MAY 2022

SAMANTHA XIA
ZARNI HTET
KRISTIN E. PORTER
MEGHAN MCCORMICK

EXPLORING THE VALUE OF PREDICTIVE ANALYTICS FOR STRENGTHENING HOME VISITING

Evidence from Child First

Predictive analytics, or the use of historical data to forecast future outcomes, has long been a feature of business and marketing research. But increasingly, predictive analytics is being applied in the social service and education domains. Program administrators can use relatively simple approaches, such as models with a limited number of measures, to predict outcomes of interest. Or they may consider more complex machine-learning models, taking advantage of the large amounts of data that organizations often collect, to improve their ability to make predictions. Either way, the promise of predictive analytics is to help programs identify those clients who could most benefit from targeted interventions—facilitating effective service delivery at an efficient cost.

An ongoing research partnership between MDRC and Child First, a home visiting program that provides therapeutic support and services to families with young children, offered the opportunity to examine the potential benefits, if any, of using predictive analytics to improve service delivery. To date, these methods have had limited applications in the home visiting domain. This brief offers results from that proof-of-concept exercise, in which the team faced challenges in identifying a reliable model. Additionally, the brief provides much-needed information on the value of predictive analytics for similar organizations and may be a helpful guide for future researchers and practitioners, as more programs seek to implement these cutting-edge analytics tools.

USING PREDICTIVE ANALYTICS TO STRENGTHEN SOCIAL PROGRAMS

Although predictive analytics has rarely been used in studies of individual home visiting programs, there is a growing body of work that has applied these methods to programs in related fields, with various degrees of success. For example, the Allegheny Family Screener Tool, first implemented in 2016, was developed in a partnership between researchers from Auckland University of Technology and the Allegheny County Office of Children, Youth, and Families, a Pennsylvania child welfare system. The research team wanted to use predictive analytics to help inform and improve decisions made by staff when determining whether reports of possible child abuse and neglect should be marked for further investigation, rather than replace human decision making altogether. The tool summarizes vast amounts of information across multiple databases to provide a risk score to child welfare call screeners. The researchers worked closely with the child welfare agency and partner organizations to discuss implementation of the tool and results, and feedback from community meetings informed how the tool was developed. An independent evaluation found

mdrc
BUILDING KNOWLEDGE TO IMPROVE SOCIAL POLICY

that it increased the staff's ability to accurately screen reports and pursue investigations. Also, the tool did not increase the rate of children screened in for investigation. That is, using it resulted in a *different* pool of children being identified as needing child welfare intervention, but did not substantially increase the proportion of children investigated among all children referred for maltreatment. The model and its implementation have been updated over time so that its predictions reflect contemporary information on families currently being served.[1]

Using predictive analytics in the Pittsburgh Public Schools yielded more mixed results. An evaluation compared the district's prior performance early warning system to an algorithm that used a range of in-school and out-of-school data to identify students at risk of chronic absenteeism, course failure, low grade point average, or suspension. The evaluation found that the prior performance early warning system and the risk scores created by the algorithm were just as accurate for some purposes, but there were also tradeoffs, such as the need to balance performance and complexity.

There were two key takeaways from this evaluation. First, because the warning system and the algorithm performed similarly, the cheaper prior performance warning system might be sufficient if resources are limited. Second, if using algorithm risk scores, districts must choose between over-including and under-including students in a given intervention. For example, if districts over-include, they could risk spending too many resources on costly and labor-intensive services and potentially stigmatizing a large group of students. It may be better, the research team concluded, to include a narrow group of students if the intervention could be perceived negatively, whereas over-inclusion may be acceptable for cheaper interventions with little risk of stigma.[2]

However, using predictive analytics has not always yielded useful results. The Fragile Families Challenge (FFC), an offshoot of a large-scale longitudinal study designed to understand the lives of families formed by unmarried parents, set out to develop models that would predict six different outcomes by making vast amounts of data available to a large group of researchers. FFC received 160 different submissions from teams targeting a wide range of outcomes, such as material hardship, household eviction, and primary caregiver participation in job training. Ultimately, even the best predictions were not very accurate and were only slightly better than those from a simple benchmark model that used regression and four predictor variables selected by an expert. Thus, there seem to be practical limits to the predictability of life outcomes in certain settings.[3] At the same time, FFC relied on longitudinal survey data. Other types of data sources, like those from program administrators, often include rich information about clients' *current* situations and therefore may provide more predictive information for near-term events.

THE CHILD FIRST HOME VISITING PROGRAM

Home visiting programs connect directly with families in their homes, providing a broad set of services to support caregivers' and children's health and well-being. Child First is a home visiting program that aims to promote high-quality relationships between caregivers and children in families

experiencing challenges related to caregiver mental health and child behavior. Staff members provide intensive in-home clinical services to both the caregiver and child. They also connect families to additional services such as financial and housing support, health care, and treatment for disorders such as substance abuse. To accomplish its goals, Child First must ensure that families remain consistently engaged in program services over time.[4] When families leave before being officially discharged from the program, they receive truncated interventions that are likely to be less effective at improving outcomes.[5] Early disengagement is also expensive, given the large, fixed costs of enrolling new families into the program.

Child First prioritizes collecting high-quality data to understand the population of families it serves. For example, the program collects information on a range of family characteristics assessed at intake, including sociodemographic information on both the caregiver and child, health-related information like insurance status and child DSM-5 diagnoses, as well as several risk measures and assessments such as adverse childhood experiences (ACEs) that capture experiences with violence, abuse or neglect, and household instability.[6] The availability of these data provided researchers with a singular opportunity to explore whether predictive analytics could be a useful tool to summarize the large amounts of information the program collects and use it to help staff members identify families at particularly high risk of **early disengagement**, defined as being enrolled in the program for fewer than 90 days. That information could allow Child First staff members to triage families better at intake and provide more intensive support and services to those families at risk of early disengagement.

