What Works *for* Children's Social Care

MACHINE LEARNING IN CHILDREN'S SERVICES

SUMMARY REPORT

Vicky Clayton
Michael Sanders
Eva Schoenwald
Lee Surkis
Daniel Gibbons

SEPTEMBER 2020

ACKNOWLEDGEMENTS

We are grateful to data teams at the four local authority partners, to the participants in a series of events over the last eighteen months which have informed our thinking, Michael Yeomans from Harvard's Kennedy School of Government, and anonymous peer reviewers. We are also grateful to colleagues at the Department for Education.

What Works *for* Children's Social Care

MACHINE LEARNING IN CHILDREN'S SERVICES DOES IT WORK?

CONTENTS

FOREWORD
ANNE LONGFIELD, CHILDREN'S COMISSIONER FOR ENGLAND

Ensuring that all vulnerable or disadvantaged children get the right help at the right time has been one of the cornerstones of my agenda as Children's Commissioner. It is why I have constantly drawn attention to the many children with unidentified or unmet needs, falling through the gaps and under the radar of services, until a crisis hits and the state has no choice but to step in.

MY OFFICE HAS FOUND THAT THERE ARE 2.3 MILLION CHILDREN IN ENGLAND GROWING UP WITH A VULNERABLE FAMILY BACKGROUND – FAR BIGGER THAN THE NUMBER OF CHILDREN BEING SUPPORTED BY CHILDREN'S SOCIAL CARE AT ANY ONE TIME

My office has found that there are 2.3 million children in England growing up with a vulnerable family background – far bigger than the number of children being supported by children's social care at any one time. Around a third of these children are on the radar of local services, but what, if any, support they get is unclear. Another third are not even known to services – effectively 'invisible.'

Not all of these children will need a social worker or a child protection plan. Many of them might just need a helping hand, a trusted adult, and a stable source of support when times are difficult. But far too many children who need help risk being missed completely, until it is too late. In some cases, the result is a serious case review that highlights the all-too-common issues of 'failing to spot the signs', 'cases being deemed low-risk at the time', or 'information not being shared.' In many other cases, the result will be a child or young person who never quite reaches their potential, carrying the failures of a system that did not give them what they needed to thrive.

That is why I welcomed the creation of What Works for Children's Social Care. There is so much to learn and embed in front-line practice about how we can intervene at crucial points to improve outcomes for children who need help and support, and evidence is right at the heart of it. I firmly believe that innovative uses of data – be they better analysis, sharing or recording – can unlock considerable benefits, helping local agencies make better and more effective decisions.

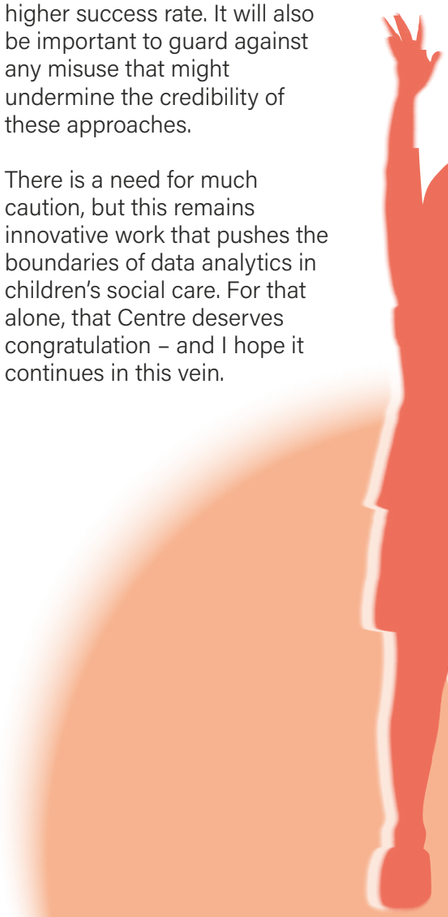
Better prediction of risk can help front-line professionals to quickly assess relatively straightforward cases, and spend more of their valuable time on the complex and nuanced cases where a statistical model is no match for a human. Nobody is suggesting that data science or algorithms can ever replace professional judgement. But what they can do is provide objective evidence, at scale, that makes professional judgement easier.

We will have all seen, in the context of this year's A Level results, the issues caused by so-called 'decision-making by algorithm.' It is an important warning that we must all heed. But it is not a reason not to use data science or algorithms; rather, it is a reason to use them carefully, understand their limitations, and test and refine them; while continuing to treat people as individuals.

This report summarises a fascinating project which I have been following closely since its inception. It is refreshingly candid, focussing on the limitations and challenges of applying machine learning methods in this way. Clearly there is more work to do to improve the effectiveness and suitability of these approaches in children's social care, and to establish the circumstances where predictive models have a higher success rate. It will also be important to guard against any misuse that might undermine the credibility of these approaches.

There is a need for much caution, but this remains innovative work that pushes the boundaries of data analytics in children's social care. For that alone, that Centre deserves congratulation – and I hope it continues in this vein.

CLEARLY THERE IS MORE WORK TO DO TO IMPROVE THE EFFECTIVENESS AND SUITABILITY OF THESE APPROACHES IN CHILDREN'S SOCIAL CARE, AND TO ESTABLISH THE CIRCUMSTANCES WHERE PREDICTIVE MODELS HAVE A HIGHER SUCCESS RATE



EXECUTIVE SUMMARY

In this project, we worked with four local authorities to develop models to predict eight outcomes for individual cases. The predictions all focused on a point within the children's journey where the social worker would be making a decision about whether to intervene in a case or not and the level of intervention required, and looked ahead to see whether the case would escalate at a later point in time. We used natural language processing techniques to turn reports and assessments into information that can be used as input to a model. We then used machine learning techniques to learn patterns in historical data associated with risks and protective factors, and examine whether those factors were present in unseen cases. We sought to understand whether machine learning models, applied in this way, correctly identify the cases at risk of the outcome and those that are not and whether they do this equally well for different groups. We also compared four different ways of designing the models.

WE SOUGHT TO UNDERSTAND WHETHER MACHINE LEARNING MODELS, APPLIED IN THIS WAY, CORRECTLY IDENTIFY THE CASES AT RISK OF THE OUTCOME

WE FIND THAT

On average, if the model identifies a child is at risk, it is wrong six out of ten times. The model misses four out of every five children at risk.

None of the models' performances exceeded our pre specified threshold for 'success' .

Adding information extracted from reports and assessments does not improve model performance.

Our analysis of whether the models were biased was unfortunately inconclusive.

There is a low level of acceptance of the use of these techniques in children's social care amongst social workers.

In summary, we do not find evidence that the models we created using machine learning techniques 'work' well in children's social care. In particular, the models built miss a large proportion of children at risk which - were the models to be used in practice - risks discouraging social workers from investigating valid concerns further, potentially putting children and young people at risk. For just over half of the models, adding more cases may improve the model performance; however, using data further back in time is unlikely to help. Machine learning techniques may be more suitable for a different type of outcome which doesn't reflect social worker decisions and which has a higher percentage of the population at risk of the outcome.

We did not seek to definitively answer the question of whether machine learning will ever work in children's social care but we hope to have shown some of the challenges faced when using these approaches in children's social care. For local authorities already piloting machine learning, we encourage them to be transparent about the challenges they experience.

