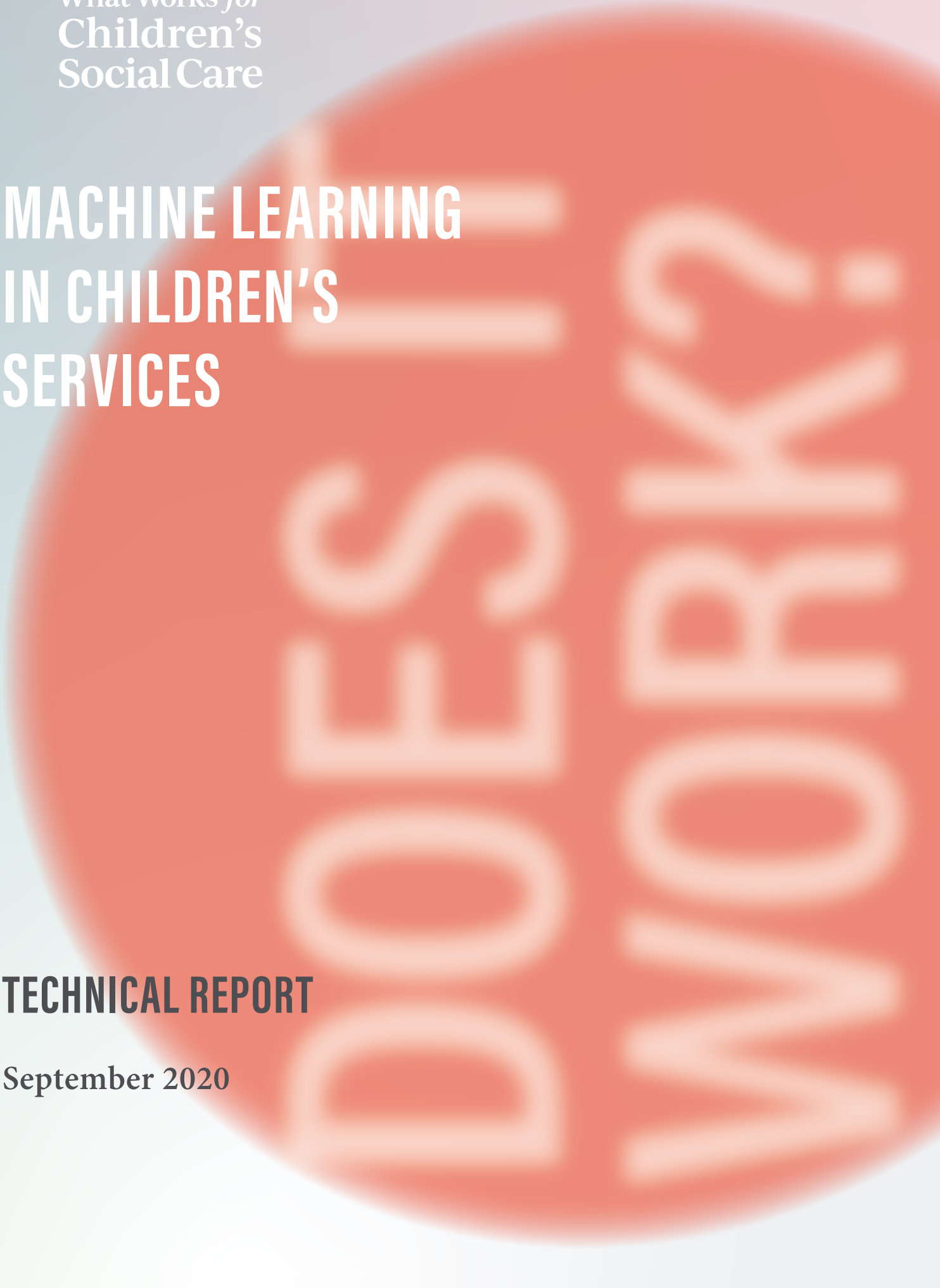
What Works *for* Children's Social Care

MACHINE LEARNING IN CHILDREN'S SERVICES



TECHNICAL REPORT

September 2020

What Works *for* Children's Social Care

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.

Funding

Department for Education, England.

Authors

Vicky Clayton

Michael Sanders

Eva Schoenwald

Lee Surkis

Dan Gibbons

About What Works for Children's Social Care

What Works for Children's Social Care seeks better outcomes for children, young people and families by bringing the best available evidence to practitioners and other decision makers across the children's social care sector. We generate, collate and make accessible the best evidence for practitioners, policy makers and practice leaders to improve children's social care and the outcomes it generates for children and families.

To find out more visit our website: whatworks-csc.org.uk

If you'd like this publication in an alternative format such as Braille, large print or audio, please contact us at: info@whatworks-csc.org.uk

CONTENTS

CONTENTS

LIST OF TABLES AND FIGURES

EXECUTIVE SUMMARY

Introduction

In addition to fostering the relationship with the child and family, much of assessment is about the social worker evaluating risk and predicting future outcomes as accurately as possible. Social workers draw on their experience and the experience of their colleagues to make such judgements. In foreign jurisdictions[1] and a small but increasing number of local authorities in England (we think approximately 10%), predictive models are being piloted to assist social workers by predicting outcomes relating to child protection. In their favour, they may be helpful as decision aids to social workers when undertaking assessments, or identifying the cases most at risk to team managers. However, at worst inaccurate models can contribute to unnecessary intervention or a lack of necessary intervention.

Research Aims

The aim of this research was to help local authorities understand whether it's worthwhile investing in developing these types of models and associated tools to assist social workers in practice, and what are the important choices to be made. We assessed the performance of predictive models for a range of different predictions, for example, whether a child enters care or becomes subject to a child protection plan, across four local authorities (who have remained anonymous and whom we refer to as LA1, LA2, LA3 and LA4). The local authorities were of a range of sizes, practice models and case management systems which helped us understand how some of the practical obstacles (for example, the ease of data extraction from the case management system, quality of data etc) might play out in different settings.

Key research questions were:

1. Do the models perform accurately enough to be useful for local authorities to commission or develop?
2. What practical obstacles would local authorities face in building or commissioning the building of these models?
3. Does including text data improve the performance of predictive models?
4. Does restricting predictive models to just learning patterns from earlier cases (rather than subsequent cases) worsen performance? (The former being a better simulation of how well the model would predict on new cases).
5. Does increasing the amount of data used in a model improve the model performance? This was chosen to feed into local authority decision making about when to invest in these models (should they be convinced that it was worthwhile to do so).
6. Does model performance differ for people with particular characteristics (age group, gender, disability, ethnicity)? Higher error rates for children with particular characteristics would be of concern if the model were to be used in practice to assist social worker decision making because the model would overstate or understate the level of concern for children with these characteristics.

1 For example, Allegheny County, Pennsylvania using predictive analytics in screening. (Allegheny County Department of Human Services. May, 1st 2019. *Developing Predictive Risk Models to Support Child Maltreatment Hotline Screening*. Decisions. https://www.alleghenycountyanalytics.us/index.php/2019/05/01/developing-predictive-risk-models-support-child-maltreatment-hotline-screening-decisions/) and in Florida (ECKERD CONNECTS | ECKERD RAPID SAFETY FEEDBACK, https://eckerd.org/family-children-services/ersf/).

We were also interested in knowing when the model returns a probability, over what range can we expect the 'true' value to lie. This helps us understand what level of granularity (e.g. buckets of 5% or 10%) would be appropriate for presenting the predictive probabilities should probabilities be the method of presentation.

Research Design

The research involved building models to predict outcomes in children's social care, for example, a child entering care, and evaluating their performance of the models. The performance of the model tells us how many errors the models make and what kind of errors, which is important to know if we are testing whether they would be useful to assist decision-making. In order to test whether including text data improved the models, we conducted 'topic modelling', a natural language processing (NLP) technique to find human-recognisable themes, and used how well each topic represented the documents associated with the case as predictive features (variables) in the models. We anticipated that being able to incorporate the information from reports would be particularly important to the children's social care context because much of the nuanced information about the child and family is contained within these reports - their context, risk factors, strengths and interactions with other services.

We assessed the performance of the models by predicting the outcomes for children and young people from unseen historical data ('holdout data'). This is a common workflow when building predictive models which provides a safe environment to experiment with ways to improve the model without conducting any 'live' predictions on cases that social workers were making decisions about. No decisions regarding any individual cases were taken by, or as a result of, the model.

Predictions

We predicted two outcomes per local authority (LA1: predictions 1-2; LA2: predictions 3-4; LA3: predictions 5-6; LA4: predictions 7-8). All the predictions related to predicting whether a case would escalate but from different points of a child's journey and to different levels of concern, for example, a child protection plan or the child becoming looked after. The aim of a child protection plan is to ensure the child is safe from harm and prevent them from suffering further. A child is considered 'looked after' when a child is placed somewhere other than with their legal guardian by the local authority. (Appendix 1 includes a glossary of children's social care terms and acronyms.)

