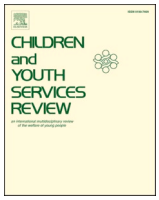
Children and Youth Services Review 153 (2023) 107117

Contents lists available at ScienceDirect

Children and Youth Services Review

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



Predicting successful placements for youth in child welfare with machine learning

Kimberlee J. Trudeau *, Jichen Yang, Jiaming Di, Yi Lu, David R. Kraus

Outcome Referrals, Inc., Framingham, MA, USA

ARTICLE INFO

Keywords:
Machine learning
Placements
Child welfare
Behavioral health
Clinical decision support system (CDSS)

ABSTRACT

Out-of-home placement decisions have extremely high stakes for the present and future well-being of children in care because some placement types, and multiple placements, are associated with poor outcomes. We propose that a clinical decision support system (CDSS) using existing data about children and their previous placement success could inform future placement decision-making for their peers. The objective of this study was to test the feasibility of developing machine learning models to predict the best level of care placement (i.e., the placement with the highest likelihood of doing well in treatment) based on each youth's behavioral health needs and characteristics. We developed machine learning models to predict the probability of each youth's treatment success in psychiatric residential care (i.e., Psychiatric Residential Treatment Facility [PRTF]) versus any other placement (AUROCs > 0.70) using data collected in standard care at a behavioral health organization. Placement recommendations based on these machine learning models distinguished between youth who did well in residential care versus non-residential care (e.g., 80% of those who received care in the recommended setting with the highest predicted likelihood of success had above average risk-adjusted outcomes). Then we developed and validated machine learning models to predict the probability of each youth's treatment success across specific placement types in a state-wide system, achieving an average AUROC score of >0.75. Machine learning models based on risk-adjusted behavioral health and functional data show promise in predicting positive placement outcomes and informing future placement decisions for youth in care. Related ethical considerations are discussed.

1. Introduction

In 2017, a total of 669,799 children were confirmed victims of maltreatment in the United States; 34% of the 442,733 children in foster care had been in more than one placement and 11% were in a group home or institution (The Annie E. Casey Foundation, KIDS COUNT Data Center, 2020). Stakes are extremely high for making the best out-of-home placement choices as some placement types and multiple placements are associated with poor outcomes (e.g., Rubin et al., 2007). In the past few years, legislation has been created to guide placement decisions for children in the care of the state, including Federal law 42 U.S. Code 675 (requires that children are placed in the least restrictive safe setting) and the Family First Prevention Services Act signed into law in 2018 (see Title VII in the Bipartisan Budget Act of 2018; U.S. Congress, 2018; requires that children are placed in the most family-like setting).

Psychiatric Residential Treatment Facility (PRTF) settings are among the most restrictive congregate care settings. They were intended for acute cases and are associated with poor outcomes but have become a long-term "solution" for many children between placements in the child welfare system (U.S. Children's Bureau, 2015). The Family First Prevention Service Act (FFPSA) shifts the risk of not having community and family-like placement settings onto states and counties. States will be required to pay the average $88,000 per year cost (ODJFS, Special Request, Median Per Diem Costs as of 10/1/17 cited in Public Children's Services of Ohio, 2017) to keep a child in residential care if not determined to be "the right level of care" by a qualified expert. This research is the first in a series of studies to provide empirical evidence to support expert determinations of the optimal level of care.

Currently available tools used by states to inform level-of-care (LOC) placement recommendations for children in care include: the Child and Adolescent Level-of-Care Utilization System (CALOCUS; Sowers et al., 2003), the Child and Adolescent Functional Assessment Scales (CAFAS; Hodges, 2000), the Child and Adolescent Needs and Strengths (CANS, Lyons, 2008) and the Treatment Outcome Package (TOP, Kraus et al.,

* Corresponding author.
E-mail address: ktrudeau@outcomereferrals.com (K.J. Trudeau).

https://doi.org/10.1016/j.childyouth.2023.107117
Received 28 March 2023; Received in revised form 31 July 2023; Accepted 3 August 2023
Available online 4 August 2023

2015). The first three tools listed require a single, expert rater to review existing evidence to recommend a LOC per child, e.g., Level 1: Foster home/kin with basic supports, Level 2: Foster home/kin with extra supports, Level 3: Therapeutic Foster Home or Foster home/kin with therapeutic supports, Level 4: Residential, and Level 5: Hospital.

CALOCUS, CAFAS, and CANS are limited by their dependence on a single, expert rater and a unidimensional recommendation (Kraus et al., 2015). Unfortunately, current child welfare decision-making includes well-documented biases, such as differential placement by race and ethnicity (e.g., Lee et al., 2015). Structured decision making (SDM; California Department of Social Services, Children and Family Services Division, 2014) has been used to guide decisions about risk; however, SDM is limited by the use of a smaller group of factors to generate recommendations (Teixeira & Boyas, 2017). Examples of other limitations of SDM according to an evaluation study in Los Angeles, CA (Nash, 2017) were: caregiver-centric data vs. family-based; manually-entered data source; lack of credibility associated with lack of transparency; and lack of documentation for overrides. There is also a well-documented lack of inter-rater reliability for non-technical solutions to support decision-making: Clinician-determined placement decisions for the same profiles using standard level-of-care criteria were near zero ($r = 0.007$) between clinicians (Bickman et al., 1997). If unsolved, a near zero reliability of placement determinations will thwart the goal of FFPSA to place all possible children in family-like settings.

Existing systems may benefit from the rich data and innovative methodological approaches being explored in other fields. For example, simulation modeling is currently being used to assist with scheduling of healthcare procedures; coordinate care across mental health services for individuals with serious mental illness (Kuno et al., 2005); assign available, proximal, and compatible (i.e., previously reported interpersonal ease and surgical outcomes) surgeon-nurse teams based on the clinical needs of incoming surgical patients (Canonico et al., 2018); match kidney donors with patients (National Academy of Sciences, 2016). Machine learning has also been used to predict psychiatric treatment outcomes from self-reported, clinician-reported, and brain imaging data (see review by Gillan & Whelan, 2017) and for decision-making in pediatric care (see Ramgopal et al., 2023), including asthma management (Seol et al., 2021).