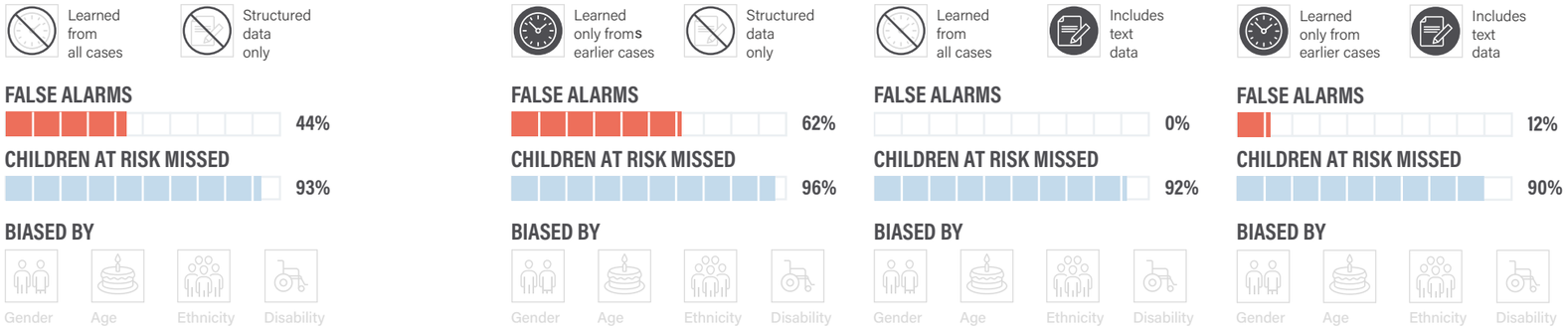
Given these challenges and the extent of the real world impact a recommendation from a predictive model used in practice could have on a family's life, it is of utmost important that we work together as a sector to ensure that these techniques are used responsibly if they are used at all.

WE DID NOT SEEK TO DEFINITIVELY ANSWER THE QUESTION OF WHETHER MACHINE LEARNING WILL EVER WORK IN CHILDREN'S SOCIAL CARE BUT WE HOPE TO HAVE SHOWN SOME OF THE CHALLENGES FACED WHEN USING THESE APPROACHES

PREDICTION ONE:

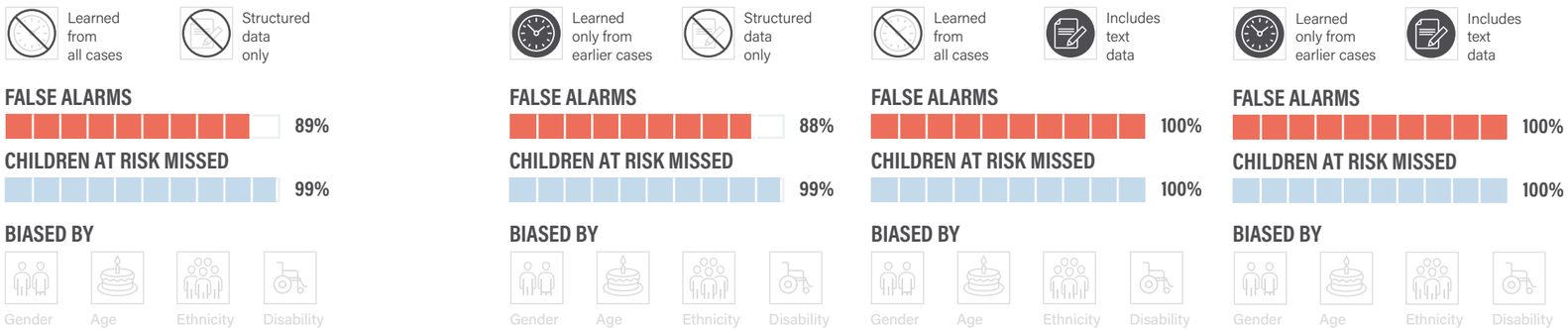
Does a child / young person's case come in as a 're-contact' within 12 months of their case being NFA-ed ('no further action'-ed), and does the case then escalate to the child being on a Child Protection Plan (CPP) or being Looked After (CLA)?

A greyed out icon without a strike-through denotes the unavailability of bias data ▶



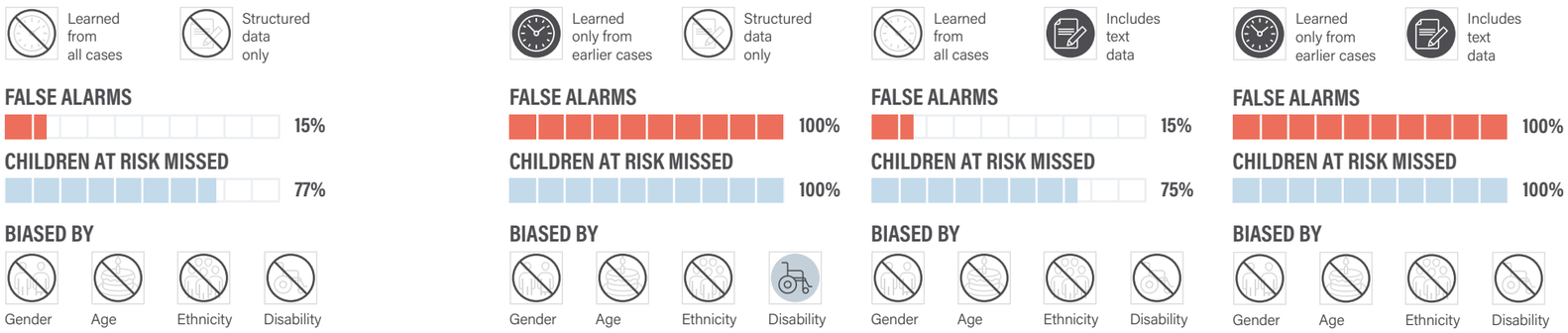
PREDICTION TWO:

Does the child / young person's case progress to the child being subject to a Child Protection Plan (CPP) or being looked after (CLA) within 6-12 months of a contact?



PREDICTION THREE:

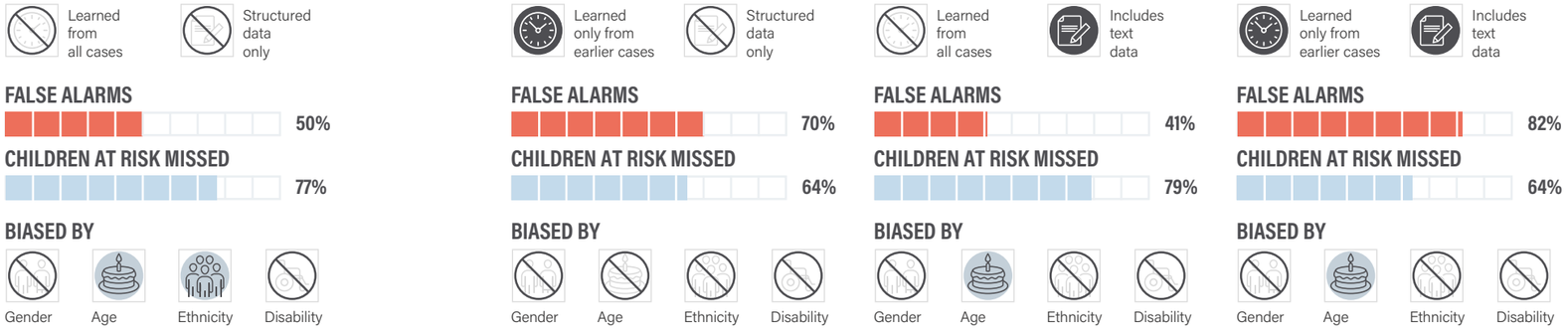
Is the child / young person's case open to children's social care - but the child / young person not subject to a Child Protection Plan (CPP) or being Looked After (CLA) - within 12 months of their case being designated 'No Further Action'?



PREDICTION FOUR:

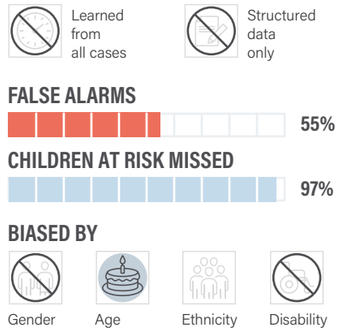
Is the child or young person's case which is already open to children's social care being escalated (to the child being subject to a Child Protection Plan, being Looked After, being adopted, being subject to a Residence Order or being subject to a Special Guardianship Order) between three months and two years of the referral start date?