MDRC's ongoing work with Child First includes a replication study of earlier program impact findings. As a natural outgrowth of that work, MDRC presented Child First with a proposal to learn more about predictive analytics and the problem of family disengagement. Child First agreed that this would be a worthwhile enterprise because predictions of family disengagement could help staff members target limited resources to new participants most in need. The predictive models could be used in conjunction with staff members' knowledge and experience as a check to ensure that families did not fall through the cracks, similar to the implementation of predictive analytics by child welfare agencies in Allegheny County.

The research team then accessed de-identified data on thousands of families who enrolled in the Child First program from 2017 to 2020 and used this information to "train" and "test" a series of predictive models. The team trained models to predict the outcome of interest on one set of data and then tested those models on a set of unseen data to evaluate how well they correctly predicted whether families disengaged early from the program or not. Because these methods were particularly novel in the field of home visiting, the research team's work aimed to understand whether predictive analytics could reasonably be applied to a program like Child First and strengthen program operations. In line with these goals, the team sought to answer the following questions:

1. To what extent could Child First intake data be used to reliably predict family disengagement?

2. Could a well-specified regression model predict family disengagement as well as a model developed with machine learning?

KEY FINDINGS AND CONCLUSIONS

The next section previews the research team's key findings and conclusions on how the results can be used to support Child First and better serve families who are at risk of early disengagement. The remainder of the brief presents further detail on the sample of families, data, and methodology used to come to these conclusions, exploring the added value of predictive analytics for a home visiting program like Child First.

- **It was possible to use Child First intake data to build predictive analytic models that identified early family disengagement from the Child First program better than by random chance.** However, there was always some amount of unacceptable error when making predictions, and models could not be adjusted to substantially reduce or eliminate these errors. It was difficult to justify the number of families who might be affected by misclassification errors and could be excluded from extra intervention when needed.

- **Using more complex methods and including larger numbers of measures did not result in better predictions of families' early disengagement.** A simple regression model with a selected set of variables—which were either programmatically important or had been shown to be correlated with the outcome—was able to predict early disengagement as well as the machine learning models using the same predictors or all available intake data. The more advanced and complicated methods did not substantially improve the ability to make predictions. In other words, machine learning models added little value beyond what could already be identified through relatively simple and more transparent methods.

- **The research team did not identify a predictive analytics model that could forecast early family disengagement with sufficient reliability and so did not recommend integration into the Child First program.** The team presented a framework for evaluating how well predictive analytics worked in this context, drawing conclusions that caution against assuming that newer and more sophisticated techniques will always perform better than simpler methods. These conclusions may be a helpful guide for future researchers, as more programs seek to implement cutting-edge analytic tools.

The remainder of the brief provides further detail on the sample of families, data, and methodology used to come to these conclusions, exploring the added value of predictive analytics for a home visiting program like Child First.

THE CHILD FIRST SAMPLE

Child First provided the research team with de-identified data on 3,750 families who enrolled in the program in three states—North Carolina, Connecticut, and Florida—from July 2017 through September 2020. All families included in the dataset had either been formally discharged from the

program or had been enrolled for more than 90 days at the time of the data extraction; this made it possible to determine whether or not families had disengaged early from the program. Twenty-two percent of the families disengaged early from the program.

The research team partitioned information on families who had already gone through the program, mimicking the predictive analytics process of using historical data to create predictions that could be applied to new individuals. More specifically, the team trained models using one set of data and then tested model predictions in another set of unseen data, thus evaluating how well these models could predict known outcomes. The team used the data on families enrolled in 2017 and 2018 (N = 2,099) to develop the predictive models (the training sample) and then used the data on families enrolled in 2019 (N = 1,154) to test how well the models could predict outcomes in a "future" cohort (the testing sample). Because the team conducted the analysis in late 2020, it was known whether or not the families enrolled in 2019 experienced early disengagement.

The team considered families enrolled in 2020 as a separate COVID-19 sample (N = 497) and excluded them from the testing sample because of the disruption the pandemic caused to Child First clients, services, and data collection. Notably, although the 2020 testing results are not described in this brief, the findings were similar to the 2019 testing results.

MAKING AND EVALUATING PREDICTIONS

As an initial introduction to predictive analytics and how to evaluate predictions, consider the 2X2 confusion matrix shown in Figure 1. It illustrates using a single characteristic, in this case, the caregiver's relationship to the child, to predict early disengagement from Child First services. The rows indicate the known engagement outcome of the families, the columns indicate the predicted outcome based on caregiver relationship to child (defined here as whether the caregiver is the child's birth mother or not), and counts of sample members are provided in parentheses. In a hypothetical scenario, the program might plan to use information about the caregiver's relationship to the child to identify which families to provide with additional support to prevent early disengagement. For example, newly enrolled families in which the caregiver is not the birth mother might receive more home visits, follow-up calls, and incentives from staff members compared with families where the caregiver is the birth mother.

Cell A contains families who disengaged from the program early and were predicted to do so, also known as **true positives**. Cell D contains families who did not disengage early and were predicted *not* to disengage from the program early, also known as **true negatives**. Here, "true" refers to the prediction being correct and "positive" or "negative" refers to the predicted status, which is the same as the known outcome. Thus Cells A and D contain accurate predictions, and an ideal model would maximize the number of cases that fall into these cells.