There are an inordinate number of factors to consider in placement decision-making (Chor, 2013), making machine learning an ideal methodology to address this challenge. Machine learning is a subset of Artificial Intelligence (AI) in which historical data are used to make future predictions with varying degrees of human supervision of the modeling. People, either consciously and unconsciously, use historical data to make future predictions in everyday life: e.g., we look for discernable patterns in cause and effect, canvass others about their related experiences, and/or trust our intuition. With sophisticated analytical methods, a machine can maximize predictive validity using key data sources as inputs then provide a summary for the decision-maker's consideration.

Use of machine learning to generate recommendations is called a non-knowledge-based Clinical Decision Support Tool (Berner [2007] cited in Sutton et al., 2020) because it does not use expert-driven rules (e.g., if/then statements) to guide the recommendations. The data analyst chooses historical data sources and a specific approach for machine learning algorithm development: e.g., supervised or unsupervised. The "supervised learning" approach used in this project requires the input of specific, defined features (e.g., potential predictors like baseline well-being scores) for consideration by the machine learning models to *predict* a specific, defined target (i.e., outcome) that is either categorical ("classification") or continuous ("regression"); in contrast, an "unsupervised learning" approach means that that there is no pre-filtering of inputs in modelling to *describe* patterns in the data (Jiang et al., 2020).

Of the current placement tools identified above, only one, the TOP, maximizes inputs that benefit machine learning modeling. Rather than requiring one rater to simplify or distill the case in an attempt to create a "shared reality," the TOP encourages and seeks out discrepancies among raters. With TOP, the child has an identified voice and often self-discloses new issues and problems hidden by single-rater systems. Outliers are also seen as a signal that more team building and consensus discussions are needed, providing a roadmap for family team meetings. For machine learning predictions these additional datapoints from multiple raters often improve model accuracy.

Predictive modelling techniques in which existing data are used to predict future outcomes are already being used to assist with decision-making in child welfare. An environmental scan study by the US Department of Health and Human Services (Teixeira & Boyas, 2017) identified that predictive analytics are being used to predict many problems: risk of fatality/near fatality, repeat reporting, re-entry into the child welfare system, etc. For example, Alleghany County in Pennsylvania is using an algorithm to facilitate screening decisions about referral calls made to a child maltreatment hotline (Chouldechova et al., 2018).

The objective of this two-part study was to test the feasibility of using big data and machine learning techniques to help counties and states predict each child's likelihood of success in high end, congregate care (i.e., residential treatment program), compared to lower-cost, wraparound services like outpatient therapy, school-based therapy, etc. If successful, these models could provide additional, valuable information for consideration by teams during the high stakes, placement decision-making process that occurs around children in care.

2. Methods

This machine learning project used de-identified data that was collected in the course of clinical treatment for youth-aged behavioral health clients.

2.1. Outcome measure

The Treatment Outcome Package (TOP) is a comprehensive, validated (e.g., Kraus et al., 2015) behavioral health well-being assessment used in behavioral health and child welfare settings. TOP was also used as the outcome measure to identify the therapeutic strengths of each clinician in a RCT (Constantino et al., 2021) based on the treatment outcomes of their previous clients using risk-adjusted values developed from machine learning models. Results indicated that clients who were assigned to clinicians who had successfully treated previous clients with similar issues had better outcomes than clients assigned to clinicians through usual case assignment methods.

TOP includes three forms: the Consumer Registration, the Case-Mix, and the Clinical Scale. The TOP Consumer Registration (TOP-CR) form consists of twelve items regarding demographic characteristics (e.g., race, education). The TOP Case-Mix (TOP-CM) form includes 54 items about stressful life events, physical health, and medication use in the past 12 months and the past 30 days. The Adolescent TOP Clinical Scale (TOP-CS) is a 58-item scale for adolescents between ages 11 and 21. The Child version of the TOP-CS is a 48-item scale for children between ages 3 and 12. The TOP-CS assesses the client's past 2-week experience on 12 domains including Depression, Attention Problems, Conduct Disorders, and Suicidality, and provides a total well-being score. The TOP-CS is available to collect data in 5–10 min from raters who have knowledge of the child: caseworkers, foster parents, parents, grandparents, teachers, therapists, probation officers, and the child him/her/themselves.

Raters indicate "All" to "None of the Time" for each item on a 6-point Likert scale. Each domain score represents the number of standard deviations from the general population score (i.e., scores 1.5 and above indicate higher pathology; scores of zero indicate being on par with the general population; scores below 0 indicate potential strengths). Multiple studies have indicated that the TOP has reliability and validity (e.g., Baxter et al., 2016; Boswell et al., 2009; Kraus et al., 2010; Kraus et al., 2005). TOP Total Score is an overall score that summarizes the

child's well-being across the functional and symptom domains.

2.2. Placement data

De-identified placement data with placement name, placement start date, and placement end date for each child per treatment episode were provided by the behavioral health organizations from their electronic health record (EHR).

2.3. Procedure

The TOP is completed via self-report, as well as from the perspective of other raters who know the child or adolescent well, during the course of standard clinical care. TOP data and placement data for clients were merged in one de-identified database for analysis. This project was approved as exempt category 4 by the WCG Institutional Review Board.

2.4. Data analyses

A key component in evaluation outcomes research is to predict risk-adjusted outcomes using available data such as baseline variables and behavioral health information from the past two weeks. This risk-adjustment allows for fair comparisons across providers to identify their strengths and weaknesses. This approach is supported by previous research that concluded that a nested modelling approach is superior to a multi-level modelling approach because it provides more accurate variance estimates (Kraus et al., 2016). The algorithm that is currently used for risk-adjustment is a parallel tree boosting model implemented by an optimized distributed gradient boosted library known as XGBoost. This algorithm was trained using a de-identified database of 1.4 + million administrations. The models consider all information recently submitted for each client, then use that information to make predictions about their continued care. On average, the models explain about 35% of the variance in follow up scores for clients. This is a significant amount of the variance of future state of health for each child and can be compared to other tools like CANS (6%; Sieracki et al., 2008).