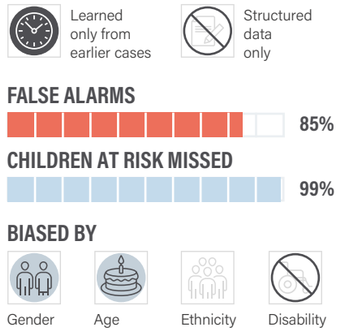
Source: Four local authorities (March 2012 - July 2019). Sample: c. 700 -24,000

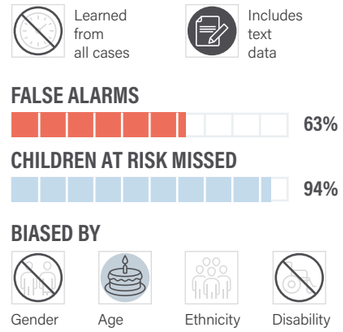


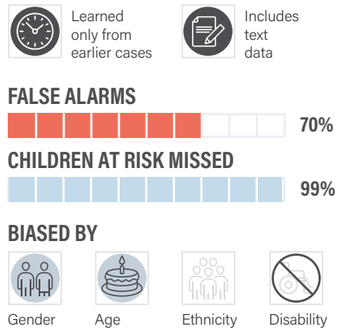
PREDICTION FIVE:

Does the child / young person's case progress to the child being subject to a Child Protection Plan (CPP) or the child being Looked After (CLA) within 6-12 months of a contact?



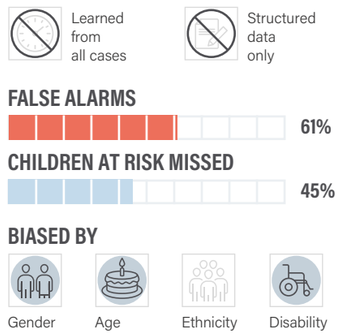


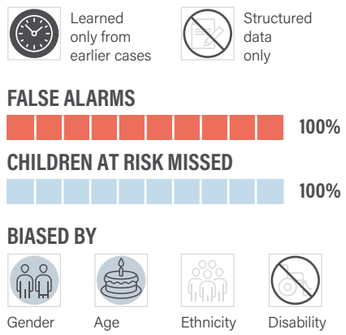


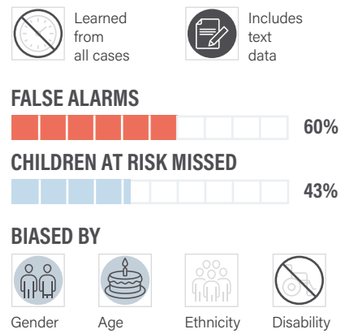


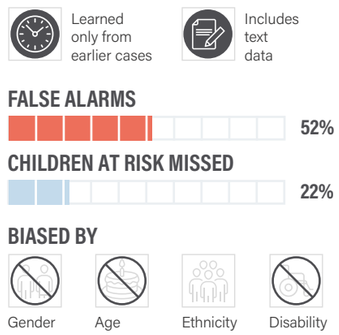
PREDICTION SIX:

After successfully finishing early help, is the child / young person referred to statutory children's services within 12 months?



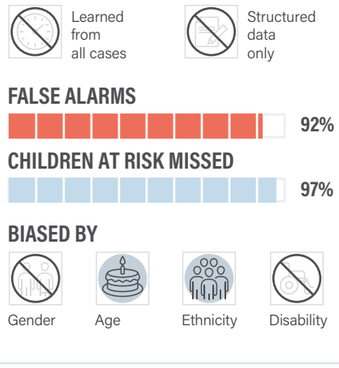


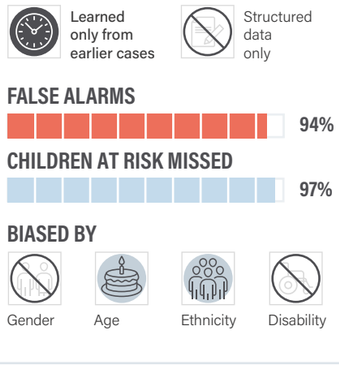


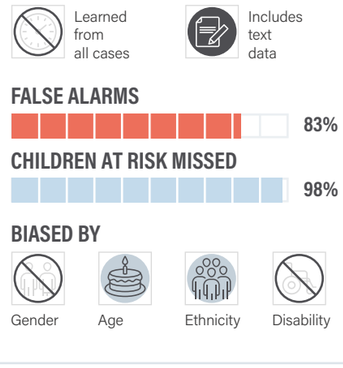


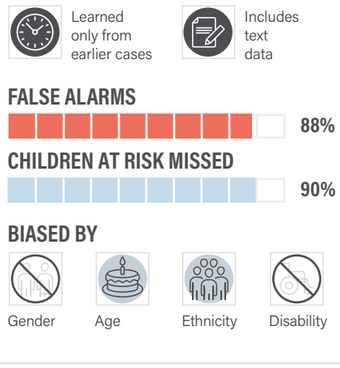
PREDICTION SEVEN:

Does the child / young person's case progress to the child being subject to a Child Protection Plan (CPP) within 1-12 months of the assessment authorisation date?



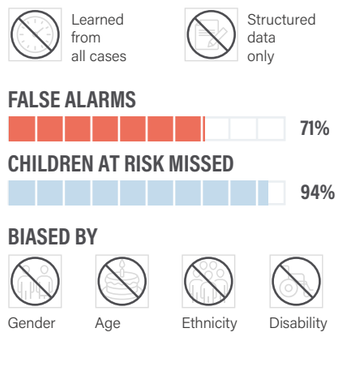


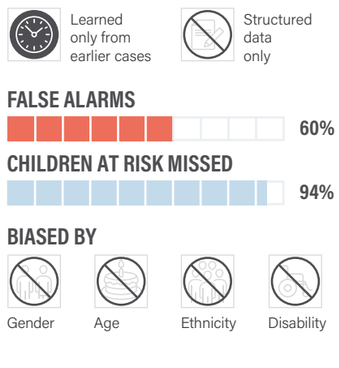


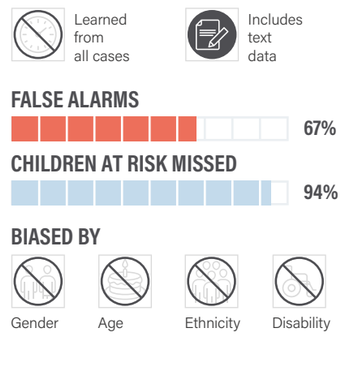


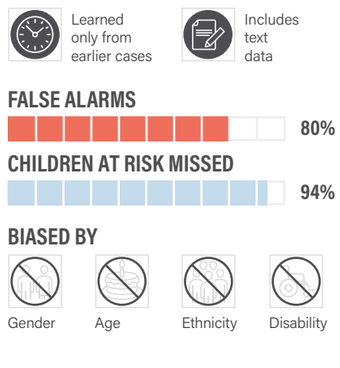
PREDICTION EIGHT:

Does the child / young person progress to the child being Looked After (CLA) within 1-12 months of the assessment authorisation date?









INTRODUCTION

Data is everywhere; from responses to polls and surveys, to our social media activity, to what we buy in our weekly shop. Many of these forms of data are new, or have surged in recent years. By contrast, the data available to social workers is much the same as it always has been; their case notes and records of important historical events, and the case notes of the social workers who worked with a family previously, form the bulk of this.

Even if the form of the data is old, what can be done with it is not. Advances in computing power, and the algorithms used in machine learning, mean that this often huge body of text can be treated as a dataset for analysis, and can be used in a similar way to other forms of data.

The promise of these approaches is obvious to many. By analysing the huge amount of data that exists about a young person and their family, data scientists could predict the likelihood of a case escalating, or a placement breaking down. Armed with this information, social workers and others could better support these families who need it most earlier on.

