our results from Section 6.2 show a higher specificity rate for classifiers built using only the TM dataset compared to models built using only RA data. As a *formalizer* [1], this suggests that compared to RA data, casenotes provide different signals surrounding cases and indicators that can correctly identify 'NR' outcomes. A closer dive into the topics in our topic model solution outlined in Section 6.3 highlights that unlike the focused risk domains measured by the RAs, the topics presented nuanced details on interactions between multiple agents that can influence the trajectory and outcome of cases. However, as a *rebuttal* [1], the lower than 50% specificity rate of the classifiers to correctly predict 'NR' outcomes in Section 6.2 raises caution against using topic models for predictive tasks. We noted that when we transformed topic model probabilities into continuous features to train classifier models, this step inevitably obfuscated much of the nuanced details recorded in narrative texts. Furthermore, manual inspections of casenotes highly associated with the topics outlined in Section 6.3 showed that although caseworkers often wrote detailed casenotes in a neutral tone based on observations and facts, they were still contextualized *caseworker perspectives* that may be biased because they could choose to include or omit specific case details. For example, we found different caseworkers described transporting children to bio-parent visits differently; one could record emotions exhibited by children during transportation, while others focused on the logistics.

Consistent with findings by Saxena et al. [94] and Field et al. [46], applying textual data for prediction tasks can therefore be problematic as that would mean relying on individual casenotes for prediction. Casenotes are not uniform nor the ground truth. They record socially situated experiences where different participants in the interactions might experience the same events differently due to structural power asymmetries [36, 40, 41, 85, 86]. Roberts [86] notes that police reports can entail contradictory records on the same CW case where one officer records a house is in fair condition with

food while another records the same home is dirty with no food. In a similar vein, caseworkers may omit or underplay the oppression, surveillance, and coercion experienced by biological parents that fundamentally unpin CW interactions [94]. Furthermore, we noted that casenotes do not include all case details that may be critical to a case due to legal reasons. For example, we found that when a case involved an ongoing criminal investigation, caseworkers recorded that they omitted details following privacy laws [102].

Despite limitations on applying topic modeling on casenotes for predictive tasks, we identify key opportunities for using computational text analysis on casenotes. Nguyen et al. [78] argue that compared to smaller units of texts, texts in the aggregate can often unveil high-level themes. While caseworkers may exhibit different writing styles, our findings suggest that topic modeling on the casenote corpus can be a useful exploratory tool to uncover the dynamic and temporal dimensions in CW cases where family relations are complex and not self-isolated and exogenous organizational, legislative, and policy tensions pose systemic risks [44, 97]. A common theme that emerges from HCI research that studies how CW workers use CW algorithms is how workers do not only rely on decision-making algorithms because they deem contextual knowledge of cases [60, 61, 91], and theoretical knowledge of trauma [91] with an awareness of organizational or placement constraints [90, 91] as critical factors that affect case decisions. Consistent with Saxena at al. [95], our findings show that child welfare narratives can unveil these contextual factors mentioned above and reveal invisible labor patterns and day-to-day power dynamics between stakeholders of child welfare. Understanding such details can be helpful for developers and researchers to examine how we can better support caseworkers in managing heavy caseloads and improve the quality of services offered to families [65].

7.3 Implications for algorithms in the public sector (RQ3)

Serving as a *synecdoche* [1], our research brings attention to the long-standing tenet in child welfare, which is: striving to protect children from maltreatment by assessing child maltreatment risk [21] without fully accounting for systemic risks [86]. Through a *diagnostic* [1], we quantitatively deconstructed RA scores and casenotes, finding that pre-defined risk parameters cannot correctly predict case outcomes, nor are casenotes suitable for predictive tasks in CW. From these findings, we add quantitative support to existing scholarship on decision-making algorithms used in CW [25, 89, 91, 92], the criminal justice system [31, 42, 58], and unemployment sector [3, 56, 98] which have been criticized for using proxies that violate construct validity [75], focusing on risk factors in cases, and leading to outcomes that are biased by disproportionately affecting historically marginalized population groups [4, 44].