- **Prediction 1** ('Escalation to CPP or CLA after 'No further action''): 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 ('Escalation to CPP or CLA after contact'): Does the child / young person's case progress to the child being subject to a CPP or being looked after (CLA) within 6-12 months of a contact?

- **Prediction 3** ('Open case after 'No further action''): 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** ('Escalation to CPP, CLA, RO or SGO after open case'): 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** ('Escalation to CPP or CLA after contact'): 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** ('Referral after finishing early help'): After successfully finishing early help, is the child / young person referred to statutory children's services within 12 months?
- **Prediction 7** ('Escalation to CPP after assessment'): 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** ('Escalation to CLA after assessment'): Does the child / young person progress to the child being Looked After (CLA) within 1-12 months of the assessment authorisation date?

Analyses

We developed four models for each outcome predicted (creating a total of 32 models). To investigate the impact on the performance metrics of including text data, we compare models with and without text data. We also test whether restricting the models to just learning patterns from earlier cases (rather than subsequent cases) worsens performance: including cases which occur after the case being predicted exposes the model to additional information that it would not have available if the model were to be deployed in practice. Imposing this restriction is a stricter test but a more accurate reflection of a real world scenario. These comparisons involved the development of four models:

	Ignoring the restriction on learning from earlier cases in validating and evaluating the model		Restricting the model to learning from earlier cases in validating and evaluating the model	
Including structured data	Learned from all cases	Structured data only	Learned only from earlier cases	Structured data only
Including structured and text data	Learned from all cases	Includes text data	Learned only from earlier cases	Includes text data

For each model, we tried three binary classification algorithms (decision tree, logistic regression, and gradient boosting) which differ in complexity and cost[2] to develop and maintain.

We optimised for the average precision which allowed us to balance the trade-off 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.
- A model with **high recall** which identifies most of the cases at risk of the outcome of interest.

There is an inherent tradeoff between the two because being more confident that the model is

2 The local authorities contributed 'in kind' to this project through the provision of staff time for managing the project internally, extracting the data etc. and so did not incur any cost in monetary terms for this project. But we wanted to keep in mind the cost factor if the local authority were to commission or develop these models themselves.

correct when it predicts a case at risk requires setting the threshold for what 'counts' as at risk pretty high, which means identifying cases with a lower probability of the outcome as not at risk, and hence missing more borderline cases.

We validated the model using cross-validation (a technique which helps reduce the chance that the model learns patterns from the training dataset which aren't generalisable to other cases) and evaluated the average precision of the final model chosen on unseen data.

Results

None of the models performed well enough to exceed our minimum threshold of a 'well performing' model (which we defined before starting the analysis) and the threshold was not set ambitiously.[3] The models tend to miss children at risk. It appears that adding more data may improve the model performance for just over half of the models. However, the input data changes enough over time that gathering more data from the archive or waiting for more data to accrue is unlikely to help the model. We find weak evidence that the models which learn only from cases preceding them perform worse. Because we believe it to be a more accurate reflection of how well the model would perform if deployed in practice, we suggest that this is an important metric to report for future models designed for the children's social care context where the time horizons for outcomes are relatively long. The models including text data did not perform better than the models using just structured data and tended to learn patterns which didn't generalise more than the models including just structured data.

We also evaluated whether the models perform about the same for subgroups with protected or sensitive characteristics (different age groups, genders, disabilities or ethnicities). However, our results were very sensitive to the type of analysis we conduct. One methodology (non-parametric tests) suggested that the models perform differently for all the groups whilst another (checking whether the confidence intervals of the average precision overlap) suggested that the models perform indistinguishably for most subgroups. Given the sensitivity to how bias is tested, we would encourage a great deal of transparency around how the model is evaluated for bias.

Discussion

In summary, we do not find evidence that machine learning techniques 'work' well in children's social care. In particular, the models miss a large proportion of the children at risk and as such would not fulfil an often made promise of identifying children and young people who would benefit from early intervention support. 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: other data teams may be able to build more useful models using the same data and techniques are likely to improve over time. However, we have reason to think that there may be ceilings on how well models can perform based on the amount of data being a limiting factor and there being practical constraints on obtaining more data which would aid performance. 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.

3 Please note that we do not include the results of LA1 when discussing the overall results from the project because of problems of information leakage identified after the analysis was complete (please see the 'Handling Data' section of the Detailed Methodology).

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.

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.

This is not just a question of technical interest. Children that the model misses may also be deprioritised by any decision making process that uses it and as such children can be left at risk of real harm. 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.

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

BACKGROUND

Introduction

Over the last 18 months, What Works for Children's Social Care has worked with four local authority partners, developing models on outcomes they were interested in predicting with the aim to also draw more general conclusions about how well these techniques work in the context of children's social care and what are the important choices to be made.

Partners

Local Authorities

We issued an open call[5] for local authorities to work with us on the project. We worked with four local authorities[6]. 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.

The local authorities extracted the data, chose the outcomes to predict and provided expertise in interpreting their data. Two of the local authorities hosted WWCSC researchers onsite to process the text data whilst the other two set aside IT resource to facilitate data access.

What Works for Children's Social Care (WWCSC)

WWCSC managed the project overall, conducted the analysis on behalf of each local authority and evaluated what the results collectively suggest for machine learning in children's social care.

Approach

Keeping the data separate for each local authority

We kept each local authorities' data separate and built a model for each local authority using their own data. Combining the data from different local authorities would have meant much more complex data governance arrangements, complexity matching data fields across case management systems and substantial checks that the distributions of the data were alike enough to be used in the same predictive models.

Building models for each local authority using their own data meant that the number of cases each model can learn from was small (at the smallest c.700 cases for a prediction in the early help context to c. 24,500 cases for the largest sample size) relative to the complexity of the prediction we're asking the model to make. This constraint focused our attention on questions early in a child's journey through children's social care where there are a larger number of cases to maximise the potential sample size available.

5 What Works for Children's Social Care. February, 7th 2019. *Call for partners to pilot potential of analysing case notes to assist social workers*. https://whatworks-csc.org.uk/blog/call-for-partners-to-pilot-potential-of-analysing-case-notes-to-assist-social-workers/

6 In the research protocol, we set out the intention to work with six local authorities. Unfortunately but understandably one local authority felt the need to reassign the resources dedicated to this innovation project to more pressing day-to-day needs. The sixth local authority pilot could not go ahead because of the complications due to the Covid-19 crisis.

Conducting the analysis sequentially

We conducted the analysis at each local authority sequentially (rather than in parallel). There was considerable learning in terms of the technical approach as well as changes of circumstances as we continued with the project and we shall outline where we deviated from the plan set out in the research protocol. We also note that we had a limited amount of time at each local authority (particularly onsite time) and because of this we had to be pragmatic about what is achievable in the time available. Because of this approach, some of the more exploratory analysis we developed later on is not available for LA1, for whom we conducted the analysis first.

Research Design and Analytical Approach

The research involved building predictive models to predict outcomes in children's social care, for example, a child entering care, and evaluating the performance of the models. The performance of the model tells us how many errors the models make and what kind of errors, which is important to know if we are testing whether they would be useful to assist decision-making.

We compared four models for each prediction (creating a total of 32 models). To investigate the impact on the performance metrics of including text data, we compared models with and without text data. We also test whether restricting the models to just learning patterns from earlier cases (rather than subsequent cases) worsens performance: including cases which occur after the case being predicted exposes the model to additional information that it would not have available if the model were to be deployed in practice. Imposing this restriction is a stricter test but a more accurate reflection of a real world scenario.

These comparisons involved the development of four models:

	Ignoring the restriction on learning from earlier cases in validating and evaluating the model		Restricting the model to learning from earlier cases in validating and evaluating the model	
Including structured data	Learned from all cases	Structured data only	Learned only from earlier cases	Structured data only
Including structured and text data	Learned from all cases	Includes text data	Learned only from earlier cases	Includes text data

or each model, we tried three binary classification algorithms (decision tree, logistic regression, and gradient boosting) which differ in complexity.

We set out our approach in a pre-registered research protocol which can be found here: https://osf.io/jwtf4/ registrations. Where our approach differs substantively from the intended approach, we discuss the reasons why.

Data

Focus on process-oriented outcomes at the start of the child's journey

The types of models we explored predict a binary (yes/no) outcome for an individual e.g. 'will child A be referred to children's services within 12 months of their case being closed?'

Predictive modelling focuses on predicting *at the level of the individual child or young person.* Different techniques would be appropriate for forecasting demand at a service level.

We focused on outcomes which are *binary* (yes/no) because the questions we considered which have continuous outcomes tended to require many more years of data than we could access e.g. the duration of being on a CIN (child in need) plan, or the number of CIN (child in need) episodes.