THE PROMISE OF THESE APPROACHES IS OBVIOUS TO MANY. BY ANALYSING THE HUGE AMOUNT OF DATA THAT EXISTS ABOUT A YOUNG PERSON AND THEIR FAMILY, DATA SCIENTISTS COULD PREDICT THE LIKELIHOOD OF A CASE ESCALATING, OR A PLACEMENT BREAKING DOWN

THE RESULTS OF OUR EFFORTS SHOW THE CHALLENGES OF USING MACHINE LEARNING WITH SUCH LARGE, UNSTRUCTURED DATASETS IN AN ENVIRONMENT THAT IS BOTH AS COMPLEX, AND AS FAST MOVING, AS CHILDREN'S SOCIAL CARE

Using text data - such as that found in case notes - is especially appealing, as it uses data generated by social workers and already held by a local authority. Social workers need to discuss a case and gather information from partner agencies. This information enters their case notes, and replaces an often expensive, time consuming and complex governance process of merging data from across multiple agencies. It is also likely that case notes provide a more nuanced picture of a family's circumstances, and their strengths, than any number of quantitative data points.

This promise is matched on the other side by a substantial risk; that algorithmic outputs will be used as a crutch and an alternative to professional thought and judgement; that inputs to a case management system will become less about recording and reflecting on the facts and as a statement of facts agreed with the family, and more about trying to elicit a response from an algorithmic risk score; that the predictions will be biased along racial, or other, lines; or that the algorithm creates a self fulfilling prophecy, failing to recognise or allow families and children to speak beyond the facts of their past.

Married to this are very real concerns about the ethics and legality of using machine learning in this way without the consent of the people to whom the data relate, and the people to whom it will be applied, as well as about the quality of the data that inputs into any decisions. An additional concern is that these tools will simply tell social workers what they know already, making the development of these tools a waste of public resources. Where the models make wrong predictions, this could lead to unnecessary intervention by social workers or children and families not getting the support that they need.

For the last eighteen months, What Works for Children's Social Care has attempted to provide the sector with a stronger basis on which to make decisions about the use of machine learning in this context.

This report focuses on the important question of the value of these approaches in terms of one of the key tenets of their appeal; their ability to accurately predict future events. Where our previous research has focused on the ethics of using machine learning, here we quantitatively assess whether machine learning achieves its stated goals, and whether it is biased, when using historical data.

We worked with four local authorities from across England to identify how well models performed on two predictions at each local authority. The results of our efforts show the challenges of using machine learning with such large, unstructured datasets in an environment that is both as complex, and as fast moving, as children's social care.

These results may not represent the best that can be done with the available data. While important, this project was modest in its size and its scope. We cannot guarantee that other researchers, in other local authorities, using different methods could not produce better, and more useful models. However, our researchers, with advice and support from experts and academics, have produced these models transparently - from our research protocol, to our code, to the results, all of which we've published alongside the report. We hope that this shows partners in local authorities, and in organisations producing predictive analytics tools, what is possible in terms of this transparency, and the extent to which testing the effectiveness of a model is vital, both at the outset of its use and continuing its use.

ABOUT THE LOCAL AUTHORITIES INVOLVED

The four local authorities ranged in size, from c.200-300 referrals per 10,000 to c.500 referrals per 10,000, and location: they were situated across the North West, South West, West Midlands, and South East. They had Good or Outstanding Ofsted ratings. Unusually for our partnerships, the local authorities have remained anonymous for this project. This allowed the local authorities to participate in an innovative but sensitive project. We shall refer to them as LA1, LA2, LA3 and LA4

WHAT IS MACHINE LEARNING?

Machine learning (ML) is a technique which finds statistical patterns in data. Specifically, we use a subset of machine learning techniques called 'predictive analytics.' In a predictive analytics framework, models learn patterns from historical data about how input data is associated with a particular outcome (for example, are younger children more likely to enter into care?). The models then use these learned patterns to predict the outcome on observations where the input data is known but the outcome is not known. A model is a combination of the input data, the decision rules used to transform the input data into a prediction ('the algorithm') and the parameters used to make adjustments to the rules.

When the model is used in practice, the outcome is not known because it has not happened yet. For example, if the model was predicting whether a child's case escalates, one would have to wait to see whether the child's case escalated over the time period specified to see whether the model was correct or not. This isn't a practical way of testing how well the model works so in order to test how well the model would perform in the real world, some of the available data is separated prior to training the model (when the model learns the patterns), and the model is then tested on this unseen data to simulate how well the model would work in practice.

IS USING MACHINE LEARNING ETHICAL?

What Works for Children's Social Care commissioned the Rees Centre at the University of Oxford and The Alan Turing Institute to undertake a review of the ethics of using machine learning in children's social care. The review made some preliminary recommendations for steering machine learning (ML) in children's social care:

Mandate the responsible design and use of ML models in children's social care at the national level;

Connect practitioners and data scientists across local authorities to improve ML innovation and to advance shared insights in applied data science through openness and communication;

Institutionalise inclusive and consent-based practices for designing, procuring, and implementing ML models;

Fund, initiate, and undertake active research programmes in system, organisation, and participant readiness;

Understand the use of data in children's social care better so that recognition of its potential benefits and limitations can more effectively guide ML innovation practices;

Use data insights to describe, diagnose and analyse the root causes of the need for children's social care, experiment to address them;

Focus on individual- and family-advancing outcomes, strengths-based approaches, and community-guided prospect modelling;

Improve data quality and understanding through professional development and training.

WHAT DOES CHILDREN'S SOCIAL CARE DATA LOOK LIKE?

When conducting an inspection of local authority children's services, Ofsted requests child-level data (the "Annex A" dataset). The data contains demographic information and information about their interactions with children's services. Ofsted provides guidance with standardised column headings and codes, as well as a template for the data requested. This means that the structured data available is relatively standardised, but the definitions of different interactions with children's services may be slightly different by local authority. As thresholds for various levels of intervention are set out by the Children Act 1989, the documentation required for each stage of the child's journey is also relatively standardised although the exact format of each document may vary by local authority.

ABOUT THIS PROJECT

In February 2019, we issued an open call to local authorities to join as partners on this project. This led to partnerships with four local authorities whom we refer to as LA1, LA2, LA3 and LA4. The local authorities provided us with 3-7 years of data extracted from their case management systems: Structured data: usually in the form of the Annex A report which they prepare for Ofsted. Text data: accompanying notes from early help contacts, referrals, assessments, initial and review child protection conference reports and strategy discussion (depending on the outcome being predicted)

ED

MS FOR INDIVIDUALS OF DIFFERENT GENDERS, DISABILITY STATUSES, ETHNICITIES AND AGE GROUPS

This gave us datasets ranging in size between c.700 and c.24,000 cases, with the smallest dataset focusing on an early help context. We completed the analysis for each local authority separately, keeping the datasets separate. We prepared the datasets for the model, validated the data and created new ways of summarising or categorising the data. Processing the text data involved automatic redaction of personal information and turning the documents into tabular data (which words appear, the topics which summarise the documents and other linguistic features). We also extracted from the text 'vulnerabilities' the child was experiencing e.g. food poverty, parental substance abuse (as classified by the Office for the Children's Commissioner[1]) - this painted a richer picture than the structured data alone could. The 'raw' data was processed on the local authority IT systems, and in two cases, the pseudonymised data was then analysed 'offsite' at WWCSC offices.

Local authority colleagues identified two outcomes to predict for their own local authority and in total we predicted eight outcomes:

PREDICTION 1:
Does a child / young person's case come in as a 're-contact' within 12 months of their case being NFA-ed ('no further action'-ed), and does the case then escalate to the child being on a Child Protection Plan (CPP) or being Looked After (CLA)?