Study 1: Develop Models to Predict Placement Success in One Placement. The objective of Study 1 was to apply the new risk-adjusted models to predict doing well in PRTF (Psychiatric Residential Treatment Facility) or doing well in non-PRTF then pilot various recommendation models (e.g., predictions based on individual TOP domain scores or TOP Total Score).

The supervised Machine Learning Model development entailed three steps: Pre-processing, Model-Training, and Evaluation. A fourth step – Validation – was also implemented.

Step 1: Data Pre-processing. Data pre-processing included integrating a behavioral health outcome dataset with demographic and stressful life events datasets using Python (Van Rossum & Drake, 1995).

Missing data for each selected predictor with numeric values were replaced by that predictor's mean value. Missing categorical values were imputed with most frequent value. Treatment episodes were divided into two groups: Psychiatric Residential Treatment Facility (i.e., "ever-PRTF" –most severe level-of-care) or non-PRTF treatment (i.e., "non-PRTF") using recent demographic and clinical data. Each treatment episode was labeled as "did well in treatment" or "did not do well in treatment" in their group per domain based on whether the difference between Actual Follow-up Score and Risk-Adjusted (predicted) Score was better than the risk-adjusted clinical population mean (did well in treatment) or worse than the risk-adjusted clinical population mean (did not do well in treatment). If the unique client's outcomes were better than this risk-adjusted average, the client "did well in treatment." We have evidence of the predictive validity of this operationalization of "did well in treatment" from a recent randomized controlled trial (Constantino et al., 2021): matching new clients with clinicians who had clients who did well in treatment for similar issues approximately doubled the effect size.

There were 228 ever-PRTF treatment episodes (185 unique clients) and 2,621 non-PRTF treatment episodes (1,787 unique clients). A total of 190 features were included in the original raw dataset; from this list, a subsample of variables was created (e.g., date-related variables) and removed from further analysis. Statistical tests were used to a) evaluate if the characteristics associated with success in each type of placement group were different and b) identify the variables that were important per domain prediction. Important categorical variables selected by chi-square statistics and standardized domain features (continuous variables) were used to build each prediction model.

Step 2: Preliminary studies had indicated that a) random forest models outperformed boosting tree models in both ever-PRTF and non-PRTF predictions and b) the addition of baseline behavioral health domain data increased the performance of the modeling. Therefore, in this project we built random forest classification models per domain that included baseline domain data. Model Training with Cross-validation was used for hyperparameters optimization in developing tree-based models which can improve accuracy and stability. Repeated random sub-sampling validation was also applied during model training and evaluation in a 75% training dataset and 25% testing dataset. Instead of a single model, 50 models were built and their performances were reviewed to help understand predictive power.

Step 3: For the model Evaluation step, we used the random forest classifier machine learning models to predict the likelihood of positive treatment outcomes per domain for the ever-PRTF group and the non-PRTF group using recent demographic and clinical data. An Area Under the Receiver Operating Characteristic (AUROC) was calculated for each model. AUROC takes into consideration the True Negative, False Negative, False Positive, True Positive rate.

Step 4: Validation. For the validation step, domain-specific likelihood estimates for doing well in treatment were computed for ever-PRTF and non-PRTF for all treatment episodes. Based on looking at the difference distribution between the two likelihood estimates, we set the following threshold: If the difference between the two likelihood estimates was $<5\%$, no recommendation was made. If the difference was higher than 5%, the higher of the two likelihoods (ever-PRTF v. non-PRTF) was the recommended treatment. If the recommended treatment aligned with the treatment received by the youth, it was labeled a match. Outcomes for the client in treatment were labeled as "good" for a follow-up score that was better than the risk-adjusted value and "bad" for a follow-up score that was worse than the predicted value.

2×2 Chi squares with Yates correction were conducted to test the hypothesis that the ratio of clients with good/bad outcomes was higher for the treatment episodes that matched the treatment recommendation based on the likelihood. The percentage of clients who experienced improved outcomes in the placement setting with the highest likelihood of success was calculated for each model (per domain; TOP Total Score; Aggregate Score); 70% was our pre-stated minimum feasibility criterion.

Study 2: Develop Models to Predict Placement Success Across Multiple Placement Types. The objective of study 2 was to develop and test customized placement prediction models for multiple placement types in a state-wide child welfare system. The supervised Machine Learning Model development entailed three of the four steps described above for Study 1: Pre-processing, Model-Training, and Evaluation.

Analysis. We used machine learning modeling techniques to build random forest models fitting the data for three different placement groups: Qualified Residential Treatment Program (QRTP); Supervised Apartment Living (SAL); and Family Foster Care (FFC). Each treatment episode was labeled as "did well in treatment" or "did not do well in treatment" as defined in Study 1.

Step 1: Data Pre-processing. Data pre-processing included integrating a behavioral health outcome dataset, a demographic and stressful life events dataset, and the placement data set using Python (Van Rossum & Drake, 1995). The same data imputation methods used in Study 1 were applied here: missing numeric values were replaced by that predictor's mean values and missing categorical values were

imputed with most frequent value. Treatment episodes were divided into three placement groups (i.e., QRTP, SAL, and FFC) using the placement data from Iowa Department of Human Services (Iowa DHS) then associated with recent TOP demographic and clinical data for each client. Each treatment episode was labeled as "did well in treatment" or "did not do well in treatment" in their group, per domain.

There were 1,254 QRTP treatment episodes, 116 SAL treatment episodes, and 1,395 FFC treatment episodes. The number of features was reduced from 200 to 133 (19 numerical features and 119 categorical features) when ignored variables (e.g., date-related variables) were removed. Statistical tests were used to a) evaluate if the characteristics associated with success in each type of placement group were different and b) identify the variables that were important per domain prediction. To reduce dimensionality, we used chi-square tests to obtain only the most significant features for the model training.