FIGURE 1

EXAMPLE CONFUSION MATRIX

Known Outcome	Predicted Outcome by Caregiver Relationship: Early Disengagement	Predicted Outcome by Caregiver Relationship: No Early Disengagement
Early disengagement	**A** (295)	**B** (388)
No early disengagement	**C** (1099)	**D** (1968)

SOURCE: Adapted from Child First intake data.

NOTE: Numbers in parenthesis are counts of families. Cell A contains true positive cases, Cell B contains false negative cases, Cell C contains false positive cases, and Cell D contains true negative cases. N=3,750.

Cell B contains participants who disengaged from the program early but were not predicted to do so. These are **false negatives**, or families who were not identified as needing intervention to help avoid early disengagement but could have potentially benefitted from it. Cell C contains families who did not disengage early but were predicted to disengage early. These participants are **false positives**, or those who would have received additional services but did not need them. Here, "false" refers to the prediction being wrong, and again, "positive" or "negative" refers to predicted status. But Cells B and C are misidentified cases, where the prediction did not match with the known outcome. A good predictor or a good model will have minimal numbers of cases in cells B and C.

The overall **accuracy** (defined in Table 1) of caregiver relationship to the child in this hypothetical scenario is 0.60, which would imply that it is a fairly accurate predictor. However, accuracy considers both the true positive *and* true negatives, and because the outcome is skewed, with only about 20 percent of families disengaging early from services, the large proportion of correctly predicted negative cases is driving the accuracy score.

To address this limitation, it is also important to assess predictive performance using **precision** and **sensitivity**, also defined in Table 1. These measures capture how many positive cases were correctly predicted while also considering misclassified cases. Either high precision or high sensitivity alone is not sufficient. For example, a system could reach 100 percent sensitivity by identifying *all* families

TABLE 1

CLASSIFICATION METRICS IN PREDICTIVE ANALYTICS

KEY TERM	DEFINITION
Accuracy	• Accuracy captures the proportion of correct predictions out of the total number of cases examined. • In Figure 1, accuracy is equal to (A+D)/(A+B+C+D).
Precision	• Precision captures the proportion of correct predictions that are predicted to be positive. Higher precision means fewer false positives. • In Figure 1, precision is equal to A/(A+C).
Sensitivity	• Sensitivity captures the proportion of correct predictions out of all cases that are positive. Higher sensitivity means fewer false negatives. • In Figure 1, sensitivity is equal to A/(A+B).

as needing extra services. But this would not result in precise or useful predictions, as it would be expensive and likely unnecessary to provide all families with additional supports. Thus, precision and sensitivity should be reviewed together to determine how well a predictor performs. In this example, caregiver relationship to child as a predictor of early disengagement alone has a precision of 0.21, and a sensitivity of 0.43, both indicating more misclassified cases than correctly identified positive cases.

Another way to think about the low precision in this example is that, for every five families predicted to disengage from the program, only *one* was correctly identified as being at risk for this outcome using the information assessed at intake. Under the hypothetical scenario, allocating program resources to target families for extra supports based on this predictor alone would have resulted in substantial costs, both in dollars and staff time, with arguably low return. The low sensitivity of the predictor also indicates that more than half of the families who were at risk of early disengagement were not correctly identified. These are cases that did not receive additional intervention but would have benefited from such. These metrics suggest that using only caregiver relationship to child to make predictions would not help the Child First program.

This simple example demonstrates how outcomes could be predicted based on a single characteristic assessed at program intake. However, in many cases the program cannot reliably assess a family using only a single attribute, because no single attribute can reflect a family's complex set of background characteristics. A common practice is to base predictions on multiple attributes, but selecting, combining, and weighting those attributes is a challenge. For example, researchers could combine subject-matter knowledge and simple modeling methods. Or they could harness more advanced techniques using a machine learning algorithm (a method of data analysis that automates analytical

model building) to make predictions. Considering innovative modeling approaches and comparing how well they predict families' outcomes could strengthen home visiting programs.

METHODS AND ANALYSIS

Incorporating more pieces of information can theoretically improve risk predictions by capturing variation in the unique set of characteristics that fit different participants. The research team considered two sets of predictors to assess the utility of predictive analytics for the Child First program if the staff were to predict early disengagement by relying on data collected during the intake process. Table 2 describes these groups of attributes.

The first group includes the predictors that were "programmatically important," meaning they were already being used by Child First to assess general risk for a range of outcomes when families first enroll. This group also includes six measures that had a theoretical justification for being potential predictors as well as a correlation of 0.10 or higher with early family disengagement. The team was

TABLE 2

CHILD FIRST PREDICTOR SETS BASED ON INTAKE DATA

NAME	DESCRIPTION		
Programmatically important	• Child age • Child abuse • Child depression/anxiety • Child development • Child behavior • Child experience of physical abuse	• Caregiver is birth mother • Caregiver emotional well-being • Caregiver is a single parent • Caregiver medical problems • Caregiver substance abuse • Caregiver childhood trauma	• Family has low income • Family is in North Carolina • Family economic stability • Family housing stability • Family history of incarceration • Family child welfare involvement
Kitchen sink[a]	• Programmatically important variables • Child race and ethnicity • Other child developmental diagnoses • Child languages spoken • Child insurance status • Child usual place of medical care	• Caregiver education level • Other caregiver relationships with the focal child • Caregiver race and ethnicity • Other caregiver relationship statuses • Caregiver insurance status • Caregiver languages spoken • Caregiver prenatal health history	• Family is in Connecticut or Florida • Family income sources • Family assistance receipt

SOURCE: Based on Child First intake data.