PREDICTION 2:
Does the child / young person's case progress to the child being subject to a Child Protection Plan (CPP) or being looked after (CLA) within 6-12 months of a contact?

PREDICTION 3:
Is the child / young person's case open to children's social care - but the child / young person not subject to a Child Protection Plan (CPP) or being Looked After (CLA) - within 12 months of their case being designated 'No Further Action'?

PREDICTION 4:
Is the child or young person's case which is already open to children's social care being escalated (to the child being subject to a Child Protection Plan, being Looked After, being adopted, being subject to a Residence Order or being subject to a Special Guardianship Order) between three months and two years of the referral start date?

PREDICTION 5:
Does the child / young person's case progress to the child being subject to a Child Protection Plan (CPP) or the child being Looked After (CLA) within 6-12 months of a contact?

PREDICTION 6:
After successfully finishing early help, is the child / young person referred to statutory children's services within 12 months?

PREDICTION 7:
Does the child / young person's case progress to the child being subject to a Child Protection Plan (CPP) within 1-12 months of the assessment authorisation date?

PREDICTION 8:
Does the child / young person progress to the child being Looked After (CLA) within 1-12 months of the assessment authorisation date?

To test each outcome, we split the historical data into 'training' and 'test' data, training the model on the training data and then testing whether the patterns learned generalise to the test data. This simulates how well the models would perform if used in practice to predict the outcome for new cases. We built four different models for each of the eight outcomes (so 32 models in total) to investigate two questions of interest:

i) whether including pseudonymised text data improved the performance of the models;

ii) what the effect on performance is of only allowing case data to be used to predict future cases.

To answer the first question, we first built the model using just the structured data and measured its performance. We then built a second model adding in information extracted from the text.

To answer the second question, we first built the model, allowing it to learn cases irrespective of whether these cases were before or after the case at hand, then tested the model on the outcome for the remaining unseen cases, measuring its performance. We then built a second model but restricted it to learning patterns from just earlier cases. Imposing this restriction is a stricter test but a more accurate reflection of a real world scenario.

Given that there may be justice concerns if the models misidentified individuals with particular sensitive characteristics as at risk or failed to identify those at risk, we also examined how well the model performs for individuals of different genders, disability statuses, ethnicities and age groups.

We report the performance of all the models. Please note that we exclude the model performance from LA1 (predictions 1 and 2) when evaluating whether the models work overall. This is because of 'information leakage' - which likely artificially inflates the performance metric - identified in subsequent analysis after the modelling was complete. However, this is likely to affect each of the models within LA1 equally and so comparing LA1's models to each other is still helpful.

THE RESULTS

Once we had built the models, we tested whether the patterns learned enabled them to predict well on cases the model hadn't yet 'seen' but whose outcome was already known. The number of misclassifications the model makes of cases as at risk or not at risk of the outcome give an indication of the numbers and types of misclassifications the models would make if they were used to assist social workers in practice.

Metrics which summarise the overall model performance count the correct and incorrect classifications of the model. In the table overleaf, we report two different ways of summarising the performance of the model: average precision and 'area under the curve' (AUC). Both metrics are measured on a scale of 0 to 1 with 0 being the worst possible model and 1 being the best possible model.

The average precision metric is more appropriate for our predictions because we are looking for a 'needle in a haystack' (the proportion of cases at risk of the outcome is quite small, ranging from 2%-17% with seven of the outcomes being 2%-7%).

OVERALL FINDINGS

We specified in our research protocol[2] before beginning the analysis that we would deem the model a 'success' if it scored above 0.65 average precision. This is lower than the threshold we would recommend for putting a model into practice but provides a useful low benchmark.

Overall, none of the 24 models had average precision scores which exceeded the threshold for success. Ten of the 24 models have AUCs greater than 0.65 but this 'success' reflects that the model correctly identifies most of the cases not at risk as not at risk, an 'easy win' when most cases are not at risk of the outcomes we're predicting.

Average precision focuses on the tradeoff between two goals:

A precise model
which - when identifying cases as at risk of the outcome of interest - is right the majority of the time. For example, if a model with a precision of 0.9 flags 100 cases as at risk, then 90 (0.9 x 100) of the cases it flags are at risk of the outcome. The other ten it flags will be false alarms.

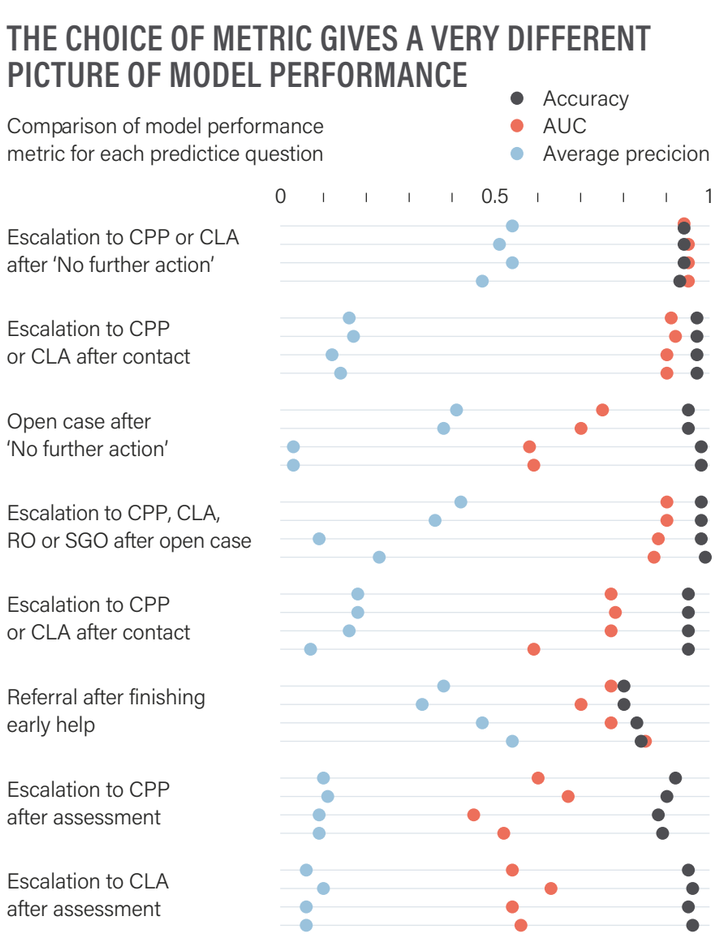
A model with high recall
which identifies most of the cases at risk of the outcome of interest. For example, if there are 100 cases at risk of the outcome, and the model has a recall of 0.9, then the model will flag 90 (0.9 x 100) of those 100 cases as at risk. It will miss the other ten.

OVERALL, NONE OF THE 24 MODELS HAD AVERAGE PRECISION SCORES WHICH EXCEEDED THE THRESHOLD FOR SUCCESS

HOW IMPORTANT IS CHOICE OF METRIC?

How well the model performs can be measured by examining the unseen cases on which the model predicts correctly, and the unseen cases which the model gets wrong. There are a number of ways to measure the performance of a model but they are all some combination of the number and types of correct classifications and misclassifications. The choice of metric depends on whether you are more concerned about false alarms or missing children and young people at risk of the outcome, as well as a few other factors.

THE CHOICE OF METRIC GIVES A VERY DIFFERENT PICTURE OF MODEL PERFORMANCE



HOW WELL THE MODEL PERFORMS CAN BE MEASURED BY EXAMINING THE UNSEEN CASES ON WHICH THE MODEL PREDICTS CORRECTLY, AND THE UNSEEN CASES WHICH THE MODEL GETS WRONG

We use the average precision as our key metric. This balances the tradeoff between false alarms and missing children at risk and is useful in situations where you see very few observations of the outcome being predicted. In contrast, the 'area under the curve' (AUC) does not take into account that it is an 'easy win' for the model to correctly identify most of the cases not at risk as not at risk when most cases are not at risk of the outcome we're predicting. For this reason, AUCs tend to be higher for datasets where the vast majority of cases are not at risk - as is the case in our models. Accuracy is also measured from 0 to 1 with 0 being the worst possible model and 1 being the best possible model. The accuracy scores show the proportion of cases the models correctly identify. Again, the accuracy is always reasonably high when the vast majority of cases are not at risk of the outcome because the model easily correctly identifies those not at risk.