Through the topics uncovered in our topic model solution outlined in section 6.3, we identify opportunities for using computational text analysis to investigate and support workers in public sector sociotechnical systems. The topics outlined in section 6.3 showed that aggregated narrative texts could unveil motivations, lived experiences, and work practices of stakeholders – holistic details not provided in RA data [95]. While individual narratives carry inherent limitations wherein they detail potentially biased *perspectives* [39], on the flip side, aggregated *perspectives* offer a unique, interpretive lens into contextual lived experiences that form a part of the complexities of sociotechnical systems, including invisible labor patterns, strengths in cases, and daily power asymmetries between stakeholders [95].

The nature of public sector work involves frontline workers making discretionary decisions which are often influenced by normative and socially constructed judgments regarding risk and safety [15, 55, 56, 96]. For example, in CW, caseworkers conduct home safety inspections to determine if a home is safe for children based on culturally influenced indicators of safety and neglect (e.g., cleanliness of a home, presence of furniture) [33, 49]. As the agency trains caseworkers to include observations of the home that *back up* their safety assessments in casenotes [22, 49], we argue narratives can provide a contextual rationale for decisions made. Similar to the child welfare sector, collecting documentation is a key task in the public sector, including job placement centers and the criminal justice system [3]. Here, records pertaining to cases record the dynamic negotiations between lawyers and district attorneys in criminal cases [31] and memos between job placement caseworkers and unemployed individuals [3]. Therefore, extending our study's opportunities for computational text analysis, we propose harnessing NLP techniques on narrative texts when investigating complex public sector sociotechnical systems. Our topics uncovered in section 6.3 suggest using narrative texts as a data source can help *humanize* actors in public sector domains and help us critically and holistically understand how computational or technical interventions may or may not support workers [23]. Furthermore, by moving away from using RA and administrative data points for predictive tasks, we can avoid abstracting ourselves to data points that distill and flatten social and economic problems in the public sector [6, 23, 101].

However, here we add a note of caution. Seidelin et al. [98] find that public sector algorithms are not always used the way they were initially intended, as is the case for RAs used in child welfare [91, 92]. Further, when technical interventions are repurposed from one context to the next they may cause unintended consequences [91, 93, 99] and automation bias [44, 95]. Just as AAPI and NCFAS scores were reappropriated to predict risk even though they were initially developed as support tools to help caseworkers identify appropriate services and goals for families [11, 64], it is always possible that new interventions may diverge from their original intentions. Recently, the private sector has introduced products that claim to harness NLP to assist child welfare caseworkers in case management [7, 8]. We argue that it is possible that such tools can diverge from their original intentions and be reappropriated amidst political and policy-related factors as has been done in the past [83, 91, 98]. In light of these concerns, we suggest researchers adopt multiple lenses involving both a quantitative and qualitative lens to investigate stakeholder perceptions and the downstream effects of technical interventions.

8 Conclusion

In this study, we quantitatively deconstructed RA scores and child welfare narratives to assume the different roles of computing outlined by Abebe et al. [1]. We critique and add support to recent research that questions the construct validity of existing child welfare algorithms built on RA data by using casenotes and RA scores for cases in a US child welfare agency. In our study, we applied topic modeling to casenotes and trained random forest and support vector machines on RA scores and casenotes. We examined the predictive validity of these data sources on CW discharge outcomes. Our results show RAs could not inform discharge outcomes. We also show narrative casenotes contain caseworker perspectives on child welfare cases that are not uniform and may be biased, making casenotes unsuitable for case outcome prediction. We offer supporting evidence to move away from predictive tasks using RA data and instead find support for using contextual data such as casenotes to study sociotechnical systems in child welfare and, more broadly, in the public sector.

Acknowledgments

This research was supported by the NSERC Discovery Early Career Researcher Grant RGPIN-2022-04570 and the Connaught New Researcher award. Any opinions, findings, conclusions, and recommendations expressed in this material are those of the authors. We are grateful for the anonymous reviewers whose suggestions and comments helped improve the quality of this manuscript.