NOTE: [a]Kitchen sink includes programmatically important variables plus all other items available at intake, as listed here.

cautious when considering variables like race and ethnicity, aware that biases could be introduced into the models if included. The team examined a set of predictors that did include race and ethnicity, but this predictor set was not found to improve predictions beyond the programmatically important list so the results are not highlighted in this brief.

The second group of predictors is described as the "kitchen sink." This group includes the programmatically important predictors *and* the rest of the information collected at intake when families initially enroll in Child First. In total, the Child First program shared over one hundred variables with the team to use for this work.

To synthesize data on the families, the team considered a series of modeling approaches: regression, random forest, and support vector machine with radial basis function (SVM-RBF).[7] Through modeling, information from a large set of predictors could be distilled into one risk estimate for each family. The research team wanted to explore what insights machine learning models could provide above and beyond regression. Regression is widely used and is also relatively easy to understand. However, more complex machine learning methods such as random forest and SVM-RBF could take advantage of the rich data collected by Child First to synthesize a risk score from a large number of predictors.

Below are three increasingly complex approaches that the research team evaluated on predictive performance with respect to identifying families at risk for early disengagement from Child First:[8]

1. **SIMPLE APPROACH:** Programmatically important indicators combined with simple regression
2. **HYBRID APPROACH:** Using the machine learning algorithm random forest to combine and weight programmatically important indicators
3. **ADVANCED APPROACH:** Using the machine learning algorithm SVM-RBF to combine and weight the full set of measures available (often referred to as the "kitchen sink")

The team started with many combinations of measures and modeling methods, eventually narrowing its discussion to the three described above. Of the machine learning methods examined, the random forest algorithm performed best with the programmatically important indicators, while the SVM-RBF algorithm performed best with the "kitchen sink." Because this is a proof-of-concept exercise, the team wanted to track the performance of more than just one "best" model.

By comparing these three models, the team could examine the potential tradeoffs between improvements in predictive performance and loss of simplicity and transparency. For example, in this work, the first model is a regression that includes the programmatically important set of predictors. This model is the simplest and most straightforward, both in terms of number of variables in the predictor set and the modeling method. Though it uses the same programmatically important set of variables, the second model is more complex in that it applies random forest, a machine learning method, to

make predictions. This model provides important insight into the added value of machine learning even when the same predictor set is used. The third model is the most advanced and not only includes *all* available characteristics measured at intake but also presents results from the highest performing machine learning method with these predictors, SVM-RBF. This was the best machine learning model out of all different combinations of covariate sets and algorithms considered, many of which are not discussed here.

The team optimized each of the machine learning algorithm's tuning parameters using cross-validation and moved forward with models with the highest Area Under the Curve-Receiver Operating Characteristics metric (AUC-ROC). AUC-ROC is a general measure of how well a model can predict the outcome that summarizes the overall success of a model. Scores range from 0 to 1, and a score higher than 0.5 indicates that the model can make predictions better than the flip of a coin. This brief does not focus on the interpretation of the AUC-ROC beyond the fact that the team used it to determine which models to move forward with in the analysis. Table 3 shows the scores in the training and testing data.

TABLE 3

CHILD FIRST PREDICTIVE MODEL AUC-ROC SCORES

	2017-2018 TRAINING DATA		2019 TESTING DATA
MODEL	**Mean AUC-ROC**	**SD AUC-ROC**	**AUC-ROC**
1 Programmatically important indicators + regression	0.802	0.027	0.818
2 Programmatically important indicators + random forest	0.794	0.025	0.807
3 Kitchen sink + SVM RBF	0.821	0.028	0.795

SOURCE: Calculations based on Child First intake data.

NOTES: SD = Standard deviation, AUC-ROC = area under the curve-receiver operating characteristics. SVM-RBF = support vector machine with radial basis function.

Mean and standard deviation are provided for the training data AUC-ROC because these are summary statistics based on the five-fold validation performed on the sample. In the testing data, a single AUC-ROC score is calculated for each model.

All three models performed well and were comparable to one another on this metric. The results were also consistent between the training and testing data. However, the AUC-ROC measure is limited in that it is sometimes considered difficult to interpret and does not provide any information about the distributions of predictions. Accordingly, this brief revisits precision and sensitivity in the testing data to comprehensively evaluate the actual performance of the model.