PERFORMANCE METRICS

Average precision exceeds threshold (0.65)

AUC exceeds threshold (0.65)

	Local Authority	Learned from all cases / Learned only from earlier cases	Structured data only / Text data included	Average precision	AUC
PREDICTION 1: Escalation to CPP or CLA after 'No further action'	1	⊘	✎	0.54	0.94
	1	⊘	⊘	0.51	0.95
	1	🕒	✎	0.54	0.95
	1	🕒	⊘	0.47	0.95
PREDICTION 2: Escalation to CPP or CLA after contact	1	⊘	✎	0.16	0.91
	1	⊘	⊘	0.17	0.92
	1	🕒	✎	0.12	0.9
	1	🕒	⊘	0.14	0.9
PREDICTION 3: Open case after 'No further action'	2	⊘	✎	0.41	0.75
	2	⊘	⊘	0.38	0.7
	2	🕒	✎	0.03	0.58
	2	🕒	⊘	0.03	0.59
PREDICTION 4: Escalation to CPP, CLA, RO or SGO after open case	2	⊘	✎	0.42	0.9
	2	⊘	⊘	0.36	0.9
	2	🕒	✎	0.09	0.88
	2	🕒	⊘	0.23	0.87
PREDICTION 5: Escalation to CPP or CLA after contact	3	⊘	✎	0.18	0.77
	3	⊘	⊘	0.18	0.78
	3	🕒	✎	0.16	0.76
	3	🕒	⊘	0.07	0.59
PREDICTION 6: Referral after finishing early help	3	⊘	✎	0.38	0.77
	3	⊘	⊘	0.33	0.7
	3	🕒	✎	0.47	0.77
	3	🕒	⊘	0.54	0.85
PREDICTION 7: Escalation to CPP after assessment	4	⊘	✎	0.1	0.6
	4	⊘	⊘	0.11	0.67
	4	🕒	✎	0.09	0.45
	4	🕒	⊘	0.09	0.52
PREDICTION 8: Escalation to CLA after assessment	4	⊘	✎	0.7	0.56
	4	⊘	⊘	0.1	0.63
	4	🕒	✎	0.06	0.54
	4	🕒	⊘	0.06	0.56

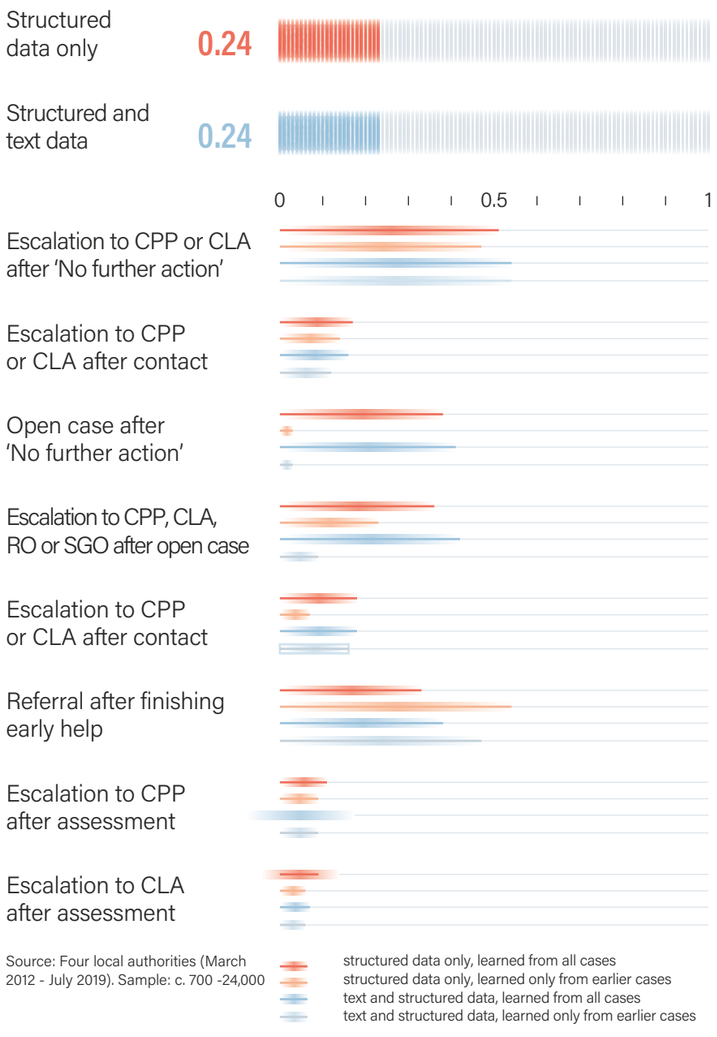
Source:
Four local authorities (March 2012 - July 2019).
Sample:
c. 700 -24,000

DOES INCLUDING TEXT DATA IMPROVE MODEL PERFORMANCE?

Including text data improves the model performance in six out of 16 like-for-like comparisons but worsens the model in a further six. In the cases where the model performance is worse, the models are learning patterns from the cases that don't generalise well to other cases. Text does not seem to help with interpreting the predictions made by the models: the 'topics' identified in the reports and assessments do not clearly distinguish between cases at risk and those not at risk of the outcome. Given these findings, the additional costs associated with handling text data, particularly in data governance, seem unlikely to be rewarded.

INCLUDING TEXT DATA DOES NOT IMPROVE MODEL PERFORMANCE

Average precision for each outcome predicted: comparing models including and excluding text data



Source: Four local authorities (March 2012 - July 2019). Sample: c. 700 -24,000

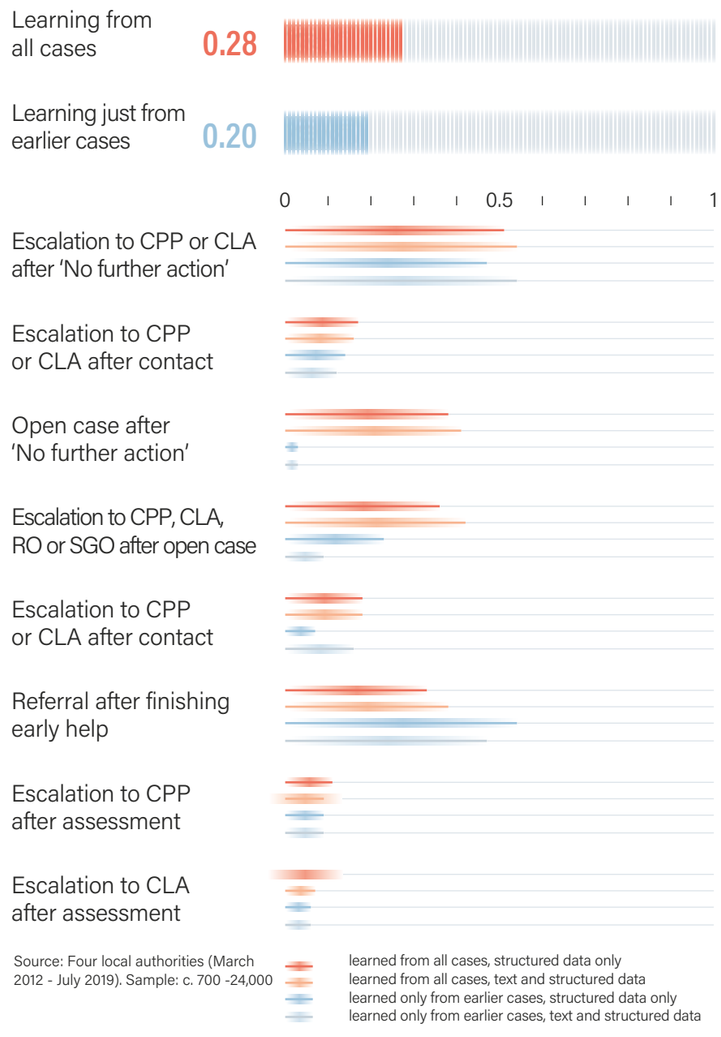
DOES RESTRICTING THE MODEL TO LEARNING FROM EARLIER CASES HARM MODEL PERFORMANCE?