Step 2: Model Training. We explored recursive feature elimination to obtain most significant features for our model training. Unfortunately, the accuracy was not significantly higher than our current model. Considering the runtime cost of recursive feature elimination, we concluded with the models that were a combination of chi-square tests and random forest classifier. Model Training with Cross-validation was used for hyperparameters optimization in developing tree-based models which can improve the accuracy and stability. Repeated random subsampling validation was also applied during model training and evaluation. As in Study 1 above, instead of a single model, 50 models were built per recommendation type candidate (75% training dataset and 25% testing dataset) and their performances were reviewed to help understand the prediction power.

Step 3: Evaluation: For the model evaluation step, we used the random forest models to predict the likelihood of positive treatment outcomes per domain for QRTP, SAL, and FFC groups using recent TOP demographic and clinical data. Again, an AUROC was calculated for each model.

3. Results

3.1. STUDY 1 models predicting success in one placement type

Demographics. Children's Hope Alliance (CHA), is a large non-profit behavioral healthcare agency that provides services for clients involved in the Juvenile Justice, Mental health, Child Welfare, and Public Health systems across the state of North Carolina. Data were collected between 2017-07-17 and 2021–10-07 from 1,145 youth (average age of 15.86 years old; SD = 1.53; range = 7.3 to 19.53) through standard clinical practice. Their demographic characteristics were: female 39%, male 59%, and transgender 1%; Asian 1.2%, Black/African American 25.5%, Hispanic/Latino 7.3%), Mixed Race < 1%, Native American 4.3%, White 53.0%, and Unreported 8.6%.

Model Development: The range of AUROC is from 0 to 1. The average of AUROCs for 50 models per domain was recorded and plotted. Non-PRTF models had a mean AUROC = 0.70 (SD = 0.03) with a range of 0.66 to 0.77. See Fig. 1 for the AUROCs of models predicting success in non-PRTF.

Ever-PRTF models had a mean AUROC = 0.70 (SD = 0.09) with a range of 0.46 to 0.79, thereby achieving our a priori target of 0.70 or higher. See Fig. 2 for the AUROCs of models predicting success in PRTF.

Validation: All Chi square values were significant. The percentage of clients who experienced improved outcomes in the placement setting with the highest likelihood of success ranged between 51% (Lack of Resiliency, a Child TOP domain) and 87% (Substance Abuse, an Adolescent TOP domain). In addition, the models predicting doing well on other domains (individual domains like Depression or combined like TOP Total Score) were also supported in this validity test: the TOP Total Score model (80%) and the Aggregate Score model (average of Conduct, Substance Abuse, Suicidality, Violence, Worrisome Sexual Behaviors, and Psychosis; 79%). These domains were selected for the Aggregate Score because they are factors that are strongly considered in expert-based placement decisions for PRTF. The most severe score model had the lowest value (57%). See example of evaluation computations for TOP Total Score in Table 1.

Values for all domains are included in Table 2.

The models met our a priori feasibility criterion for model evaluation of a minimum AUROC of 70%. For application of the models in practice, it is also necessary to have a) variability in recommendations for PRTF versus non-PRTF and b) many cases with a decisive PRTF v. non-PRTF recommendation (>5% difference between likelihood predictions and minimal missing data). Models that a) predicted either ever-PRTF or non-PRTF in almost all cases (e.g., the Aggregate Score models recommended non-PRTF for 99% of cases) or b) lacked placement recommendations for ¾ of the cases (e.g., Substance Abuse) were not good candidates for implementing the model recommendations in practice.

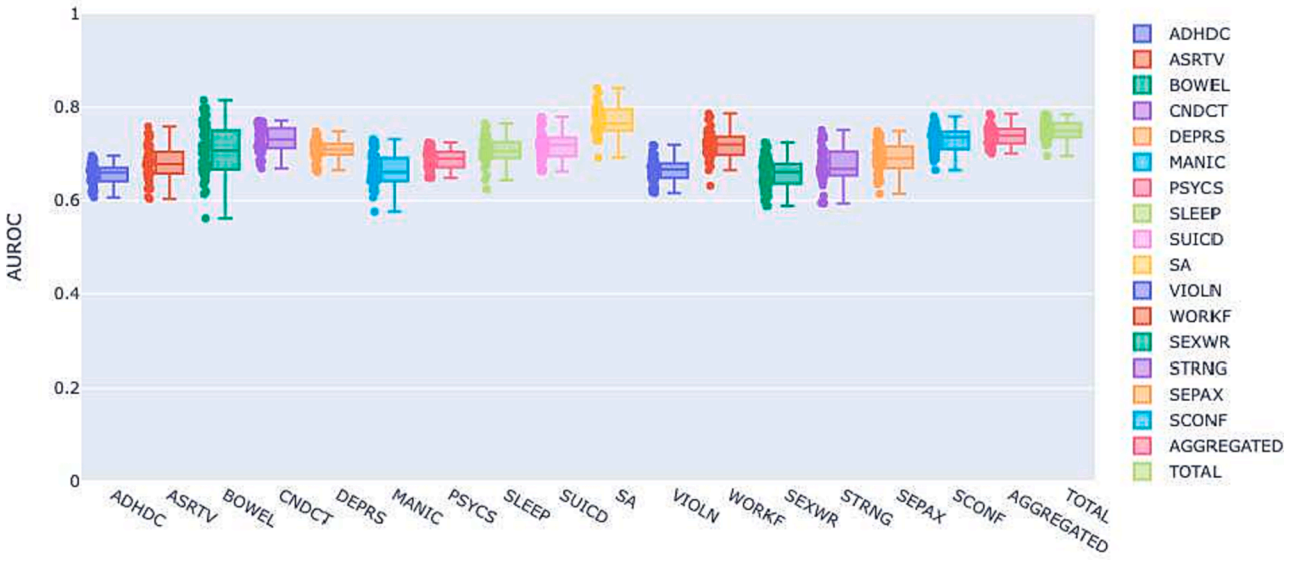


Fig. 1. Random forest models for nonPRTF Users with Stratified Sampling.