The questions focused on predicting *process-oriented outcomes* e.g. stepping up or down or closing a case rather than impact-oriented outcomes, e.g. safety or wellbeing, because predictive models need well-defined outcomes which are collected in the course of day-to-day operations and therefore fit into existing processes instead of, for example, survey data.

The outcomes differ slightly from those we specified in the trial protocol. Although we did attempt to estimate the sample size of the relevant populations for the years available and the percentage of cases at risk of the outcome to enable us to evaluate whether it was feasible to pursue predicting the outcome, this proved more difficult than initially anticipated because the criteria for the populations were quite specific. For example, the time frames of interest were sometimes different from the time frames for the publicly available figures. As another example, restricting the population to those who had been re-referred after 'No further action' where the 'No further action' designations which were considered a proxy for 'concern not met threshold' and not all 'No further action' designations was also difficult to estimate from publicly available figures. This meant that for some outcomes we anticipated predicting, we were required to re-evaluate the question.

In some cases, this required an adjustment of the time horizon over which we would expect the outcome to occur. We agreed a time horizon with the local authorities on the basis of what would sound salient to social workers whilst also providing a large enough sample size and class balance to work constructively with.

On further consideration, we split the prediction 'Are Children in Need escalating to child protection plans & are Children on Child Protection Plans escalating to 'Looked After Children' within 12 months?'[7] into two predictions ('Does the child / young person's case progress to being on a Child Protection Plan within 1-12 months of the assessment authorisation date?' and 'Does the child / young person progress to being Looked After within 1-12 months of the assessment authorisation date?') because the base population was conflating two populations which should be considered separately.

Data included in models

We only used information about the particular case that is available to the social worker and recorded in the case management systems at the time of the decision which the model is designed to assist.

We did not include sibling or other relative's data as inputs for the individual's predictions. This is because there may be 'information leakage' between relatives. Often, social workers copy and paste descriptions of incidents or updates across siblings' files because the information will be relevant to all siblings. Knowing the outcome for one sibling could unfairly inflate the performance of the models. Figure 1 shows the thought process for data fields to include.

7 This was an original question for LA4, which had not been decided at the time of publishing the research protocol.

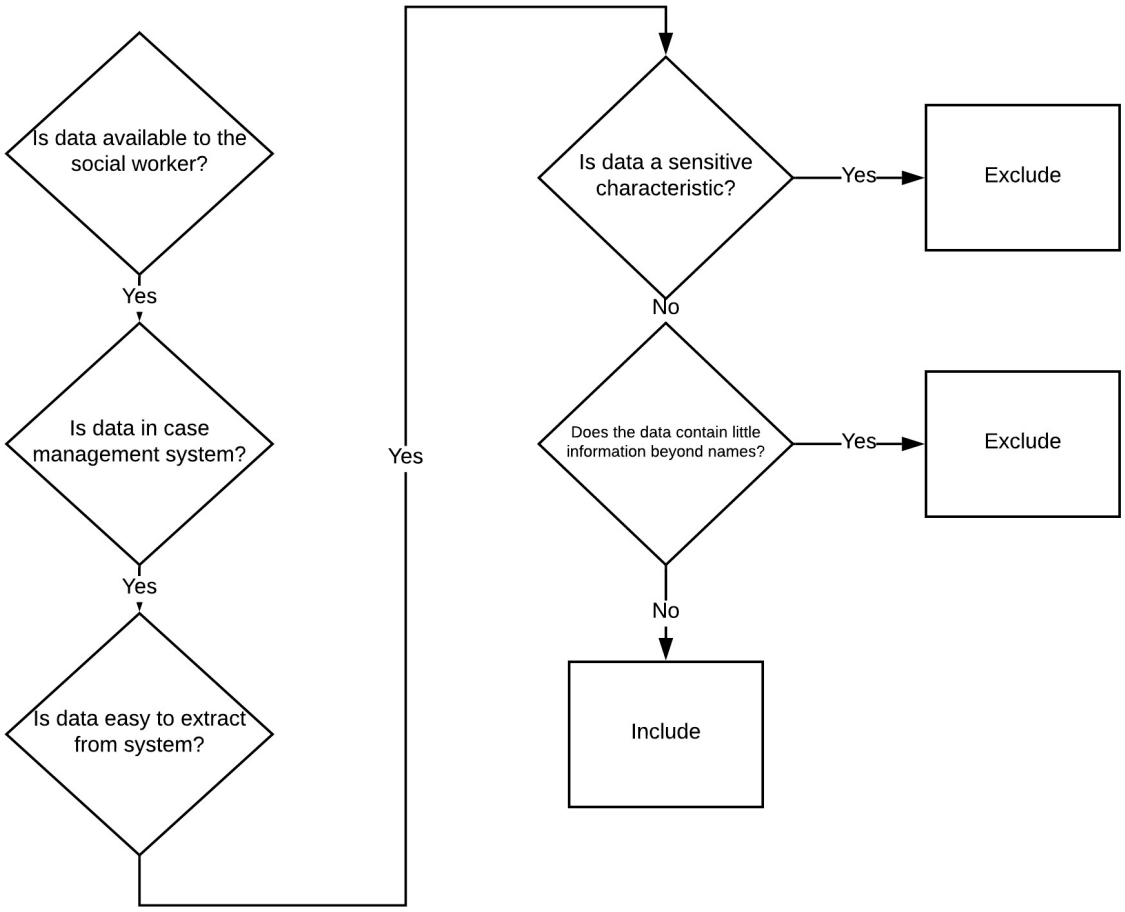


Figure 1: Decision-making process for which data to include in the model

Structured Data

The structured data were extracted from the case management systems, often using queries written to extract data for the purposes of creating the 'Annex A' dataset to share with Ofsted for inspection. When conducting an inspection of local authority children's services, Ofsted requests child level data. The data contains demographic information and information about their interactions with children's services. Ofsted provide guidance with standardised field headings and codes, as well as a template for the data requested.[8] 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. We chose to request these lists (or an adaptation of them) so that the data extraction request was not too onerous. In general, we used some combination of the following: list 1 (Contacts), list 3 (Referrals), list 4 (Assessments), list 5 (Section 47 enquiries and Child Protection Conferences), list 6 (Children in need), list 7 (Child Protection), list 8 (Children in care), and for Prediction 6, list 2 (Early help). All features which we specified in the trial protocol were available with the exception of school related data (attendance and exclusions) which is often stored in a different database than children's social care data so proved non-trivial

8 Please see the template for more detailed information about which variables are included: Ofsted. (2019, August). Child-level data: additional guidance and template for Annex A. https://www.gov.uk/government/publications/inspecting-local-authority-childrens-services-from-2018

to extract. We included the data where available and where it wouldn't cause the model to be prescient i.e. have knowledge of events about the individual's case before the point in time at which the model is making the prediction. An example of this is where you are predicting whether a child will become subject to a Child Protection Plan at the point of contact and the outcome for assessment is recorded as 'Section 47 enquiry' (the enquiry to decide whether to make a child subject to a child protection plan). The model would not know the outcome of the assessment at the point of contact (as assessments are made after contact).

To exclude 'prescient' data, we worked with the data teams at the local authorities to identify what information would not be known to a social worker at the time of the relevant decision. For LA2, LA3 and LA4, we also evaluated the feature importance during the model building process, and inspected features with high feature importance. Once consulting with local authority colleagues, we excluded features from this subset which they indicated would be prescient. As mentioned, the analysis was conducted sequentially, and we had not yet developed this approach for identifying prescient features when conducting the analysis at LA1. Subsequent evaluation of feature importance estimated the importance of the number of Child Protection Plans at between 0.14 and 0.6. Given that we were predicting escalation, it looks likely that the escalation we were predicting was included in the calculation of this feature instead of the feature referring to all historical child protection plans. Unfortunately, it is not possible to go back to the local authority to re-run the analysis. Since this represents considerable information leakage (thus artificially inflating the performance metrics), we exclude the results from predictions 1 and 2 when aggregating the performance metrics across the predictions. However, the information leakage is likely to affect each of the models within LA1 equally and so comparing the eight different models to each other is still helpful.

We excluded social worker ID. Although there is some evidence that there is variability in social worker decision-making (even with standardised test case vignettes)[9] and so we expect social worker ID to be a useful feature in prediction, we were interested in understanding the patterns of historical social worker decisions because of what they can tell us about the risk to the child rather than replicating the historical decision of the social worker. Additionally, this avoided the risk of this research being used as a historical performance tool.

We then generated additional 'features' (also known as columns, data fields, variables, predictors) from the structured data such as months (e.g. whether an incident occured in January, February etc.) and years (2013, 2014 etc.) relating to dates; the number of days between two events (e.g. a referral) or since the latest event; classifying information in data fields with a large number of options (e.g. contact sources, assessment team types); averages and counts of data associated with previous interaction with children's services (e.g. the number of times and the proportion of times the referral source was the school for previous referrals). For all of the local authorities, we had data over a certain number of years rather than all historical data for children who had current 'live' cases. For the children and young people whose cases appeared early in our sample then, the data about the number of and proportions of previous interactions with children's services may not truly reflect such interactions because the data simply was not available to us.