In 13 out of the 16 like-for-like comparisons, model performance drops when restricting the model to learn just from earlier cases - in other words, when more accurately simulating the real world of standing in the present and predicting the future. This finding is important from a perspective of transparency: presenting models which ignore whether cases were earlier or later when learning patterns in the data may artificially inflate the performance of the model.

IN 13 OUT OF THE 16 LIKE-FOR-LIKE COMPARISONS, MODEL PERFORMANCE DROPS WHEN RESTRICTING THE MODEL TO LEARN JUST FROM EARLIER CASES - IN OTHER WORDS, WHEN MORE ACCURATELY SIMULATING THE REAL WORLD

MODEL PERFORMANCE IS POORER WHEN RESTRICTING THE MODEL TO LEARNING FROM EARLIER CASES

Average precision, averaged over all outcomes predicted: comparing models restricted to learning from earlier cases and not restricted



Source: Four local authorities (March 2012 - July 2019). Sample: c. 700 -24,000

IS THE MODEL BIASED AGAINST CERTAIN GROUPS?

One major concern about the use of predictive analytics is that it is biased against those with particular characteristics. There are two main forms of this bias. First, that the model perpetuates existing biases that exist in reality; for example because of structural or other forms of racism, sexism etc. It is difficult to handle this type of bias as the models can only learn from the patterns of data as they are recorded.

The second form of bias is that the model performs worse for some groups than for others. In this section, we consider whether the models suffer from this form of bias.

If a model which makes systematically more errors for children with particular protected characteristics were to be used in practice to assist social worker decision-making, we would have concerns about just access to support when needed and/or just freedom from interference in family life. For this reason, we investigated the performance of the models broken down by particular protected / sensitive characteristics, namely, gender, disability status, ethnicity and age group. Overall, the models perform equally well (or, indeed, poorly) for each subgroup with the exception of under one year olds for whom they perform better, and 16+year olds and those identifying their ethnicity as "Other ethnicity" for whom the models perform worse (on average). However, whether the models are biased is very sensitive to the type of test we carry out. From one angle, the models perform differently for 90% of the subgroups when comparing them 'head-to-head' and from another angle the models perform differently for only 10% of the subgroups compared 'head-to-head'. Given the sensitivity of the results to the type of test carried out, we encourage extensive testing before models are considered for being used in practice.

MODEL PERFORMANCE DOESN'T VARY MUCH BY SUBGROUP

Comparison of mean average precision for subgroups

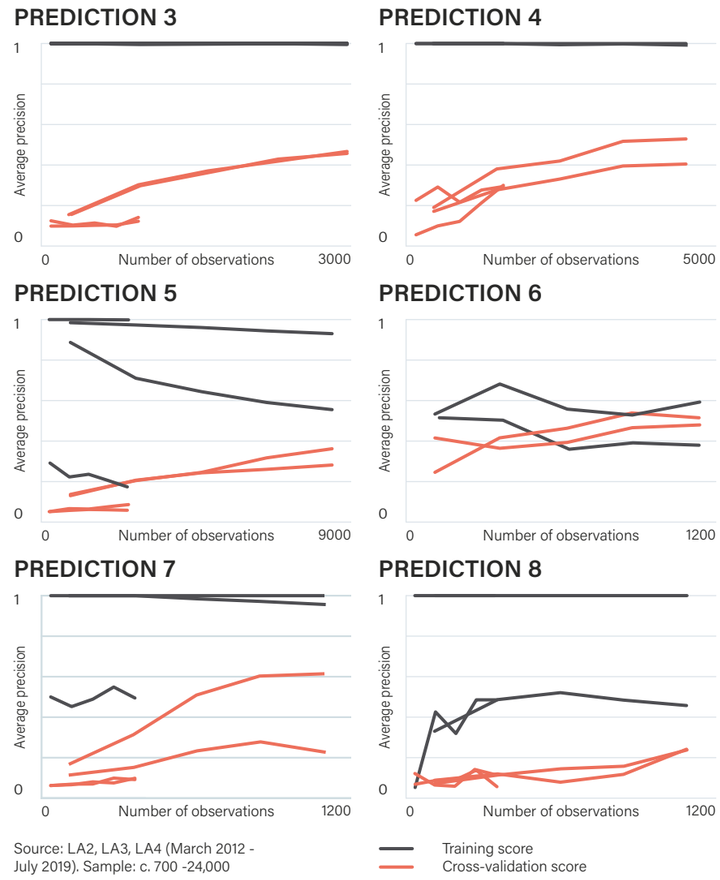
Source: Four local authorities (March 2012 - July 2019). Sample: c. 700 -24,000

WHAT WOULD HELP IMPROVE THE MODELS?

In some cases, giving the data more observations to learn from can help. How many observations of cases is 'enough' depends on multiple factors. Whilst datasets of the same size may be sufficiently large in a different context, outcomes in the children's social care context are more complex than a typical ML use case and are like looking for 'a needle in a haystack,' in that a small proportion of the population go on to experience the outcome being predicted. Furthermore, the children's social care context involves families, and because siblings tend to have similar outcomes, the model can learn less from observations about related children and young people, which 'deflates' the size of our datasets.

To give us an insight into whether increasing the sample size would help, we trained the model using progressively larger subsets of the data available. For 14 out of the 24 models, it seemed that increasing the number of observations available to the model would improve the model's performance (the model performance doesn't plateau in the graphs to the right). However, additional analysis suggests that the data changes sufficiently over time and that the model may learn patterns which don't match the current context if data that spans too great a time period is used.

INCREASING THE NUMBER OF OBSERVATIONS AVAILABLE TO THE MODEL WOULD IMPROVE THE MODEL'S PERFORMANCE FOR 14 OUT OF THE 24 MODELS



Source: LA2, LA3, LA4 (March 2012 - July 2019). Sample: c. 700 -24,000

Given that children's social care is the responsibility of local government, and local authorities have different practice models and systems for recording the data, it seems unlikely that more observations could be usefully sourced from other local authorities. This presents a ceiling in the useful data available to local authorities to build models to predict social care outcomes.

In other cases, providing a richer dataset is sometimes offered as a solution. Adding richer data about each individual child is helpful when the model isn't picking up enough of the nuance of the patterns to perform well on data it has already seen as well as unseen data. However, this does not correctly diagnose the problem with our models which try to over-generalise the patterns they observe (as shown above by the large gap in performance between the training and validation curves).

WHAT DO SOCIAL WORKERS THINK?

We polled the WWCSC social workers panel[3] to get a sense of the views of the profession. 129 of the panel responded. Around three in ten participants (29%) did not know what predictive analytics was. Only 10% of participants thought that predictive analytics had a role to play in decision making in social care, while 29% thought it didn't and 32% weren't sure. We then asked "If predictive analytics were to be introduced in children's services, which of the following uses might it be acceptable?" 34% thought that it should not be used at all. The use which the highest percentage of participants thought was acceptable was producing a tool which would support social workers to identify early help for families (26% thought this use would be acceptable). Just 14% of participants thought producing a tool which would help social workers prioritise cases to discuss with their managers would be acceptable. Another 14% thought that producing a tool that helps decide whether a family meets the threshold for social work intervention would be acceptable. Only 12% thought that producing a tool which helps to decide whether to close a case would be acceptable.