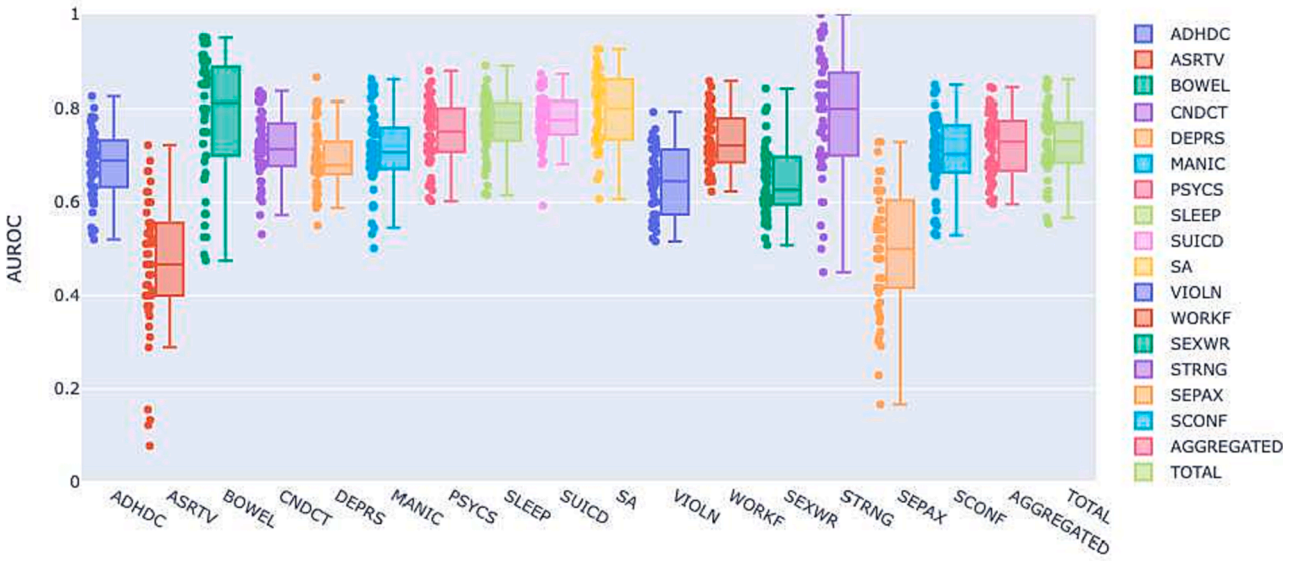


Fig. 2. Random forest models for PRTF Users with Stratified Sampling.

Table 1
Evaluation computations for TOP Total Score.

Example: TOP TOTAL SCORE	Bad (outcome worse than prediction) on TOP Total Score	Good (outcome better than prediction) on TOP Total Score	Good/ Bad ratio
Treatment **Matched** Recommendation	128	522	522/128 =4.08
Treatment **Mismatched** Recommendation	131	164	164/131 =1.25

Chi sq = 61.06, p = 5.54e-15 Summary: 80% (522) of the total evaluations (650) that were placed in the matched condition had a TOP TOTAL SCORE outcome score better than the predicted value.

4. Study 2: Models predicting success in multiple placement types

Demographics. Data were collected from clients in the Iowa DHS child welfare system between 2017-05-22 and 2022-06-14. A total of 2,225 youth (average age of 13.94 years old; SD = 3.24) were included. Their demographic characteristics were: female 40.3%, male 59.7%; Asian < 1%, Black/African American 16.6%, Hispanic/Latino 8.1%), Mixed Race 5.6%, Native American 2.5%, White 63.1%, and Unreported 3.1%.

Model Development. The average of AUROCs for 50 models for TOP Total Score was recorded and plotted for each placement type: Qualified Residential Treatment Program (QRTP); Supervised Apartment Living (SAL); and Family Foster Care (FFC). QRTP models had a mean AUROC = 0.80 (SD = 0.03) with a range of 0.73 to 0.89; SAL models had a mean AUROC = 0.88 (SD = 0.10) with a range of 0.62 to 1.00. Although it is unexpected to see a maximum range of 1.0 for an AUROC score, we hypothesize that it is associated with the small sample size for SAL (N = 116). FFC models had a mean AUROC = 0.74 (SD = 0.03) with a range of 0.68 to 0.81 for TOP Total Score and a mean AUROC for Aggregate Score = 0.78 (SD = 0.037) with a range of 0.69 to 0.87, suggesting that the Aggregate Score would be the preferred recommendation type for generating models for placement in FFC. The range of AUROCs for the TOP Total Score are presented in Fig. 3.

5. Discussion

Choosing the right treatment for a youth in care is a very important and challenging decision that can have long-term impact on a youth's well-being. Traditionally, consensus-based expert systems like the Child and Adolescent Level-of-Care Utilization System (CALOCUS; Sowers et al., 2003) have been used to inform placement decision-making but there is a growing recognition of the power and benefits of machine learning tools to inform decision-making in child welfare (Chouldechova et al., 2018; Teixiera & Boyas, 2017). In other words, static and linear decision trees built on the consensus of a panel of experts can be enhanced with a growing dataset of actual results from a growing knowledgebase of previous decisions made and their consequences (i.e., child outcomes).

With access to more digital data than ever before, we have an opportunity to use innovative big data techniques to improve decision-making of treatment options for youth in care. The goal of this project was to assess the feasibility of creating machine learning models to predict placement success for youth in care using data from youth, their caregivers, and providers.

Two studies were conducted with two different samples from two different states (behavioral health in NC; child welfare in IA). Multiple machine learning models achieved the a priori feasibility criteria of 0.70 or higher metrics (Area Under the Receiving Curve; AUROC). In this study, success was not only defined as making statistically significant progress in care. Success was further defined as a child achieving above-average risk-adjusted outcomes. When children were placed in a level of care that the model would *not* have recommended, the chance of above average success statistics did not change much (the good outcome to bad outcome ratio for TOP Total Score was 1.25; see Table 1). However, when the child was placed in the setting concordant with the model's recommendation the number of children who had above average outcomes were four times higher than those that had below average outcomes.

With the success of this first study, in future modeling we plan to expand the level of care decisions to be non-binary and include all levels of care or placement types that a specific jurisdiction has available. This will increase the chances that the recommendations are useful and relevant to each locality.