After creating these features, we had between c.150 and c.330 features for the models using just the structured data and c.250 to c.650 for the models incorporating both structured and text data, depending on the outcome being predicted. The models expect numerical rather than missing data or words. We null imputed missing values (for numerical features replacing null values with an arbitrary number and creating a missing indicator for the feature, and for

9 Keddell, E. (2014). Current Debates on Variability in Child Welfare Decision-Making: A Selected Literature Review. Soc. Sci. 2014, 3(4), 916-940. https://www.mdpi.com/2076-0760/3/4/916/htm

categorical features adding a missing indicator for the feature), and so these columns include the indicators of missingness for features with missing values. The model then selected which features added the most predictive value. In Appendix 1, we provide a summary statistics for features which had the highest importance for those models.

Text Data

The text data was extracted from the case management system using more custom queries often with additional support from the local authorities' IT teams as the analysts within the children's services data teams do not regularly export the documents associated with all cases within a certain date range. Depending on the research question, this included contact and referral records, assessments, reports to initial and / or review child protection conferences, and strategy discussion reports. These documents summarise specific incidences, action taken, collaborations with multi-agency partners, analysis of the risks present and the strengths of the family and recommendations.

In the research protocol, we set out the intention to include case notes (also known as observations or case summaries) and referrals for services. However, we had not anticipated the volume of information which 'case notes' covers when writing the research protocol. Case notes include records of all interactions with the child / young person, their family / caregivers and other professionals. Over a number of years, this can add up to hundreds of interactions for some children. Extracting this information from case management systems would have been a considerable ask of local authority colleagues, and we felt that the information would be adequately summarised in the reports.

Although we had wanted to consider each text box in each form / report separately - as we thought we may derive further insight from maintaining this structure - this proved difficult because there was a high proportion of empty text boxes, for example, when a child hadn't had a Strategy Discussion or been subject to a Child Protection Plan. For LA1 and LA4 (where there were relatively small numbers of empty documents detailing previous interactions with children's social care), we combined the text into one document. For LA2 and LA3, we dropped columns with many empty documents.

As mentioned, the models expect numerical data so we needed to process the text data to extract information from the documents and present it in numerical form. Prior to extracting information from the text, we removed formatting and pseudonymised the text. We extracted the linguistic features such as sentiment and polarity (on which more information below) at this stage because the additional context of the syntax is still helpful to estimating the value of these features. We then standardised the text, removing syntax to create 'bags of words' by removing punctuation, changing the words to lowercase, removing stop words and lemmatising the words (e.g. changing from 'played' or 'playing' to 'play'), Please see Figure 2 below.

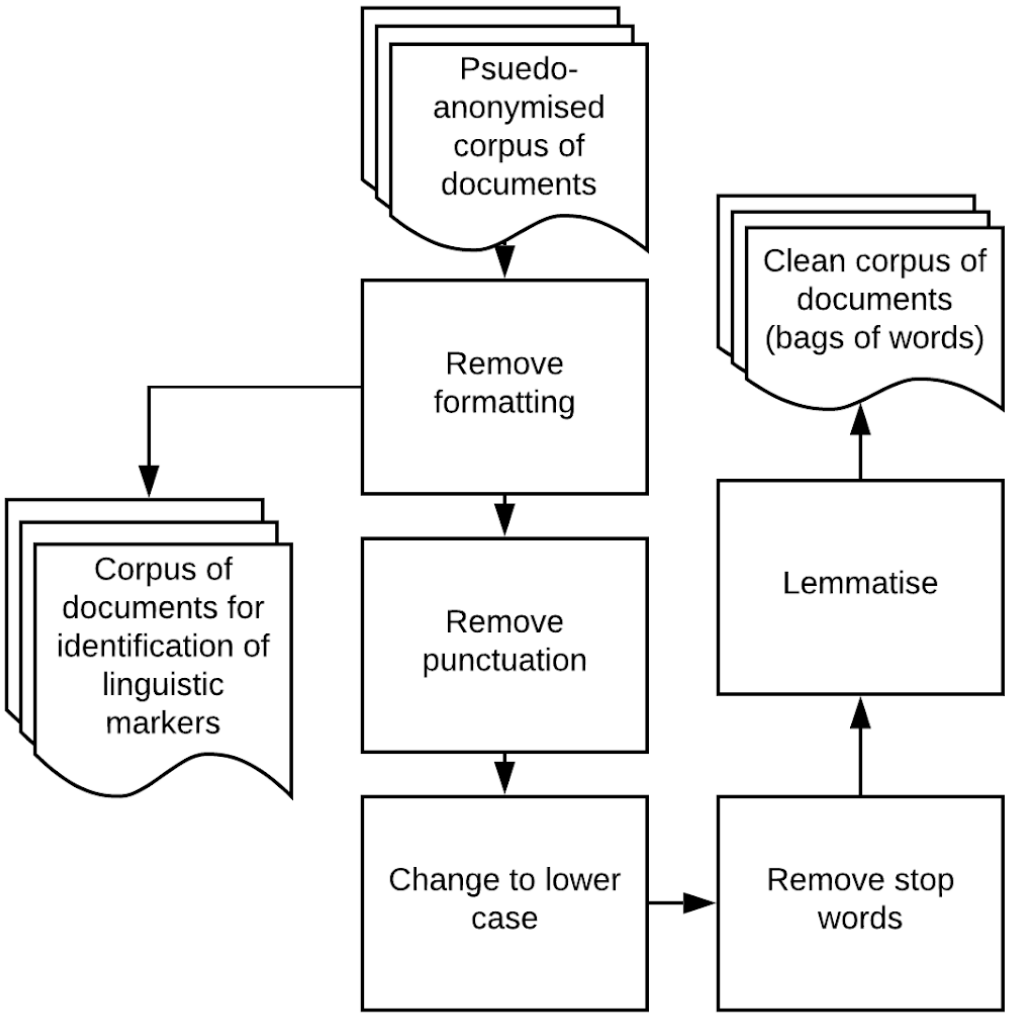


Figure 2: Text pre-processing steps

We then used several methodologies to extract information from the text:

- We analysed the frequency of words in a document relative to their frequency in documents relating to other cases (this is known as creating a 'term frequency inverse document frequency' (TFIDF) matrix). This gives an indication of which words distinguish the documents from documents about other cases. The columns of this TFIDF matrix became predictive features in our model.
- We modelled which topics best described the documents. To do so, we used a technique called Latent Dirichlet Allocation (LDA), which assumes that each document can be described by a distribution of topics and each topic can be described as a distribution of words. The topic proportions became predictive features in our model and we also explored the possibility of presenting the topics as word clouds to investigate whether they aided interpretability.
- We extracted linguistic features of the documents - whether the words in the document were associated with particular emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust). To do this

we used a lexicon (dictionary) provided by the National Research Council of Canada[10]: the Council had asked a large crowd of Mechanical Turk workers how much they thought that each word was with eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The value for the word provided in the lexicon was the majority judgement of this crowd of the public. We assigned these values to the words in the documents where they are available and average the value to get an aggregated value for the document. We took a similar approach to measuring the 'concreteness'[11] (vs abstractness) of a document. We were interested in measuring the concreteness of the text because previous (unpublished) work suggested an association between more abstract ways of describing the family's situation was associated with case escalation (the hypothesis being that abstraction reflects greater psychological distance between the social worker and the child / young person and their family and this may affect how well risk can be managed in collaboration with the family). To measure polarity (positive vs negative) and subjectivity (opinion vs fact), we use the Python library 'TextBlob'[12] which takes a slightly more sophisticated approach in that it handles different meanings of the same word in different contexts, modifiers (e.g. 'very', 'quite') and negation ('not bad', 'not great'). We included the length of text, sentiment and concreteness for LA1 and LA4 but we did not extract these features for LA2 and LA3 because all the text processing needed to happen onsite at the local authority and we had to make tradeoffs with the limited amount of onsite time we had.

- We counted the frequency of phrases which identify vulnerabilities as conceptualised by the Office for the Children's Commissioner [13](OCC). The OCC spoke with children and young people, as well as organisations who work with children, to help them choose which groups to include as 'vulnerable', resulting in a list of 72 ways of describing vulnerability. Some examples include: children in households where a parent is suffering from domestic abuse, severe mental health problems or substance addiction; children in gangs; children suffering from educational disadvantage and those outside mainstream education; children in poverty; children with caring responsibilities; children with special educational needs or disabilities. We exclude some categories which the OCC includes in their vulnerability report but which we felt would be inappropriate to include in prediction because they are protected characteristics, namely membership of the gypsy, Roma or traveller community or membership of the LGBTQ+ community. We identified vulnerabilities in the text for LA1, LA2 and LA4 but unfortunately ran out of time to do so onsite at LA3.