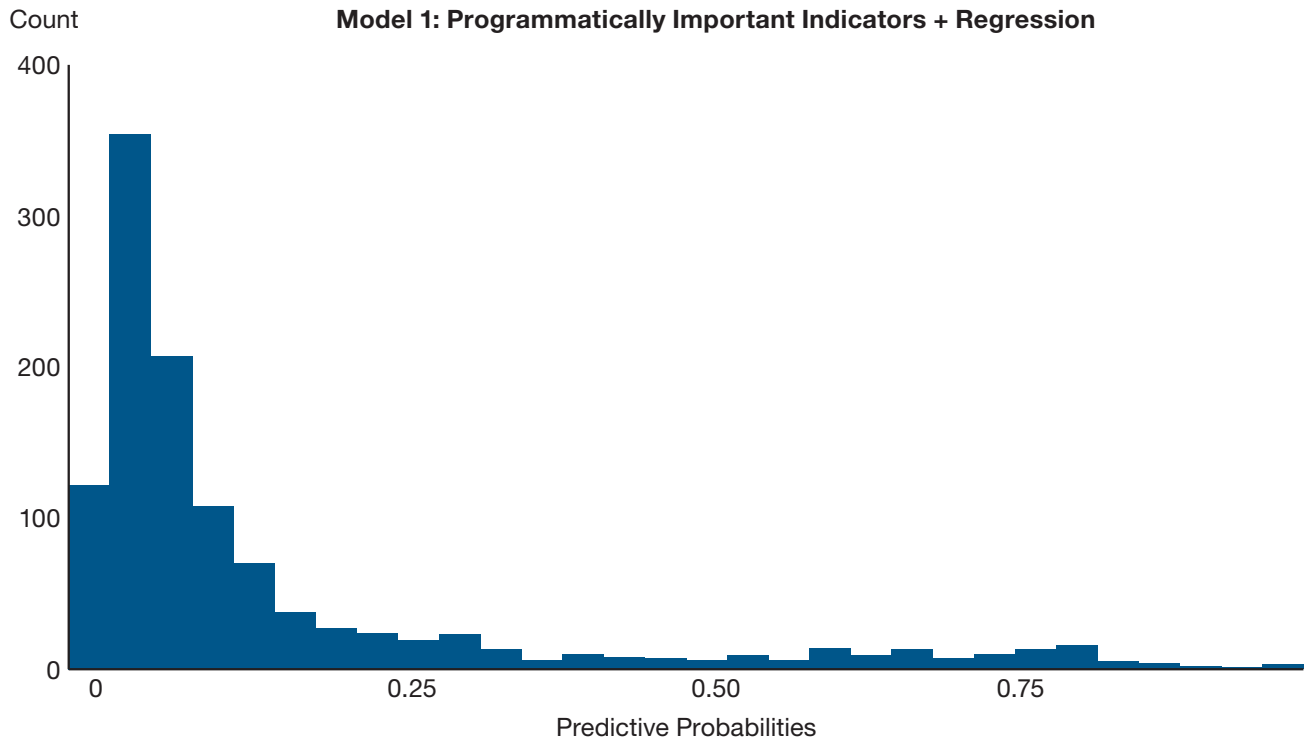
Outcome Probabilities

Rather than relying on a single characteristic, as in the simple example above, predictive models assign each family a continuous risk score or outcome probability, which falls somewhere between 0 and 1. Figure 2 presents a histogram that shows the outcome probability distributions for the simple model examined in the testing data, though all three approaches show a similar distribution.

Within a given model, the outcome probabilities or risk scores stay the same, but the classification of cases can be altered by changing the threshold used to identify which cases will be considered "at risk." A stringent threshold of 0.75 means that only those families with the highest scores are predicted as at risk, while a lower threshold of 0.50 means those with moderate to high scores are predicted as at risk. A commonly used starting point of 0.50, which may be used if there is no reason to believe the outcome is skewed, means that all families with a risk score higher than 0.50 will be predicted as at risk for early disengagement and all families below that threshold as not at risk.

FIGURE 2

DISTRIBUTION OF PREDICTIVE PROBABILITIES



SOURCE: Calculations based on Child First intake data.

Another way to view thresholds is to examine them by considering the overall mean likelihood that the event will occur.

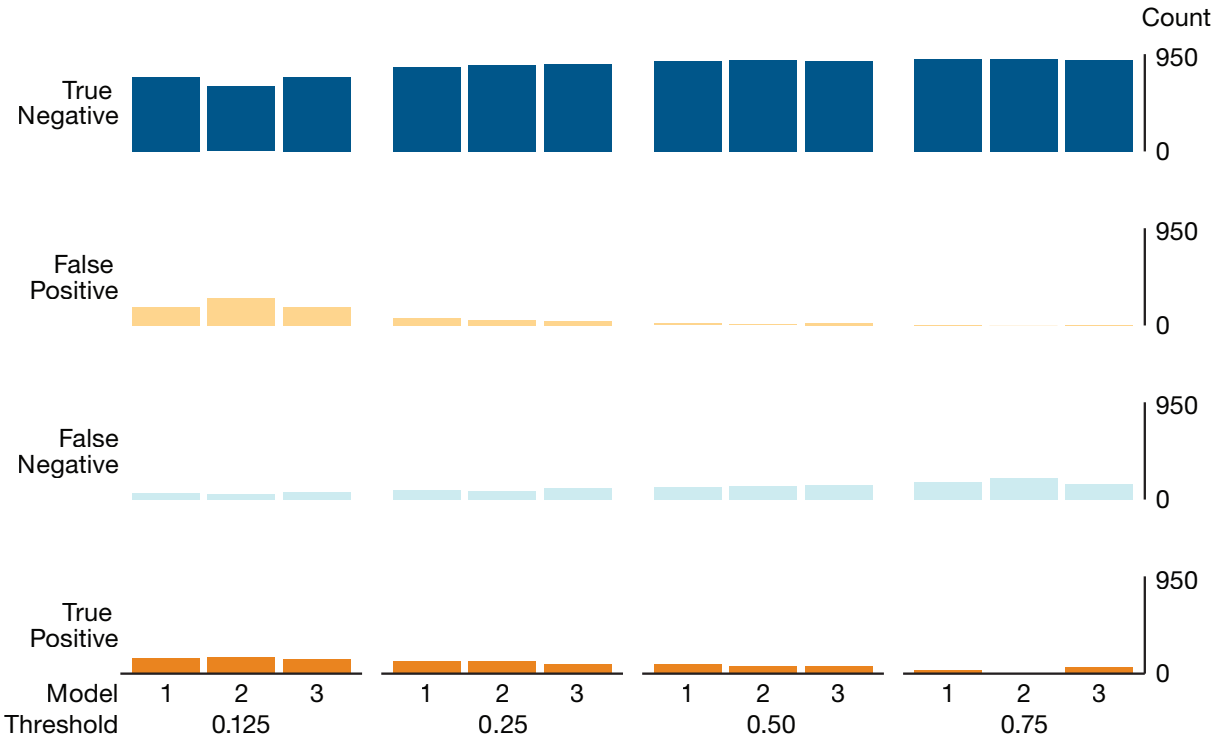
As seen in the histograms, the outcome probabilities distributions are notably skewed to the left. This brief will examine results at 0.125, 0.25, 0.50, and 0.75 to better understand how changing which families are considered at risk can alter model predictions.