5.1. Limitations

The majority of recommendations using the Total TOP Score models were indeterminate because the likelihood of success predictions for PRTF and non-PRTF were < 5% different. However, since the cost of the

Table 2
Summary of Evaluation Computations for All Domains.

DOMAIN	Precision Metric: # Good outcome with Recommended Treatment/Total # of Matched Evaluations for that domain	Treatment Matched Recommendation Good/Bad Ratio	Treatment did NOT Match Recommendation Good/Bad Ratio	Recommendation = None*	Recommendation: PRTF	Recommendation: Non-PRTF
**Substance Abuse	87%	6.77	0.98	75%	8%	17%
**Conduct Problems	85%	5.82	0.71	28%	2%	70%
Sleep Problems	85%	5.82	2.18	61%	34%	5%
Incontinence	84%	5.07	2.17	17%	30%	53%
**Psychosis	83%	4.76	0.81	75%	13%	12%
**Suicidality	83%	4.76	2.65	17%	13%	70%
Total TOP Score	80%	4.08	1.25	58%	10%	32%
Aggregate Score of ** Domains (ave)	79%	3.76	0.64	14%	1%	86%
**Sexually Worrisome Behavior	77%	3.34	0.97	15%	1%	84%
Social Conflict	76%	3.13	0.39	43%	21%	37%
Depression	75%	2.95	0.76	23%	4%	72%
Lack of Assertiveness	72%	2.54	0.32	5%	1%	94%
Separation Anxiety	67%	2.05	0.33	11%	5%	84%
Mania	66%	1.95	0.44	7%	3%	90%
Work Functioning	63%	1.73	0.42	14%	7%	80%
Attention Problems	62%	1.65	0.46	63%	26%	11%
**Violence	61%	1.59	0.37	5%	0.40%	95%
Most Severe Domain	57%	1.32	0.32	71%	8%	20%
Lack of Resiliency	51%	1.03	0.18	12%	9%	79%

* <5% difference in likelihood of success in PRTF v non-PRTF.
** Domain included in the Aggregate (ave) models.

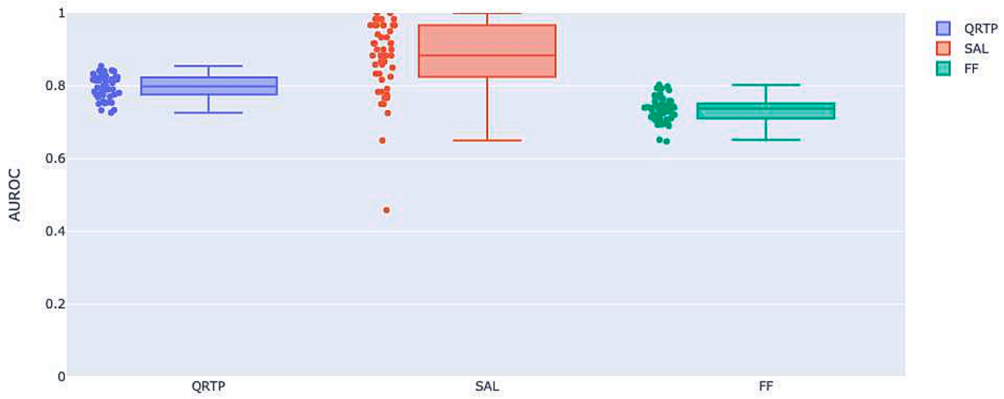


Fig. 3. Random forest models with TOP Total Score for QRTP, SAL, FF Users with Stratified Sampling.

two options are significant (PRTF costing significantly more depending on the region of the country), we submit these finding are relevant to making placement decisions. Each payor of services will need to determine the increase in success rates that warrant a higher cost and increasingly restrictive placement setting. A <5% difference may not justify the cost.

This research has significant implications on improving outcomes and reducing the cost of care. Implementing a data solution carries costs, but they are likely offset by a reduction in the cost of care. In this study only 10% of cases were judged by the predictive modeling to have a significantly higher chance of success at this higher-cost level of care.

Another limitation is the lack of generalizability of machine learning models to predict success in specific placements. Because machine learning models are developed with large quantities of historical data, it is necessary to have sufficient data to customize models for specific placements in each regional application. At least 100 clients per placement is necessary to develop supervised machine learning models based on our previous work and the literature (Beleites et al., 2013).

A potential future limitation is lack of integration of these recommendations into the placement decision-making workflow for youth in care. In a recent scoping review of implementing machine learning based tools in decision-making about patient care in hospitals (Tricco et al., 2023), barriers associated with implementing these tools were that they were time-consuming and unreliable. While we were developing and testing the feasibility of using machine learning models for this proposed Clinical Decision Support System, we conducted

interviews with professional staff who participate in placement decision-making process at a behavioral health organization. They were asked: "Where would a tool like this fit into your current workflow when making placement decisions? Five separate responses were provided: 1) When youth come into care; 2) When reviewing current clinical outcome data; 3) During discussions about levelling up in care; 4) At Monthly Child and Family Team meetings; and 5) At Discharge planning. In addition, they wanted the Placement Success Predictor tool to be part of the current routine outcome monitoring system (i.e., WellnessCheck®) so that they could generate recommendations while viewing other relevant clinical information about current challenges and supports. The state of Iowa has used this feedback in designing how this information will be available in their new child welfare data systems that put the Placement Success Predictions on every child's dashboard and prompt the user to justify use of potentially lower success placements.

Lastly, because of the potential for profiling and perpetuating existing social biases in our culture, such as racism (Glaberson, 2019), we generated our models with and without race variables and confirmed that the AUROC values were not changed significantly. Therefore, we concluded that race was not an important predictor and retained those variables in our final models.

FUTURE DIRECTIONS: ETHICAL AND LEGAL CONSIDERATIONS. While the objective of this study was to assess the feasibility of developing models to predict the likelihood of success in different treatment modalities from baseline characteristics using machine learning, the ultimate goal is to create a clinical decision support system to assist with placement decision making for children in care. The proposed system will use machine learning models customized per organization and regularly updated to ensure relevance of the results for each sample. It will also include information about the likelihood of success in every placement (not just present the "best" one with the highest likelihood of success) because there are a multitude of contextual/logistical variables (e.g., availability of slots for new clients; distance of placement from client's home; insurance eligibility) that are considered when teams are making placement decisions.