For LA2 and LA3, we had a limited amount of time onsite at the local authorities. During this time we processed the data to create a final dataset (identifying the population, creating features, pseudonymising), and then conducted the rest of the analysis (tuning of the models, and the creation of graphs and summary outputs) offsite using the final pseudonymised, structured dataset. Since it is very nearly impossible to perfectly anonymise the text data set, we agreed with the local authorities that all text would be processed within local authority IT systems to

10 For the emotion lexicon, see: National Research Council Canada. (2011, July). NRC Word-Emotion Association Lexicon (NRC Emotion Lexicon), http://web.stanford.edu/class/cs124/NRC-emotion-lexicon-wordlevel-alphabetized-v0.92.txt

11 For concreteness, see: Brysbaert M, Warriner AB, Kuperman V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. Behav Res Methods. 46(3):904-911.

12 Lorio, S. (2020) TextBlob: Tutorial: Quick Start. https://textblob.readthedocs.io/en/dev/quickstart.html

13 Office of the Children's Commissioner for England (2019) https://www.childrenscommissioner.gov.uk/vulnerability-in-numbers/

eliminate any risks arising from any potential data breach in the transfer or storage of the text data offsite. This required the transformation of all the text data into structured data (a TFIDF matrix as explained above) prior to the data being transferred offsite. For the models testing the inclusion of the text data alongside the structured data, we concatenated the TFIDF matrix onto the structured dataset (alongside the text features described above). This created a dataset which was much easier to check for identifying information, and which mitigated much of the risk of identification of individuals and their accompanying sensitive information if a data breach were to occur.

When conducting the analysis at the WWCSC offices for LA2 and LA3, we built topic models on the already constructed TFIDF matrix whilst optimising for average precision. The number of topics was set at the number of topics which gave the best results from a grid search over optimal hyperparameters involving just the TFIDF matrix creation and the Latent Dirichlet Allocation (LDA) as well as inspection of elbow plots (please see Appendix 1 under 'Elbow plots'). During the search for the best number of topics, we optimised for log likelihood as a metric instead of semantic coherence and exclusivity as we originally outlined in the research protocol. We had originally planned to do the topic modelling in R but decided that keeping the analysis in a single language was preferable. Unfortunately, semantic coherence and exclusivity are not available within Python's sklearn toolkit. For LA1 and LA4, we had access to the the text information in the form of a list of strings whilst optimising for the average precision, and so the transformation of the list of strings into a TFIDF matrix and the topic modelling took place within the pipeline and the selection of the optimal hyperparameters for TFIDF and LDA was conducted within this cross validation.

We had initially intended to explore different methodologies to model the topics of the documents - Structured Topic Models (STM) which allows the topics to vary with structured data and so allows the discussion of the same topic through different lenses e.g. child criminal exploitation may be discussed differently for children of different ages, as well as Latent Semantic Indexing (LSI) as an alternative way of reducing the dimensionality of the text data. However, subsequent consultation with natural language processing (NLP) experts suggested to us that these additional methods would not give much insight further to the LDA and so we prioritised other analyses instead. We initially were also interested in extracting the 'politeness' of the text also. 'Politeness' of the text field captures markers such as whether the notes give reasons, hedge or include qualifications to statements[14]. However, there is not a politeness library available in Python (the language in which we conducted our analysis) so we dropped this from our analysis also.

We were also initially interested in exploring the use of word embeddings (vectors of words which co-occur frequently with the word of interest in a separate large body of documents) as features. Word embeddings allow the model to learn what a word means from a much larger corpus of documents so that when the word appears in the documents at hand it can represent words which have similar meanings in a similar way. This can be helpful in that the same word can have very different meanings in different contexts and word embeddings can help avoid confusion between different meanings of the same word. However, given the extent to which the models learnt patterns which did not generalise when using the text data also, we thought that adding additional information without also increasing the number of observations would not improve model performance, and so we did not include text embeddings in our analysis.

14 For more information on politeness as a linguistic feature, please see the R package: Yeomans, M., Kantor, A. & Tingley, D. (2020, July 9). Package 'politeness', CRAN repository. https://cran.r-project.org/web/packages/politeness/politeness.pdf

Validation and Model Performance

We validated the models using cross-validation, a common methodology in machine learning to protect against models learning patterns which don't generalise well to unseen data (which 'overfit'). In cross validation, historical data is split multiple times into a subsample of the data to learn patterns about the cases in that subsample and then the model is tested on the remaining subsample of the historical data to assess whether the patterns learnt are generalisable to other cases. The simplest way to split the data is by randomly selecting cases but we required something more bespoke to take into account some of the complexities of the data in children's social care. Individuals are similar to their previous selves and siblings are similar to each other in multiple ways:

1. If a child is involved in multiple cases across time, there is likely to be a dependency in the outcomes of their cases across time.

2. If multiple children (siblings or other grouping) are involved in the same case (e.g. referral of multiple children after an incident), there is likely to be a dependency in their outcomes. It is relatively common for notes at contact and referral and assessment reports to be copied across for all the children involved in a case so there is likely to be a dependency with regard to input data too.

3. Even if multiple siblings are not referred at the same time, they likely share social or domestic circumstances.

In order to handle this lack of independence between some cases, we keep multiple cases associated with the same child, and cases of the child's siblings in the same cross-validation fold or in the holdout dataset. We handle siblings slightly differently at each local authority as we developed our approach as we went on and depending on what data was available:

- **LA1:** sibling groups were already identified in the case management system. Some children had multiple sibling groups and so we grouped together all sibling groups which shared a child.

- **LA2:** sibling groups were already identified in the case management system and were unique so not further action required.

- **LA3:** sibling groups were not identified in the case management system. 'Siblings' or at least children with related cases were identified as those with the same contact date and length of contact or referral text (indicating that the text had been copied across). Some children had multiple sibling groups so we assigned the child to the largest sibling group they were part of. There is a small risk of some siblings not being identified as such in LA3 where they did not share a contact or referral.

- **LA4:** children related to the child were already identified in the case management system and these children formed the same sibling group.

We made sure that each child had only sibling ID so that when grouping together cases by the sibling ID we also grouped together the child's own cases. Children without sibling IDs were assigned their own unique sibling ID.

Beyond the dependence of cases about the same child and their siblings across time, we also may expect some violation of independence for cases of non-related children across time. In particular, we may expect the model to perform artificially well if we allow the model to learn from cases which occur chronologically after the case the model is making a prediction on. This is because:

- The decisions made by social workers on later cases may reflect learnings from earlier cases;

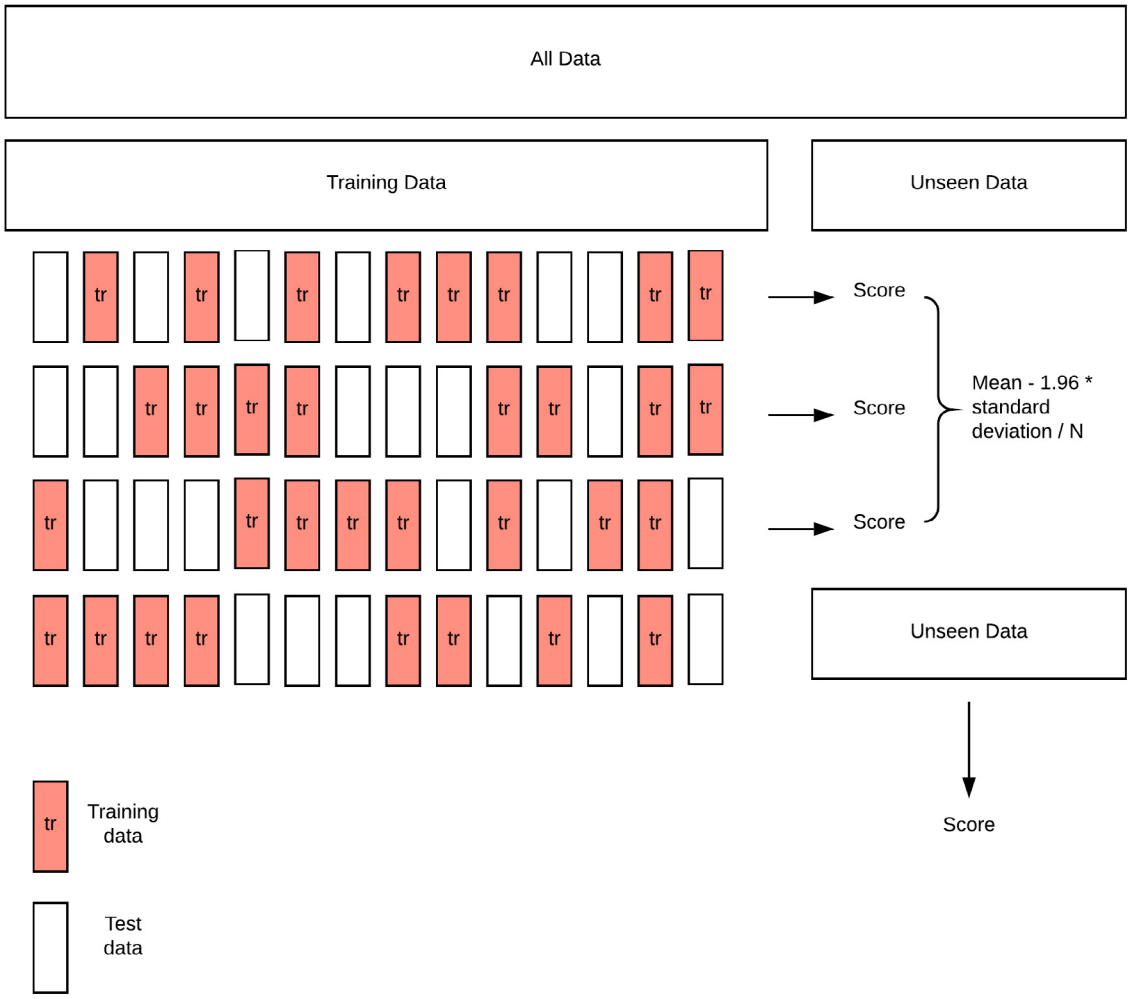
- Trends of risk and protective factors (and our understanding of them) evolve over time.