REFERENCES

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

[2] D. Allhutter, F. Cech, F. Fischer, G. Grill, and A. Mager. Algorithmic profiling of job seekers in Austria: how austerity politics are made effective. *Front. Big Data 3: 5. doi: 10.3389/fdata*, 2020.

[3] A. Ammitzbøll Flügge, T. Hildebrandt, and N. H. Møller. Street-level algorithms and ai in bureaucratic decision-making: A caseworker perspective. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–23, 2021.

[4] J. A. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine Bias, May 2016.

[5] M. Antoniak, D. Mimno, and K. Levy. Narrative paths and negotiation of power in birth stories. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27, 2019.

[6] C. Aragon, S. Guha, M. Kogan, M. Muller, and G. Neff. *Human-Centered Data Science: An Introduction*. MIT Press, 2022.

[7] Augintel. About us: Social impact meets A.I., 2020.

[8] Augintel. Allegheny county DHS case study: Unlocking the data in case notes with natural language processing, may 2022.

[9] S. Barocas and A. D. Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[10] E. P. Baumer, D. Mimno, S. Guha, E. Quan, and G. K. Gay. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, 68(6):1397–1410, 2017.

[11] S. J. Bavolek and R. G. Keene. AAPI onine development handbook the adult-adolescent parenting inventory (AAPI-2). *Family Development Resources, Inc*, 2010.

[12] S. Bekaert, E. Paavilainen, H. Scheke, A. Baldacchino, E. Jouet, L. Zablocka-Zytka, B. Bachi, F. Bartoli, G. Carra, R. Cioni, et al. Family members' perspectives of child protection services, a meta-synthesis of the literature. *Children and Youth Services Review*, p. 106094, 2021.

[13] D. M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, apr 2012. doi: 10.1145/2133806.2133826

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[15] E. A. Bosk. What counts? Quantification, worker judgment, and divergence in child welfare decision making. *Human Service Organizations: Management, Leadership & Governance*, 42(2):205–224, 2018.

[16] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. doi: 10.1191/1478088706qp063oa

[17] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001. doi: 10.1023/A:1010933404324

[18] A. Brown, A. Chouldechova, E. Putnam-Hornstein, A. Tobin, and R. Vaithianathan. 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*, p. 41. ACM, 2019.

[19] J. Burrell and M. Fourcade. The society of algorithms. *Annual Review of Sociology*, 47(1):213–237, 2021. doi: 10.1146/annurev-soc-090820-020800

[20] P. D. Caie, N. Dimitriou, and O. Arandjelović. Chapter 8 - precision medicine in digital pathology via image analysis and machine learning. In S. Cohen, ed., *Artificial Intelligence and Deep Learning in Pathology*, pp. 149–173. Elsevier, 2021. doi: 10.1016/B978-0-323-67538-3.00008-7

[21] M. J. Camasso and R. Jagannathan. Decision making in child protective services: A risky business? *Risk analysis*, 33(9):1636–1649, 2013.

[22] Capacity Building Center for States. Child Protective Services: A guide for caseworkers, 2018.

[23] S. Chancellor, E. P. S. Baumer, and M. De Choudhury. Who is the "human" in human-centered machine learning: The case of predicting mental health from social media. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019. doi: 10.1145/3359249

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, jun 2002. doi: 10.1613/jair.953

[25] H.-F. Cheng, L. Stapleton, A. Kawakami, V. Sivaraman, Y. Cheng, D. Qing, A. Perer, K. Holstein, Z. S. Wu, and H. Zhu. 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. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3491102.3501831

[26] Child Welfare Information Gateway. Family Preservation Services.

[27] Child Welfare Information Gateway. What is child abuse and neglect? Recognizing the signs and symptoms, 2019.

[28] Child Welfare Information Gateway. Concurrent planning for timely permanency for children, 2021.

[29] Child Welfare Information Gateway. The use of safety and risk assessments in child protection cases, 2022.