ONE IN TEN THINK THERE IS A ROLE FOR PREDICTIVE ANALYTICS IN SOCIAL WORK DECISION-MAKING

Do you think that predictive anayltics has a role to play in decision making in social care?

YES 10% I'm not sure 32% I don't know what it is 29% NO 29%

THERE IS NO CLEAR SUPPORT FROM SOCIAL WORKERS FOR ANY POTENTIAL USES OF PREDICTIVE ANALYTICS

If predictive analysis were introduced in childrens services, for which of the following uses would it be acceptable?

Producing a tool which would help social workers prioritise cases to discuss with their managers — 14%

Producing a tool which would support social workers to identify early help for families — 26%

Producing a tool that helps to decide whether a family meets the threshold for social work interventions — 14%

Producing a tool which helps to decide whether to close a case — 12%

I do not think it should be used at all — 34%

Source: WWCSC social worker poll, March 2020. Sample: 129

DISCUSSION

In summary, we do not find evidence that the models we created using machine learning techniques 'work' well in children's social care.

In particular, the models built miss the majority of children at risk of the outcome which - were the models to be used in practice - risks the model discouraging a social worker to support a child or young person when it is needed, potentially resulting in harm to them. Our findings provide evidence on 'what works' in the context of using administrative data to predict outcomes of children and young people with experience of children's social care in England. We do not pretend that they offer a definitive answer to whether machine learning is worthwhile pursuing in this context. However, our findings of poor predictive performance reflect the findings of a large scientific collaboration[4] of 160 teams published in the prestigious Proceedings of the National Academy of Sciences predicting life outcomes. The outcomes include outcomes related to children's protective services and the teams used 15 years of high quality data relating to a similar sample size of children (c. 4000) in the United States. Although the geographical context is different and the data is questionnaire data collected every few years, our findings and the findings of the 160 teams suggest that it is very challenging to build models to predict outcomes well in children's social care.

The responses from the small sample of social workers we polled suggest that machine learning is considered acceptable by only a small proportion of the profession. Given that the purpose of such models would be to aid social workers, for the models to be used in practice a considerable amount of groundwork would need to be done to bring social workers on board. As outlined in a report[5] by the Oxford Internet Institute, machine learning is also unlikely to save local government money, at least in the short term, because funds need to be available for early intervention for those identified as at risk.

OUR FINDINGS AND THE FINDINGS OF THE 160 TEAMS SUGGEST THAT IT IS VERY CHALLENGING TO BUILD MODELS TO PREDICT OUTCOMES WELL IN CHILDREN'S SOCIAL CARE

We did not seek to definitively answer the question of whether machine learning will ever work in children's social care across all types of outcomes and in all contexts, but we hope to have shown some of the challenges faced when using these approaches in children's social care. For local authorities already piloting machine learning, we encourage them to also be transparent about the challenges they experience. Given these challenges and the extent of the real world impact a recommendation from a predictive model used in practice could have on a family's life, it is of utmost important that we work together as a sector to ensure that these techniques are used responsibly if they are used at all.

REFERENCES

1 Office of the Children's Commissioner for England (2019), Childhood vulnerability in England 2019 www.childrenscommissioner.gov.uk/report/childhood-vulnerability-in-england-2019

2. What Works for Children's Social Care. 2019, July. Pilots of predictive analytics in children's social care: research protocol. www.whatworks-csc.org.uk/research-project/predictive-analytics

3. For more details on the polling panel, please see 'Reaching Out to Social Workers' on www.whatworks-csc.org.uk/about

4. Matthew J. Salganik, Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul, Jennie E. Brand, Nicole Bohme Carnegie, Ryan James Compton, Debanjan Datta, Thomas Davidson, Anna Filippova, Connor Gilroy, Brian J. Goode, Eaman Jahani, Ridhi Kashyap, Antje Kirchner, Stephen McKay, Allison C. Morgan, Alex Pentland, Kivan Polimis, Louis Raes, Daniel E. Rigobon, Claudia V. Roberts, Diana M. Stanescu, Yoshihiko Suhara, Adaner Usmani, Erik H. Wang, Muna Adem, Abdulla Alhajri, Bedoor AlShebli, Redwane Amin, Ryan B. Amos, Lisa P. Argyle, Livia Baer-Bositis, Moritz Büchi, Bo-Ryehn Chung, William Eggert, Gregory Faletto, Zhilin Fan, Jeremy Freese, Tejomay Gadgil, Josh Gagné, Yue Gao, Andrew Halpern-Manners, Sonia P. Hashim, Sonia Hausen, Guanhua He, Kimberly Higuera, Bernie Hogan, Ilana M. Horwitz, Lisa M. Hummel, Naman Jain, Kun Jin, David Jurgens, Patrick Kaminski, Areg Karapetyan, E. H. Kim, Ben Leizman, Naijia Liu, Malte Möser, Andrew E. Mack, Mayank Mahajan, Noah Mandell, Helge Marahrens, Diana Mercado-Garcia, Viola Mocz, Katariina Mueller-Gastell, Ahmed Musse, Qiankun Niu, William Nowak, Hamidreza Omidvar, Andrew Or, Karen Ouyang, Katy M. Pinto, Ethan Porter, Kristin E. Porter, Crystal Qian, Tamkinat Rauf, Anahit Sargsyan, Thomas Schaffner, Landon Schnabel, Bryan Schonfeld, Ben Sender, Jonathan D. Tang, Emma Tsurkov, Austin van Loon, Onur Varol, Xiafei Wang, Zhi Wang, Julia Wang, Flora Wang, Samantha Weissman, Kirstie Whitaker, Maria K. Wolters, Wei Lee Woon, James Wu, Catherine Wu, Kengran Yang, Jingwen Yin, Bingyu Zhao, Chenyun Zhu, Jeanne Brooks-Gunn, Barbara E. Engelhardt, Moritz Hardt, Dean Knox, Karen Levy, Arvind Narayanan, Brandon M. Stewart, Duncan J. Watts, Sara McLanahan.
Measuring the predictability of life outcomes with a scientific mass collaboration
Proceedings of the National Academy of Sciences Apr 2020, 117 (15) 8398-8403;

5. Bright, J., Ganesh, B., Seidelin, C. & Vogl, T. (2019, March) Data Science for Local Government. Oxford Internet Institute, University of Oxford.

GLOSSARY OF TERMS

Early Help (EH)	Pre-social care, services designed to meet the needs of families where there are lower level support needs (not child protection) to prevent them entering the social care system.
Contact	Any contact made with children's social care.
Referral	A referral, in the context of child protection, is when someone contacts Children's Services because they have concerns about the safety and well-being of a child. An initial assessment will determine if the threshold has been met.
Assessment	Unless the child requires immediate protection, the majority of cases will begin with a social worker conducting an assessment. During the assessment, the social worker gathers information and analyses the needs of the child or children and/or their family and the nature and level of any risk of harm.
No further action (NFA)	'No further action' means that there will be no statutory intervention taken by a social worker but may include giving advice, signposting or stepping down. NFA can occur at multiple stages in the child's journey through social care.
Child Protection Plan (CPP)	A child protection plan sets out how the child can be kept safe, how things can be made better for the family and what support they will need. If the child is made the subject of a child protection plan, it means that the social worker considers the child to be at risk of significant harm.
(Child) Looked After (CLA)	When a child is placed somewhere other than with their legal guardian by the local authority. Typically this means in foster care but also includes kinship, respite and residential care.
Residence Order (RO)	A residence order is a court order which establishes where a child will live.
Special Guardianship Order (SGO)	A special guardianship order is a court order appointing one or more individuals to be a child's 'special guardian.' It is intended for those children who cannot live with their birth parents but would benefit from a legally secure placement.

What Works *for*
Children's
Social Care

CONTACT

info@whatworks-csc.org.uk
@whatworksCSC
whatworks-csc.org.uk