To understand whether it is important to take into account the ordering of cases in time when validating a model, we compared the model performance of models which can learn

patterns from any case irrespective of when the case occurred in relation to the case at hand, and models restricted to only learning patterns from cases earlier than the case at hand. We conducted the comparison of models which are and are not restricted in this way instead of taking it into account for all models, as we had initially outlined in the research protocol.

Both forms of cross validation ensure that the same sibling group will not appear in two different cross validation folds. Imposing the extra restriction of only learning from earlier cases on the data in addition to the group restrictions means that the cross validation folds risk having a very small sample of cases to learn from. Because of this, we split the data into three when validating the models restricted to learning from earlier cases and five when they are not.

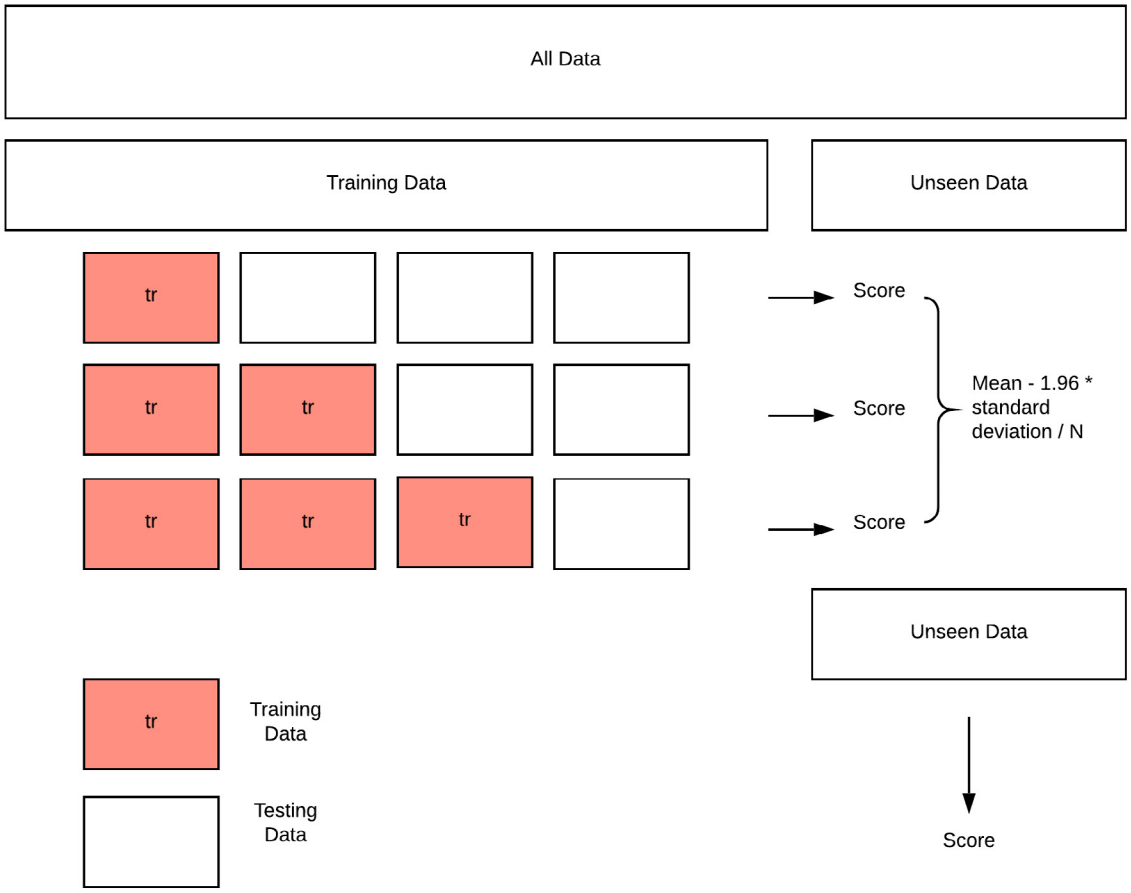
Please compare Figures 3 and 4 to see how the historical data is split under the different methodologies. (We ignore how we take into account siblings in these diagrams to focus on the comparison between a model restricted to learning from earlier cases and one not restricted in this way).



N = number of cross-validation folds

Figure 3: How the data was split to allow the model to learn from any case irrespective on when it occurred

Under this type of cross-validation, we take all the historical data and split it into training and unseen data. We save predicting on the unseen data as the final test. The training data is further split into folds randomly selected and the model is trained on a subsample (the vermillion folds) and tested on the remaining data (the white folds). This gives multiple test scores and allows for models to be compared on how well they predict on average but also how consistently they perform on different datasets.



N = number of cross-validation folds

Figure 4: How the data was split to restrict the model to learning only from previous cases

Under this type of cross-validation, we take all the historical data and split it into training and unseen data. We save predicting on the unseen data as the final test. The training data is further split into folds and the model is trained on a subsample of earlier cases (the vermillion folds) and tested on the remaining later (the white folds). This gives multiple test scores and allows for models to be compared on how well they predict on average but also how consistently they perform on different datasets.

We optimised for the average precision which allowed us to balance the trade-off 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.
- A model with **high recall** which identifies most of the cases at risk of the outcome of interest.

There is an inherent tradeoff between the two because being more confident that the model is correct when it predicts a case at risk requires setting the threshold for what 'counts' as at risk pretty high, which means identifying cases with a lower probability of the outcome as not at risk, and hence missing more borderline cases.

We searched randomly within a set range to select optimal hyperparameters (parameters which control the complexity of the model). During the randomised search cross validation, we maximised the average precision instead of maximising the average Area Under the Curve

(AUC) of the precision-recall graph as specified in the research protocol. Average precision is the mean of precisions achieved at each threshold, weighted by the increase in recall from the previous threshold. Both the AUC of the precision-recall graph and average precision summarise the trade-off between precision and recall at different thresholds and are scale invariant (so prioritise getting the ranking of cases correct instead of absolute value) but research we undertook after the publishing of the research protocol suggests that the AUC of precision-recall graph can be too optimistic.[15]

Often the model which has the highest average performance during validation is selected as the version of the model with best hyperparameters and the one to test on holdout data and put into practice (if appropriate). We instead chose the model which had the highest 95% lower confidence bound[16] to predict on holdout data. We used this decision rule instead of simply picking the maximum mean average precision score because the relatively small sample size and the small proportion of cases of children at risk of the outcome observed in the data restricts us to a small number of cross-validation folds and means that the test folds contain small numbers of positive observations. These factors decrease our confidence that the model with the highest mean will generalise well. Choosing the highest lower confidence bound allows us to incorporate our uncertainty about the distribution of test scores into the decision rule on which model performs the 'best' and should be used to predict on holdout data. We specified in the research protocol that we would monitor the variance of the score in cross validation but had not formalised how at the time of writing the research protocol.

We evaluated the average precision of the final model chosen on unseen data ('holdout data'). This approach is a core approach in building machine learning models because it allows us to simulate how well the model would perform if it were deployed in practice in a safe 'sandbox' setting and enables rapid feedback to allow for improving the model.

We report the performance of all the models (please see 'Performance Metrics' in Appendix 1. Please note that we exclude the model performance from LA1 (outcomes 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 comparing the eight different models to each other is still helpful so we include the metrics from the LA1 models in the comparative analysis.