Confusion Matrix

Model performance may be optimized by moving the thresholds at which participants are predicted to be positive on an outcome. Applying four different thresholds to the model creates four different scenarios where participants are predicted to be positive or negative on the outcome. This results in different confusion matrices for each model at each threshold. Figure 3 shows the counts of true positive, false positive, false negative, and true negative cases at four different thresholds across the three models in the testing data.

FIGURE 3

CONFUSION MATRIX FOR 2019 TESTING DATA



SOURCE: Calculations based on Child First intake data.

This figure illustrates three trends. First, when examined collectively, the three models perform very similarly at each threshold. That is, there is very little difference at each threshold, and the models seem to have a similar predictive ability. Second, the models successfully predict cases not at risk of early disengagement (true negatives). Finally, there is a noticeable amount of misclassification (false negatives and false positives) across the models at each threshold. At the lowest threshold the highest number of true positive cases are captured. However, this threshold would also catch just as many false positive cases, if not more. At the higher thresholds, true positives are being captured without the error of capturing many false positive cases, but many cases are also going undetected, as seen in the high number of false negatives.

As mentioned earlier, however, count information alone does not capture different tradeoffs in predictions, like serving the most people in need versus minimizing program costs. Focusing on metrics like precision and sensitivity allows for maximizing truly at-risk cases, while minimizing false alarms.

Precision and Sensitivity

Table 4 presents the accuracy, precision, and sensitivity for each model at the same four thresholds shown above. Significantly, there is little difference between classification metrics for the three models, and the models behave similarly at each threshold. Accuracy is high across all the models and thresholds, a result of the skewed nature of the outcome, and the models predicted the families not at risk of early disengagement very well. However, this metric does not specifically identify families the program would want to target for intervention. This is why it is important to consider other measures of successful model classification, such as precision and sensitivity. These measures emphasize capturing true positive cases, which the team wanted to maximize. At the lowest threshold the sensitivity of the models is higher than the precision. As the thresholds increase, that relationship inverts. In an ideal model, both precision and sensitivity would be high. As a result, the team concluded that none of the models performed exceptionally well.

The high sensitivity at the 0.125 threshold indicates that the models correctly identify the most families at risk of early disengagement, but the low precision indicates that the models also identify many families that are not at risk. This is because at a lower threshold it is more likely that cases will be identified as positive. However, in that process, many false positive cases will also be captured. Examining the counts for Model 1—the regression with programmatically important predictors—155 families are correctly identified as true positive cases, while 184 families are false positive cases and would have potentially received enhanced services unnecessarily. Sixty-four families are false negative cases and would not receive added interventions to keep them engaged. Key to these findings is that more false positives are being identified than true positives, so using predictions from this model would lead to inefficient use of limited resources.

Precision and sensitivity at the 0.25 threshold may seem to be acceptable, with both scores for the most part above 0.50, but misclassification still occurs. Take Model 2, the random forest model with

TABLE 4

CHILD FIRST ACCURACY, PRECISION, AND SENSITIVITY OF MODELS USING 2019 TESTING DATA

THRESHOLD	MODEL	ACCURACY	PRECISION	SENSITIVITY
0.125	**1** Programmatically important indicators + regression	0.79	0.46	0.71
	2 Programmatically important indicators + random forest	0.72	0.38	0.76
	3 Kitchen sink + SVM-RBF	0.77	0.43	0.65
0.25	**1** Programmatically important indicators + regression	0.85	0.62	0.58
	2 Programmatically important indicators + random forest	0.87	0.66	0.61
	3 Kitchen sink + SVM-RBF	0.85	0.66	0.47
0.50	**1** Programmatically important indicators + regression	0.87	0.82	0.43
	2 Programmatically important indicators + random forest	0.87	0.93	0.32
	3 Kitchen sink + SVM-RBF	0.86	0.77	0.35
0.75	**1** Programmatically important indicators + regression	0.84	0.95	0.18
	2 Programmatically important indicators + random forest	0.81	—	0.00
	3 Kitchen sink + SVM-RBF	0.86	0.88	0.30

SOURCE: Calculations based on Child First intake data.

NOTES: SVM-RBF = support vector machine with radial basis function.

The classification of cases can be altered by changing the threshold used to identify which cases will be considered "at risk." This then changes the calculations of accuracy, precision and sensitivity, as presented in this table.

Precision cannot be calculated for Model 2 at the 0.75 threshold because at that threshold, the model predicted no cases as being at risk of early discharge.

programmatically important variables: 133 individuals are correctly identified as at risk of early discharge, and there are 68 false positive cases and 86 false negative cases. This exceeds a minimal amount of misclassification. The model is not doing a good job of pinpointing families in need of intervention and would not direct services in a cost-effective way.

At the higher threshold of 0.50, precision increases, indicating that most cases predicted to be positive are in fact positive. However, the low sensitivity indicates fewer at-risk families are being predicted as such, and many families that are at risk are missed by the models at this higher threshold. Neither outcome is ideal for serving families in Child First. Take Model 3, the SVM-RBF model with all predictors: 76 cases are correctly identified as being at risk, 166 cases are misclassified, and 143 cases are false negatives, or those that would have benefited from enhanced services but were not predicted to be at risk and so missed out. As indicated by the sensitivity score, only a third of the at-risk cases are being correctly predicted as early disengagement by the model.