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

[31] A. Christin. Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society*, 4(2):2053951717718855, 2017. doi: 10.1177/2053951717718855

[32] K. Clancy, J. Chudzik, A. J. Snowden, and S. Guha. Reconciling data-driven crime analysis with human-centered algorithms. *Cities*, 124:103604, 2022. doi: 10.1016/j.cities.2022.103604

[33] O. H. R. Commission. Under suspicion: Concerns about child welfare, 2017.

[34] R. Connelly, C. J. Playford, V. Gayle, and C. Dibben. The role of administrative data in the big data revolution in social science research. *Social science research*, 59:1–12, 2016.

[35] N. A. Conners, L. Whiteside-Mansell, D. Deere, T. Ledet, and M. C. Edwards. Measuring the potential for child maltreatment: The reliability and validity of the Adult Adolescent Parenting Inventory—2. *Child Abuse Neglect*, 30(1):39–53, 2006. doi: 10.1016/j.chiabu.2005.08.011

[36] V. A. Copeland. "It's the Only System We've Got": Exploring emergency response decision-making in child welfare. *Columbia Journal of Race and Law*, 11(3):43–74, 2021.

[37] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep 1995. doi: 10.1007/BF00994018

[38] A. Coston, A. Kawakami, H. Zhu, K. Holstein, and H. Heidari. A validity perspective on evaluating the justified use of data-driven decision-making algorithms. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 690–704. IEEE, 2023.

[39] D. Das, S. Guha, and B. Semaan. Toward cultural bias evaluation datasets: The case of Bengali gender, religious, and national identity. In S. Dev, V. Prabhakaran, D. Adelani, D. Hovy, and L. Benotti, eds., *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pp. 68–83. Association for Computational Linguistics, Dubrovnik, Croatia, May 2023. doi: 10.18653/v1/2023.c3nlp-1.8

[40] A. J. Dettlaff and R. Boyd. Racial disproportionality and disparities in the child welfare system: Why do they exist, and what can be done to address them? *The ANNALS of the American Academy of Political and Social Science*, 692(1):253–274, 2020. doi: 10.1177/0002716220980329

[41] A. J. Dettlaff, S. L. Rivaux, D. J. Baumann, J. D. Fluke, J. R. Rycraft, and J. James. Disentangling substantiation: The influence of race, income, and risk on the substantiation decision in child welfare. *Children and Youth Services Review*, 33(9):1630–1637, 2011.

[42] J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, 2018. doi: 10.1126/sciadv.aao5580

[43] Edx. Adult-Adolescent Parenting Inventory-2 (AAPI-2), 2015.

[44] V. Eubanks. *Automating inequality: How high-tech tools profile,*

police, and punish the poor. St. Martin's Press, 2018.

[45] Family Development Resources. Inventory scoring system for assessing parenting practices.

[46] A. Field, A. Coston, N. Gandhi, A. Chouldechova, E. Putnam-Hornstein, D. Steier, and Y. Tsvetkov. Examining risks of racial biases in NLP tools for child protective services. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, p. 1479–1492. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3593013.3594094

[47] A. Franco, N. Malhotra, and G. Simonovits. Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203):1502–1505, 2014. doi: 10.1126/science.1255484

[48] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, 2012. doi: 10.1109/TSMCC.2011.2161285

[49] J. M. Geiger and L. Schelbe. Foster care placement. In *The Handbook on Child Welfare Practice*, pp. 219–248. Springer, 2021.

[50] M. Gerchick, T. Jegede, T. Shah, A. Gutierrez, S. Beiers, N. Shemtov, K. Xu, A. Samant, and A. Horowitz. The devil is in the details: Interrogating values embedded in the allegheny family screening tool. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, p. 1292–1310. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3593013.3594081

[51] P. Gillingham. Decision-making about the adoption of information technology in social welfare agencies: Some key considerations. *European Journal of Social Work*, 21(4):521–529, 2018.

[52] J. D. Goldhaber-Fiebert and L. Prince. Impact evaluation of a predictive risk modeling tool for Allegheny county's child welfare office. Technical report, 2019.