Modelling

Algorithms

We tested three types of algorithms (decision tree, logistic regression and gradient boosting algorithm) to understand where the tradeoff lay between interpretability and model performance: decision trees are highly interpretable but tend to have lower predictive accuracy whilst gradient boosting is a more complex tree-based model likely to have higher predictive accuracy. As the models tended to overfit we used regularisation (both ridge and LASSO) in logistic regression to try to reduce the risk of overfitting.

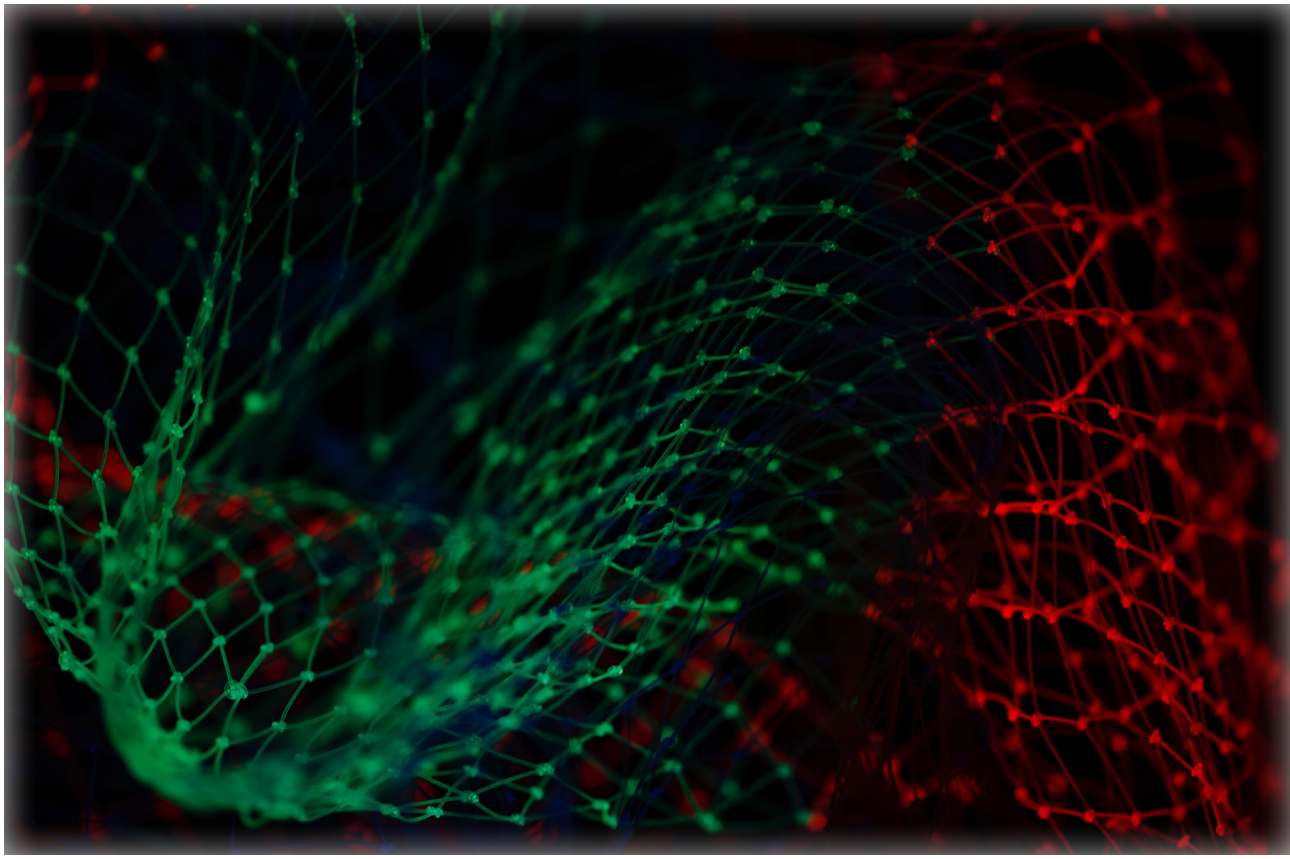
Other Choices about the Pipeline within the Cross-Validation Search

We did not outline in the research protocol choices within the cross-validation pipeline which can affect the model performance. All of the datasets had high levels of class imbalance (ranging from 2%-18%). To ensure a minimum number of observations for whom the outcome of interest was observed, we included stratification in the cross-validation splitting for the models ignoring the restriction on only learning from

15 Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html

16 Mean - (1.96 * standard deviation / square root (number of cross-validation folds))

earlier cases. We did not stratify by outcome during the cross validation where the model was restricted to only learning from earlier cases. This was because the classes were not distributed throughout time in a way that facilitated this. We also downsampled the majority class (those not at risk of the outcome) using one sided sampling which removes noisy examples and is a suitable sampling method for scenarios with high levels of class imbalance.



RESULTS

Practical Question 1: How easy is it to extract data from the case management system and get it in the required format and of sufficient quality for the model?

In order to extract the structured data, local authorities business intelligence teams adapted reports already built to deliver the Annex A datasets. There was some customisation needed to select the appropriate sample, the appropriate columns and pseudonymise the data before making it available to WWCSC researchers. Where the selection of the sample was done at the stage of data extraction, this was considerably more complex than running reports as usual because we were interested in historical information on children who had experienced a particular outcome within a certain time period (e.g. entered care from April 2018 to March 2019), requiring a selection of cases by child identifiers whilst reporting for Ofsted and the Department is usually selecting a population within a date range. The query languages for case management systems do not seem set up to handle this type request.

At least three of the children's services intelligence teams 'commissioned' extra support from internal local authority IT teams to assist with the extraction of the records, assessments and reports given that their usual reporting tasks do not require access to this type of data. In some cases, this involved IT teams writing custom SQL queries.

There was considerable variation in the time it took to extract the data, based on whether the teams had reports set up already that could easily be adapted and how easy it was for intelligence team analyst time to be redeployed from existing tasks. A median estimate of approximately half a day of IT team time and 1-2 days of intelligence team time provides an indication of resources required for data extraction. A practical consideration which influences the feasibility of this type of work is which teams hold access to which parts of the database. In some cases, IT team time was chargeable for the extraction of the data. It is worth noting that considerable additional data infrastructure would be required if the model were to be used in practice because the model would need to be connected to the live case management system (instead of a static extract from the system). From our understanding, one of the local authority partners was in the process of setting up the data infrastructure that would be required but that it would be an additional investment for the other local authority partners.

With regard to data access, the first local authority provided remote access to the data on a virtual machine via a secure VPN. Data access was provided onsite at three of the local authorities to allow the WWCSC researchers to process the text data which comes with a higher level of risk of identification in case of a data breach. For two of the local authorities, once the text data had been processed (this involved using natural language processing techniques to anonymise the documents and then converting it into a structured data as described above, for example, counting the frequency of words) and the risk of de-anonymisation significantly reduced, the remaining analysis was conducted at WWCSC offices on a WWCSC laptop and the data stored securely on an encrypted hard drive in an electronic safe. Analysis at the remaining local authority was underway when the Covid-19 outbreak led to lockdown - under these circumstances permission was granted for the WWCSC researcher to work from home on the local authority issued laptop.

The data extracts relating to different parts of the social care journey, for example, contacts, referrals etc. were made available to the WWCSC researchers in the form of separate Excel or csv files. It was then relatively straightforward to merge together the files using the unique child identifiers and dates for different stages but in some cases we needed to do a 'fuzzy match' data based on date ranges when the datasets did not share a common date which uniquely identified the observations along with the child identifier. As with any similar exercise, there was a considerable amount of 'data cleaning' to get the data in a format ready for the model.

Practical Question 2: What skills and hardware do you need to carry out this type of analysis?

Skills

As mentioned above, at some of the local authorities, the data was extracted from the case management system using SQL and required a request to the IT team. At other local authorities, analysts were able to extract the data from the case management system using existing queries or via the query system they are familiar with.

For the analysis (conducted by the WWCSC researchers), familiarity with data cleaning, machine learning and natural language processing was required. The language used was Python but it would have been possible to complete the analysis in other languages also e.g. R. Both Python and R are open source and freely available but do require quite a bit of experience to be able to use for these techniques. Because of some adaptations to 'off the shelf' methods, it would not be possible to complete this analysis in 'point and click' or 'drag and drop' software tools such as Microsoft Azure Machine Learning Studio.

Hardware

The analysis took place on standard issue computers with 8GB installed RAM and an Intel Xeon Gold 6152 processor or an Intel Core i5 processor. The average training times for each individual fit of the model are reported in Appendix 1 under 'Training Time'. Mostly the training and validation of the model with 50 iterations would complete within a few hours. We ran the anonymisation and the extraction of features from the text overnight as these processes took considerably longer and slowed down the machines substantially when running in the background. Using standard issue computers was very workable when the option to run code overnight was available, and this was possible with all of the local authorities.

Practical Question 3: What is the level of anonymisation of text data achievable by automated means?