Development of clinical decision support systems is federally supported by the 2009 HITECH ACT as "meaningful use of health information technology" (USDHHS, 2009). Yet, there is growing public concern about Artificial Intelligence-based tools having a voice in clinical decision making. E.g., What, then, is the role of the provider? How do we minimize the potential risks of bias that could be built into the models? Fortunately, these questions are already being explored in the medical field (e.g., Beam et al., 2023; Ramgopal et al., 2023; Sutton et al., 2020). Their answers take the form of professional and legal guidelines for evaluating machine learning research and machine learning-based devices. There are existing evolving tools to assess (and guide) machine learning research for risk of bias that consider factors such as appropriate data sources, well-defined predictors, relevant outcomes, and robust analytics (e.g., PROBAST; TRIPOD – see Collins et al., 2021). Depending on the application, many machine learning-based devices may require review for FDA approval when the user cannot critically evaluate the client-specific input that informed the model's recommendation (e.g., FDA Clinical Decision Support - September 2022). Guidelines plus provider critical thinking act as guardrails to protect the integrity of the objective to utilize innovative analytics as one of the tools in the provider's tool box to support, not replace, provider driven clinical decision making.

6. Conclusion

The machine learning models developed during this project appear to be robust in distinguishing clients who will do well in treatment versus not do well in treatment for various placement types. Use of a clinical decision support system (CDSS) with machine learning model-based placement recommendations could potentially improve overall outcomes while helping to reduce healthcare costs if high-cost options intended for acute care are provided only to the clients who are most likely to benefit from them.

7. Declarations

Funding

This study was funded by NIH SBIR grant # 1R43MH125486-01. Preliminary work was supported by an award from The Duke Endowment to Children's Hope Alliance (NC).

Financial **interests**

All authors were compensated by Outcome Referrals, Inc. to work on this project.

Ethics approval

This project was approved as exempt category 4 by the WCG Institutional Review Board.

CRediT authorship contribution statement

Kimberlee J. Trudeau: Conceptualization, Funding acquisition, Methodology, Writing – original draft, Writing – review & editing. **Jichen Yang:** Methodology, Formal analysis, Investigation, Conceptualization, Writing – review & editing. **Jiaming Di:** Methodology, Formal analysis, Investigation, Writing – review & editing. **Yi Lu:** Formal analysis, Investigation, Writing – review & editing. **David R. Kraus:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Declaration of Competing Interest

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

Data availability

The data that has been used is confidential.

Acknowledgements

This study was funded by NIH SBIR grant # 1R43MH125486-01. Preliminary work was supported by an award from The Duke Endowment to Children's Hope Alliance (NC). Thank you to our clinical partners – Matt Haynes (Iowa Department of Health and Human Services) and Lakisha Marelli (Children's Hope Alliance) – for inspiring, supporting, and disseminating our collaborative work.

References

Baxter, E. E., Alexander, P. C., Kraus, D. R., Bentley, J. H., Boswell, J. F., & Castonguay, L. G. (2016). Concurrent validation of the child and adolescent versions of the Treatment Outcome Package (TOP). *Journal of Child and Family Studies, 25*(8), 2415–2422.

Beam, A. L., Drazen, J. M., Kohane, I. S., Leong, T. Y., Manrai, A. K., & Rubin, E. J. (2023). Artificial Intelligence in Medicine. *The New England Journal of Medicine, 388* (13), 1220–1221. https://doi.org/10.1056/NEJMe2206291

Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., & Popp, J. (2013). Sample size planning for classification models. *Analytica Chimica Acta, 760*, 25–33. https://doi.org/10.1016/j.aca.2012.11.007

Bickman, L., Karver, M. S., & Schut, J. A. (1997). Clinician reliability and accuracy in judging appropriate level of care. *Journal of Consulting and Clinical Psychology, 65*(3), 515–520. https://doi.org/10.1037/0022-006X65.3.515

Boswell, F., Kraus, D. R., Nordberg, S., S., & Castonguay, L. G., (2009, June). *The Treatment Outcome Package (TOP): An investigation of its validity*. Poster presented at the 39th annual meeting of the Society for Psychotherapy Research, Santiago, Chile.

Canonico, L. B., McNeese, N. J., & Shuffler, M. L. (2018). Stable teamwork marriages in healthcare: Applying machine learning to surgeon-nurse-patient matching. *Sage Journals, 62*(1), 1202–1206. https://doi.org/10.1177/1541931218621276

Chor, K. H. B. (2013). Overview of out-of-home placements and placement decision-making in child welfare. *Journal of Public Child Welfare, 7*(3), 298–328. https://doi.org/10.1080/155 48732.2013.779357

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, in Proceedings of Machine Learning Research, 81,* 134-148. proceedings.mlr.press/v81/chouldechova18a/chouldechova18a.pdf.

Collins, G. S., Dhiman, P., Andaur Navarro, C. L., Ma, J., Hooft, L., Reitsma, J. B., … Moons, K. G. (2021). Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open, 11*(7), e048008.

Constantino, M. J., Boswell, J. F., Coyne, A. E., Swales, T. P., & Kraus, D. R. (2021). Effect of matching therapists to patients vs assignment as usual on adult psychotherapy outcomes: A randomized clinical trial. *JAMA Psychiatry, 78*(9), 960–969.

Food and Drug Administration. (2022). *Clinical Decision Support Software - Guidance for Industry and Food and Drug Administration Staff (fda.gov)*. https://www.fda.gov/media/109618/download.

Gillan, C.M., & Whelan, R. (2017). What big data can do for treatment in psychiatry. *Science Direct, 18*, 34–42. https://doi.org/10.1016./j.cobeha.2017.07.003.

Glaberson, S. K. (2019). Coding over the cracks: Predictive analytics and child protection. *Fordham Urban Law Journal, 46*, 307. https://ir.lawnet.fordham.edu/ulj/vol46/iss2/3/.