At the threshold of 0.75, the models have the highest precision but also the lowest sensitivity. This is because very few cases are being predicted as positive on the outcome, exemplifying what happens when too strict a threshold is applied. Looking at Model 1 again, at this high threshold only 40 cases

are correctly identified as at risk of early disengagement (out of a total of only 42 cases predicted to be at risk for the outcome) and 179 cases are misclassified as false negatives.

While the models identify better than the flip of a coin those families who are at risk of early disengagement, the research team struggled to agree upon a best model or ideal threshold. The advanced methods models did not perform better than the simple regression, and arguably, none of the models at any of the four thresholds accomplishes the goals of pinpointing families in need of intervention and directing the provision of services in a cost-effective way.

DISCUSSION

In the analyses presented in this brief, the team set out to determine whether it was possible to reliably predict early family disengagement from the Child First program by relying on measures the program collects during the intake process, to evaluate the performance of different predictive analytic methods. The team also aimed to identify whether predictive analytics techniques could be useful to staff members serving families in Child First. While it was possible to create predictive models that performed better than random chance with the data collected by Child First at intake, these models made numerous mistakes. The models were better at correctly predicting the true negatives in the sample (families who were predicted and confirmed not to disengage early) because of the skewed nature of the outcome (less than a quarter of families disengaged from the program early). However, the models could not correctly identify the true positives in the sample (families who were predicted and confirmed to disengage early) while also minimizing misclassified cases. Put another way, the models were unable to sufficiently pinpoint the families at risk while limiting extra expenditures on families not at risk.

Importantly, this exercise found that the three models performed similarly across the performance metrics. There were no substantial differences or gains to justify selecting a machine learning–based model over a more transparent regression model. Based on the results, the team concluded that advanced techniques were not necessary to make predictions. In fact, the team concluded that machine learning was not worthwhile in this context, given the increased complexity of risk scores produced by these models. Instead, using *simple regression with selected variables* could provide Child First with some helpful information about which families are likely to disengage from the program early and thus might benefit from extra support. But incorporating the regression model into the program would have to be done very carefully, since these predictions were not always reliable.

Each model exhibited some amount of misclassification at the thresholds examined, and whether or at what level that is acceptable depends on the context and the end user's desired outcome. Whether a program is willing to accept that some number of families may be served unnecessarily (which will likely have financial implications for programs with limited resources), or that some number of families may miss out on needed enhanced services will depend on risk tolerance and program priorities. Child First's current process for identifying families at risk of early disengagement needs

to be assessed more thoroughly to determine what level of misclassification is acceptable. It is likely that misclassification of families already occurs on some level during the overall assessment of risk for early disengagement when families first enter the Child First program. But it is also possible that home visitors to some extent are able to identify families at risk of early disengagement and intervene accordingly. It is unclear at this time whether additional insights from predictive analytics could improve services and reduce early disengagement above and beyond current practices. Given the unreliable performance of the models, the team would not suggest that an intervention or any decision-making rest *solely* on the results of the predictive models identified here. Rather, predictive analytics could serve as one tool—in combination with other sources of information—to help Child First staff make decisions about service provision.

It was not possible to assess how these models would compare with predictions made by individual staff members about families at risk of early disengagement. Fully understanding how staff members identify families as at risk would require an in-depth study and was outside the scope of this brief. Staff members might use a combination of knowledge, experience, and information gathered when they first meet with families to identify whether participants are likely to disengage from services early. This hands-on approach may have advantages but may also result in unintentional biases that could affect staff judgement. However, the team would need systematic information from staff on which families they thought could be at risk for early disengagement to compare the reliability of the two approaches. And the team would recommend performing a bias analysis prior to any real-world application of predictive models.

It is possible that there were not enough data available to make high quality predictions, or that the team selected an outcome at a time when not enough information was available to make good predictions. The analysis was based on the intake data collected by Child First and used around one hundred variables. It may be possible to predict a different outcome in the Child First program at a different time in the family's participation if more data were to become available. For example, early participation indicators likely are predictive of later participation challenges. However, because early disengagement occurs in such a short window, this particular outcome does not permit much opportunity to incorporate additional data. These results were not entirely surprising given the findings from the Fragile Families Challenge, which had thousands of measures available, but had roughly the same number of families or observations to work with and also resulted in poor predictions.[9] In contrast, the Allegheny Family Screener Tool has shown success predicting an adjacent but different outcome using hundreds of variables across multiple data sources with far more observations (tens of thousands of children).[10]

Finally, the analysis focused primarily on families enrolled in Child First pre-pandemic. The models in this exercise were trained on data from Child First participants in 2017 and 2018 and tested on data from participants in 2019. While the same trends identified above were observed in the data, the ongoing COVID-19 pandemic may continue to change how Child First operates, further reducing the predictive power of the models identified.

CONCLUSIONS AND AREAS FOR FUTURE RESEARCH

Ultimately, the goal of a social service like a home visiting program is to develop trusting bonds with families and provide care to meet their needs. This contrasts with sectors that have successfully turned decision-making over to predictive analytics, like advertising or entertainment, where there are minimal consequences to missed predictions. Misclassifications in these fields may result only in a consumer receiving an irrelevant ad or movie suggestion.