[53] T. Greene, G. Shmueli, J. Fell, C.-F. Lin, M. L. Shope, and H.-W. Liu. The hidden inconsistencies introduced by predictive algorithms in judicial decision making. *arXiv preprint arXiv:2012.00289*, 2020.

[54] S. Ho and G. Burke. An algorithm that screens for child neglect raises concerns., 2022.

[55] N. Holten Holten Møller, T. Rask Rask Nielsen, and C. Le Dantec. Work of the unemployed: An inquiry into individuals' experience of data usage in public services and possibilities for their agency. In *Designing Interactive Systems Conference 2021*, DIS '21, p. 438–448. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3461778.3462003

[56] N. Holten Møller, I. Shklovski, and T. T. Hildebrandt. Shifting concepts of value: Designing algorithmic decision-support systems for public services. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, pp. 1–12, 2020.

[57] Hornby Zeller Associates INC. Allegheny county predictive risk modeling tool implementation: Process evaluation. Technical report, 2018.

[58] A. Z. Jacobs, S. L. Blodgett, S. Barocas, H. Daumé, and H. Wallach. The meaning and measurement of bias: Lessons from natural language processing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, p. 706. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3351095.3375671

[59] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, pp. 429–449, 2002.

[60] A. Kawakami, V. Sivaraman, H.-F. Cheng, L. Stapleton, Y. Cheng, D. Qing, A. Perer, Z. S. Wu, H. Zhu, and K. Holstein. 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. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3491102.3517439

[61] A. Kawakami, V. Sivaraman, L. Stapleton, H.-F. Cheng, A. Perer, Z. S. Wu, H. Zhu, and K. Holstein. "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, p. 454–470. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3532106.3533556

[62] E. Keddell. 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, 2015. doi: 10.1177/0261018314543224

[63] R. S. Kirk. Psychometric properties of the trauma and post-trauma well-being assessment domains of the North Carolina Family Assessment Scale for General and Reunification Services (NCFAS G+R). *Journal of Public Child Welfare*, 9(5):444–462, 2015. doi: 10.1080/15548732.2015.1090364

[64] R. S. Kirk and K. Reed-Ashcraft. NCFAS North Carolina Family Assessment Scale Research Report. *National Family Preservation Network*, pp. 1–22, 2004.

[65] B. H. Kothari, K. D. Chandler, A. Waugh, K. K. McElvaine, J. Jaramillo, and S. Lipscomb. Retention of child welfare caseworkers: The role of case severity and workplace resources. *Children and Youth Services Review*, 126:106039, 2021.

[66] K. Kwegyir-Aggrey, M. Gerchick, M. Mohan, A. Horowitz, and S. Venkatasubramanian. The misuse of AUC: What high impact risk assessment gets wrong. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, p. 1570–1583. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3593013.3594100

[67] B. R. Lee and M. A. Lindsey. North Carolina Family Assessment Scale: Measurement Properties for Youth Mental Health Services. *Research on Social Work Practice*, 20(2):202–211, 2010. doi: 10.1177/1049731509334180

[68] Y. Lei. 3 - Individual intelligent method-based fault diagnosis. In Y. Lei, ed., *Intelligent Fault Diagnosis and Remaining Useful Life Prediction of Rotating Machinery*, pp. 67–174. Butterworth-Heinemann, 2017. doi: 10.1016/B978-0-12-811534-3.00003-2

[69] G. Lemaître, F. Nogueira, and C. K. Aridas. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[70] J. M. Lobo, A. Jiménez-Valverde, and R. Real. AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17(2):145–151, 2008. doi: 10.1111/j.1466-8238.2007.00358.x

[71] P. Marks. Algorithmic hiring needs a human face. *Commun. ACM*, 65(3):17–19, feb 2022. doi: 10.1145/3510552

[72] K. McConvey, S. Guha, and A. Kuzminykh. A human-centered review of algorithms in decision-making in higher education. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3544548.3580658

[73] N. Mickelson, T. LaLiberte, and K. Piescher. Assessing risk: A comparison of tools for child welfare practice with Indigenous families, may 2017.