Hodges, K. (2000). *Child and Adolescent Functional Assessment Scale*. Ypsilanti, MI: Eastern Michigan University.

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised Machine Learning: A Brief Primer. *Behavior Therapy, 51*(5), 675–687. https://doi.org/10.1016/j.beth.2020.05.002

Kraus, D. R., Baxter, E. E., Alexander, P. C., & Bentley, J. H. (2015). The Treatment Outcome Package (TOP): A multi-dimensional level of care matrix for child welfare. *Children and Youth Services Review, 57*, 171–178. https://doi-org.du.idm.oclc.org/10.1016/j.childyouth.2015.08.006.

Kraus, D. R., Bentley, J. H., Alexander, P. C., Boswell, J. F., Constantino, M., & Baxter, E. E. (2016). Predicting therapist effectiveness from their own practice-based evidence. *Journal of Consulting and Clinical Psychology, 84*(6), 473–483. https://doi.org/10.1037/ccp0000083

Kraus, D. R., Boswell, J. F., Wright, A. G. C., Castonguay, L. G., & Pincus, A. L. (2010). Factor structure of the Treatment Outcome Package for children. *Journal of Clinical Psychology, 66*(6), 627–640. https://doi.org/10.1002/jclp.20675

Kraus, D. R., Seligman, D., & Jordan, J. R. (2005). Validation of a behavioral health treatment outcome and assessment tool designed for naturalistic settings: The Treatment Outcome Package. *Journal of Clinical Psychology, 61*(3), 285–314.

Kuno, E., Koizumi, N., Rothbard, A. B., & Greenwald, J. (2005). A service system planning model for individuals with serious mental illness. *Mental Health Services Research, 7*(3), 135–144. https://doi.org/10.1007/s11020-005-5782-5

Lee, J., Bell, Z., & Ackerman-Brimberg, A. (2015). *Implicit Bias in the Child Welfare, Education and Mental Health Systems*. National Center for Youth Law. https://youthlaw.org/wp-content/uploads/2015/07/Implicit-Bias-in-Child-Welfare-Education-and-Mental-Health-Systems-Literature-Review_061915.pdf.

Lyons, J. (2008). Child and Adolescent Needs and Strengths (CANS) comprehensive multisystem assessment manual, 1–28. https://www.praedfoundation.org.

Nash, M. (2017). *Examination of using structured decision making and predictive analytics in assessing safety and risk in child welfare* (Item No. 49-A). County of Los Angeles Office of Child Protection. https://www.childwelfare.gov/topics/responding/child-protection/decision-making/.

National Academy of Sciences. (2016). Matching kidney donors with those who need them- and other explorations in economics. *From Research to Reward: A National Academy of Sciences Series about Scientific Discovery and Human Benefit.* https://doi.org/10.17226/ 23508.

Public Children's Services of Ohio. (2017). *Factors impacting placement costs: What drives placements, strategies to control costs, and future challenges*. https://ccao.org/wp-content/uploads/17%20Nov%202nd%20Weds%20-%20Child%20Placement%20Costs.pdf.

Ramgopal, S., Sanchez-Pinto, L. N., Horvat, C. M., Carroll, M. S., Luo, Y., & Florin, T. A. (2023). Artificial intelligence-based clinical decision support in pediatrics. *Pediatric Research, 93*(2), 334–341. https://doi.org/10.1038/s41390-022-02226-1

Rubin, D. M., O'Reilly, A. L., Luan, X., & Localio, A. R. (2007). The impact of placement stability on behavioral well-being for children in foster care. *Pediatrics, 119*(2), 336–344. https://doi.org/10.1542/peds.2006-1995

Sieracki, J. H., Leon, S., Miller, S. A., & Lyons, J. S. (2008). Individual and provider effects on mental health outcomes in child welfare: A three level growth curve approach. *Children and Youth Services Review, 30*, 800–808.

Seol, H. Y., Shrestha, P., Muth, J. F., Wi, C. I., Sohn, S., Ryu, E., … Juhn, Y. J. (2021). Artificial intelligence-assisted clinical decision support for childhood asthma management: A randomized clinical trial. *PLoS One1, 16*(8), e0255261.

Sowers, W., Pumariega, A., Huffine, C., & Fallon, T. (2003). Level-of-care decision making in behavioral health services: The LOCUS and the CALOCUS. *Psychiatric Services, 54*, 1461–1463. https://ps.psychiatryonline.org/doi/10.1176/appi.ps.54.11.1461.

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: Benefits, risks, and strategies for success. *NPJ Digital Medicine, 3*, 17. https://doi.org/10.1038/s41746-020-0221-y

Teixeira, C., & Boyas, M. (2017). Predictive analytics in child welfare. An assessment of current efforts, challenges and opportunities. *U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation*. https://aspe.hhs.gov/syste m/files/pdf/257841/PACWAnAssessmentCurrentEffortsChallengesOpportunities.pdf.

The Annie E. Casey Foundation (2020). *KIDS COUNT Data Center,* https://datacenter.kidscount.org.

Tricco, A. C., Hezam, A., Parker, A., Nincic, V., Harris, C., Fennelly, O., … Straus, S. E. (2023). Implemented machine learning tools to inform decision-making for patient care in hospital settings: A scoping review. *BMJ Open, 13*(2), e065845.

United States Children's Bureau. (2015). A national look at the use of congregate care in child welfare. *U.S. Department of Health and Human Services, Administration for Children and Families*. https://www.acf.hhs.gov/sites/default/files/cb/cbcongregatecare_brief.pdf.

United States Congress. (2018). *H.R. 1892 Bipartisan Budget Act of 2018*. https://www.congres s.gov/115/plaws/publ123/PLAW-115publ123.pdf.

United States Department of Health & Human Services. (2009). HITECH Act Enforcement Interim Final Rule. Washington, DC, US Department of Health & Human Services, 2009. https://www.hhs.gov/hipaa/for-professionals/special-topics/hitech-act-enforcement-interim-final-rule/index.html.

Van Rossum, G., & Drake, F. L., Jr (1995). Python Reference Manual. *Centrum voor Wiskunde en Informatica Amsterdam.*