Given the limitations, however, the research team did not want to recommend one of the predictive models discussed here that might introduce additional error into how the program identifies families at risk of early disengagement. The potential consequences of missed predictions in human-based services like home visiting are significant, as support and resources to improve health and well-being could be on the line. On the one hand, over-including families for intervention could result in allocating limited resources or staff time unnecessarily, while under-including families could result in missing those in need of extra support. While some unknown level of misclassification likely occurs now, any increase in misclassifications will have real consequences for the people who are chosen and the people who are left out.

The research team hopes the work presented in this brief serves as a foundation for future studies of predictive analytics in the home visiting field. The team identified major points to consider should a more successful model be developed in the future. Care should be taken to ensure that predictive analytic findings are not used against prospective participants and that biases are not built into the models. As seen in other applications, serious consideration needs to be given to the point of prediction and what the intervention would entail. Additionally, it is important to have different voices weigh in on the integration of predictive models. Good performance metrics alone are thus insufficient to justify application in human-based service programs.

NOTES AND REFERENCES

1 Allegheny County Department of Human Services, "Impact Evaluation Summary of the Allegheny Family Screening Tool" (Pittsburgh: Allegheny County Department of Human Services, 2019).

2 Lindsay Cattell and Julie Bruch, "Identifying Students at Risk Using Prior Performance Versus a Machine Learning Algorithm," September (Washington, DC: Institute of Education Sciences, 2021).

3 Matthew J. Salganik, Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim et al., "Measuring the Predictability of Life Outcomes with Scientific Mass Collaboration," *PNAS* 117, 15 (2020), https://doi.org/10.1073/pnas.1915006117.

4 Mariel Sparr and Susan Zaid, *State-Led Evaluations of Family Engagement: The Maternal, Infant, and Early Childhood Home Visiting Program*, OPRE Report 2017-39 (Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2017).

5 Chris S. Hulleman and David S. Cordray, "Moving from the Lab to the Field: The Role of Fidelity and Achieved Relative Intervention Strength," *Journal of Research on Educational Effectiveness* 2, 1 (2009): 88-110.

6 American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders: DSM-5* (Washington, DC: American Psychiatric Association Publishing, 2013).

Vincent J. Felitti, Robert F. Anda, Dale Nordenberg, David F. Williamson, Alison M. Spitz, Valerie Edwards, Mary P. Koss, James S. Marks, "Relationship of Childhood Abuse and Household Dysfunction to Many of the Leading Causes of Death in Adults: The Adverse Childhood Experiences (ACE) Study," *American Journal of Preventative Medicine* 14, 4 (1998): 245-58.

7 Random forest is a type of machine learning algorithm made up of many decision trees. Random forests can identify nonlinear relationships and interactions between predictors. SVM-RBF is a type of machine learning algorithm using classifiers used to linearly separate data using boundaries using support vectors. Radial basis functions help get better separation when there are not clean breaks by projection to higher dimensions.

8 The team started with a larger number of combinations of measures and machine learning algorithms but narrowed the discussion of this work to those listed to summarize what was learned in the process.

9 Salganik et al. (2020).

10 Allegheny County Department of Human Services (2019).

ACKNOWLEDGMENTS

The authors would like to thank several MDRC colleagues for their contributions to this brief. This includes Brit Henderson and Richard Hendra for their careful review of draft materials and suggestions for improving the brief, Dannia Guzman and Melvin Gutierrez for lending their expertise to developing the figures, Maya Goldberg for assisting with report coordination, Joshua Malbin for reviewing the brief, Jill Kirschenbaum for editing it, and Carolyn Thomas for preparing it for publication. Importantly, the work reported on in this brief was done in collaboration with leadership at Child First, including Darcy Lowell, Kim DiBella-Farber, Rebecca Parilla, Salam Soliman, Dana Hillman-Sabato and Diane Britz.

The Child First predictive analytics work was made possible through funding from The Edna McConnell Clark Foundation and The Annie E. Casey Foundation. Arnold Ventures and The Duke Endowment provided anchor funding for the broader Child First study.

Dissemination of MDRC publications is supported by the following organizations and individuals that help finance MDRC's public policy outreach and expanding efforts to communicate the results and implications of our work to policymakers, practitioners, and others: The Annie E. Casey Foundation, Arnold Ventures, Charles and Lynn Schusterman Family Foundation, The Edna McConnell Clark Foundation, Ford Foundation, The George Gund Foundation, Daniel and Corinne Goldman, The Harry and Jeanette Weinberg Foundation, Inc., The JPB Foundation, The Joyce Foundation, The Kresge Foundation, and Sandler Foundation.

In addition, earnings from the MDRC Endowment help sustain our dissemination efforts. Contributors to the MDRC Endowment include Alcoa Foundation, The Ambrose Monell Foundation, Anheuser-Busch Foundation, Bristol-Myers Squibb Foundation, Charles Stewart Mott Foundation, Ford Foundation, The George Gund Foundation, The Grable Foundation, The Lizabeth and Frank Newman Charitable Foundation, The New York Times Company Foundation, Jan Nicholson, Paul H. O'Neill Charitable Foundation, John S. Reed, Sandler Foundation, and The Stupski Family Fund, as well as other individual contributors.

The findings and conclusions in this report do not necessarily represent the official positions or policies of the funders.

For information about MDRC and copies of our publications, see our website: www.mdrc.org.

Copyright © 2022 by MDRC®. All rights reserved.

NEW YORK
200 Vesey Street, 23rd Flr., New York, NY 10281
Tel: 212 532 3200

OAKLAND
475 14th Street, Suite 750, Oakland, CA 94612
Tel: 510 663 6372

WASHINGTON, DC
750 17th Street, NW, Suite 501
Washington, DC 20006

LOS ANGELES
11965 Venice Boulevard, Suite 402
Los Angeles, CA 90066