[74] E. S. Y. Moon and S. Guha. A human-centered review of algorithms in homelessness research. 2024. doi: 10.1145/3613904.3642392

[75] R. K. Mothilal, S. Guha, and S. I. Ahmed. Towards a non-ideal methodological framework for responsible ml, 2024.

[76] M. Muller, S. Guha, E. P. Baumer, D. Mimno, and N. S. Shami. Machine learning and grounded theory method: Convergence, divergence, and combination. In *Proceedings of the 19th International Conference on Supporting Group Work*, pp. 3–8. ACM, 2016.

[77] National Family Preservation Network. NCFAS North Carolina Family Assessment Scale Scale Definitions (v. 2.0). 2009.

[78] D. Nguyen, M. Liakata, S. DeDeo, J. Eisenstein, D. Mimno, R. Tromble, and J. Winters. How we do things with words: Analyzing text as social and cultural data. *Frontiers in Artificial Intelligence*, 3:62, 2020.

[79] L. H. Nguyen and S. Holmes. Ten quick tips for effective dimensionality reduction. *JPLoS Comput Biol.*, 15(6), june 2019. doi: 10.1371/journal.pcbi.1006907

[80] W. S. Noble. What is a support vector machine? *Nature Biotechnology*, 24(12):1565–1567, Dec 2006. doi: 10.1038/nbt1206-1565

[81] Y. Ophir, D. Walter, and E. R. Marchant. A Collaborative Way

of Knowing: Bridging Computational Communication Research and Grounded Theory Ethnography. *Journal of Communication*, 70(3):447–472, 2020.

[82] I. D. Raji, I. E. Kumar, A. Horowitz, and A. Selbst. The fallacy of AI functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, p. 959–972. Association for Computing Machinery, New York, NY, USA, Jun 2022. doi: 10.1145/3531146.3533158

[83] J. Redden, L. Dencik, and H. Warne. Datafied child welfare services: Unpacking politics, economics and power. *Policy Studies*, 41(5):507–526, 2020.

[84] C. Regehr, M. Bogo, A. Shlonsky, and V. LeBlanc. Confidence and professional judgment in assessing children's risk of abuse. *Research on Social Work Practice*, 20(6):621–628, 2010. doi: 10.1177/1049731510368050

[85] D. E. Roberts. Racial harm: Dorothy Roberts explains how racism works in the child welfare system. *Colorlines*, 5(3):19, Fall 2002.

[86] D. E. Roberts. *Torn Apart How the Child Welfare System Destroys Black Families–and How Abolition Can Build a Safer World*. Basic Books, New York, NY, USA, 2022.

[87] S. Robertson, T. Nguyen, and N. Salehi. Modeling assumptions clash with the real world: Transparency, equity, and community challenges for student assignment algorithms. *arXiv preprint arXiv:2101.10367*, 2021.

[88] S. Robertson and N. Salehi. What if I don't like any of the choices? The limits of preference elicitation for participatory algorithm design. *arXiv preprint arXiv:2007.06718*, 2020.

[89] A. Samant, A. Horowitz, S. Beiers, and K. Xu. Family surveillance by algorithm: The rapidly spreading tools few have heard of, 2021.

[90] D. Saxena, K. Badillo-Urquiola, P. Wisniewski, and S. Guha. Child welfare system: Interaction of policy, practice and algorithms. In *Companion Proceedings of the 2020 ACM International Conference on Supporting Group Work*, pp. 119–122, 2020.

[91] D. Saxena, K. Badillo-Urquiola, P. Wisniewski, and S. Guha. A framework of high-stakes algorithmic decision-making for the public sector developed through a case study of child-welfare. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 2021.

[92] D. Saxena, K. Badillo-Urquiola, P. J. Wisniewski, and S. Guha. A human-centered review of algorithms used within the us child welfare system. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2020.

[93] D. Saxena and S. Guha. Algorithmic harms in child welfare: Uncertainties in practice, organization, and street-level decision-making. *ACM J. Responsib. Comput.*, sep 2023. doi: 10.1145/3616473

[94] D. Saxena, E. S.-Y. Moon, A. Chaurasia, Y. Guan, and S. Guha. Rethinking "Risk" in algorithmic systems through a computational narrative analysis of casenotes in child-welfare. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2023.

[95] D. Saxena, S. Y. Moon, D. Shehata, and S. Guha. Unpacking invisible work practices, constraints, and latent power relationships in child welfare through casenote analysis. CHI '22. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3491102.3517742

[96] D. Saxena, C. Repaci, M. D. Sage, and S. Guha. How to train a (bad) algorithmic caseworker: A quantitative deconstruction of risk assessments in child welfare. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pp. 1–7, 2022.

[97] C. S. Schwalbe. Strengthening the integration of actuarial risk assessment with clinical judgment in an evidence based practice framework. *Children and Youth Services Review*, 30(12):1458–1464, 2008.

[98] C. Seidelin, T. Moreau, I. Shklovski, and N. Holten Møller. Auditing risk prediction of long-term unemployment. *Proc. ACM Hum.-Comput. Interact.*, 6(GROUP), jan 2022. doi: 10.1145/3492827

[99] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, p. 59–68. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3287560.3287598

[100] A. Shlonsky and D. Wagner. 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, 2005.

[101] B. Shneiderman. *Human-Centered AI*. Oxford University Press, 01 2022. doi: 10.1093/oso/9780192845290.001.0001

[102] N. L. Sidell. *Social Work Documentation*. NASW Press, Washington, DC, USA, 2015.

[103] G. T. Smith, V. B. Shapiro, R. W. Sperry, and P. A. LeBuffe. A strengths-based approach to supervised visitation in child welfare. *Child Care in Practice*, 20(1):98–119, 2014.

[104] L. Stapleton, M. H. Lee, D. Qing, M. Wright, A. Chouldechova, K. Holstein, Z. S. Wu, and H. Zhu. Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pp. 1162–1177. ACM, 2022. doi: 10.1145/3531146.3533177

[105] Z. Strassburger. Medical decision making for youth in the foster care system. *J. Marshall L. Rev.*, 49(4):1103–1154, 2016.

[106] D. G. N. Y. Times. San Francisco had an ambitious plan to tackle school segregation. It made it worse., April 2019.

[107] R. Vaithianathan, D. Benavides-Prado, E. Dalton, A. Chouldechova, and E. Putnam-Hornstein. Using a machine learning tool to support high-stakes decisions in child protection. *AI Magazine*, 42(1):53–60, 2021.

[108] R. Vaithianathan, E. Putnam-Hornstein, A. Chouldechova, D. Benavides-Prado, and R. Berger. Hospital Injury Encounters of Children Identified by a Predictive Risk Model for Screening Child Maltreatment Referrals: Evidence From the Allegheny Family Screening Tool. *JAMA pediatrics*, 174(11):e202770, Nov 2020. doi: 10.1001/jamapediatrics.2020.2770

[109] R. Vaithianathan, E. Putnam-Hornstein, N. Jiang, P. Nand, and T. Maloney. Developing predictive models to support child maltreatment hotline screening decisions: Allegheny county methodology and implementation. *Center for Social data Analytics*, 2017.

[110] M. Veale and I. Brass. Administration by algorithm? Public management meets public sector machine learning. (3375391), 2019.

[111] M. Veale, M. Van Kleek, and R. Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pp. 1–14, 2018.

[112] M. Wieringa. What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, p. 1–18. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3351095.3372833

[113] B. Williamson. Digital education governance: Data visualization, predictive analytics, and 'real-time'policy instruments. *Journal of Education Policy*, 31(2):123–141, 2016.

[114] M. L. Wilson, W. Mackay, E. Chi, M. Bernstein, D. Russell, and H. Thimbleby. RepliCHI - CHI Should Be Replicating and Validating Results More: Discuss. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, p. 463–466. Association for Computing Machinery, New York, NY, USA, 2011. doi: 10.1145/1979742.1979491