Text data includes large amounts of personally identifying information. To use text data in a model, we need to pseudonymise the text in order to reduce the risk of identification in case of a data breach, and also to reduce 'noise' added by the personally identifying information (which affects the performance of the model).

- We requested a list of names for children / young people, family members and professionals in the dataset (where available).
- We tagged the parts of speech in the documents which labels words by their syntactic functions. Through this, we identified proper nouns - words which are likely to be names - and we also looked for patterns which matched the typical structure of addresses, phone numbers and email addresses.

- We used a publicly available list of common first names[17] to further identify any names to be removed.
- We removed this identifying information found and any words sufficiently similar to take into account spelling mistakes.[18]

In all of the local authorities, the name of the child was available. We derived the names of siblings through grouping together the names of children with the same sibling group identifier. In LA2, the names of the parents / carers were available. In LA1, LA3 and LA4, a list of social workers involved in the case was available. LA4 additionally provided a list of names of individuals related to the child / young person and how they were related (parent, sibling etc). None of the local authorities kept a list of nicknames or alternative names, which made it difficult to identify nicknames or alternative names that were sufficiently different from the full name (e.g. 'Abby' instead of 'Abigail'), however, most of these alternative versions also appeared in the publicly available list of first names.

A list of surnames in England is not publicly available and was not available on request from the Office for National Statistics.

In addition to the methods outlined above, we also treated all instances of capitalised words after first names which had been replaced with the 'ChildName' or 'OtherName' placeholders as surnames and replaced with the placeholder 'Surname'. Words following honorific titles (e.g. 'Mr', 'Mrs' etc) were also replaced with a 'Surname' placeholder.

In addition to removing names, addresses, phone numbers and emails as specified in the research protocol, we identified and replaced street names, postcodes, school names, dates (to remove dates of birth) and identifying numbers (e.g. NHS numbers).

We confirmed that these methods identify and replace non-Roman characters also.

Please see the Github repository for the Anonymisation code (*text_functions.py* and *Anonymisation.py*).

The pseudonymisation process is described in the figure below.

17 Office for National Statistics. Baby names in England and Wales, 1996 to 2016. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/babynamesenglandandwales/previousReleases

18 What counts as 'sufficiently similar" was mostly a matter of trial and error.

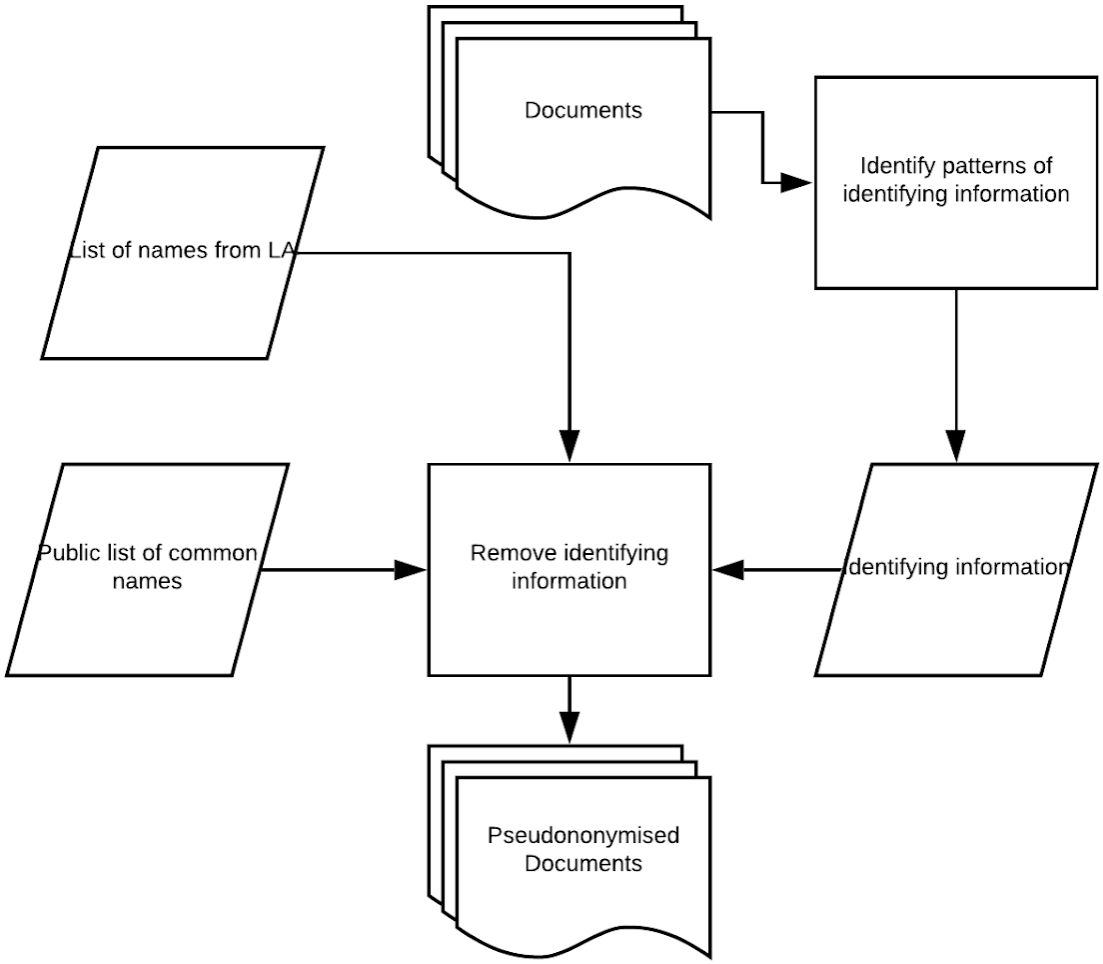


Figure 5: Pseudonymisation of text data

We conducted random checks to evaluate how well the text documents were pseudonymised. For illustration purposes: for LA4, the results for a random selection of 15 documents were as follows:

- No children's or siblings' names left in any of the documents;
- 9 / 15 documents inspected contained no identifying information at all;
- 4 / 15 documents inspected contained a name or partial name of a professional;
- 2 / 15 documents inspected contained the email address of a professional.

It is promising that no children's identifying information or siblings' identifying information remained. In the cases where the name or partial name of the professional were not identified, this was because the name or email address had merged into the following word when a space had been missed during typing. It is difficult to identify these names because the compound words are not sufficiently similar to match the list of names they are compared to. We handled such cases through:

1. Feature selection when training and validating the model. Personally identifying information is 'noise' from the perspective of the model and likely to be discarded at this stage.

2. We involved multiple individuals in conducting checks on all of the outputs prior to publishing to check that there is no identifying information displayed within the figures or the tables. These checks for identifying information (arising from the text data) were carried out in addition to the usual statistical disclosure checks conducted for the presentation of analysis on structured data. We expect the risk of identification to be very low after these checks. This quite manual process for the purposes of this publication would not be required if a tool using predictive models were to be developed as such a tool would be used by social workers already involved in the case and so no risk would arise from including identifying information in the outputs.

Overall, we concluded that the anonymisation is good enough for our purposes (reducing the 'noise' from personal data fed into the model) but would not be good enough for automatic redaction of personal information to replace the full redaction social workers undertake when required for sharing documents. It may be that such automatic redaction could give a 'head start', which the social worker could then review and amend although it is rare for social workers to be required to redact documents for identifying information on the scale needed for this project.

Technical Research Question 1: What is the performance of the models using structured data (i.e. data that would be recorded in a statutory return like risk factors)?

Technical Research Question 2: What is the performance of the models using structured and text data from assessment and referral reports?

Research questions 1-2 give an indication of whether the models would predict the outcome sufficiently well to be used in practice and the value added by including text from records, reports and assessments. Whether including text adds much predictive power is useful to know because it allows us to understand the tradeoff with the challenges of using text. As mentioned above, it is considerably more tricky to extract from databases under the current set up of case management systems (at least from the experience of our partner local authorities). Text is much more difficult to anonymise than structured data and so including text data presents a larger data governance risk. Text analysis is also an additional skill set.

Although in our research protocol we stated that we would cross validate restricting the model to learn only from earlier cases; in the course of conducting the analysis we thought that it would be helpful to report model performance for models validated with this restriction and those without to illustrate whether such a restriction affected the scores.

For each outcome predicted, we built four models to compare the inclusion and exclusion of text data, and the restriction to learning from earlier cases or no restriction:

	Ignoring the restriction on learning from earlier cases in validating and evaluating the model		Restricting the model to learning from earlier cases in validating and evaluating the model	
Including structured data	 Learned from all cases	 Structured data only	 Learned only from earlier cases	 Structured data only
Including structured and text data	 Learned from all cases	 Includes text data	 Learned only from earlier cases	 Includes text data

For each outcome predicted, we describe the population, the outcome predicted, the sample size, the years the data is drawn from and which algorithm performed best. We provide tables of:

- **Proportion of the population not at risk from the outcome:** this gives an indication of how rare the outcome is in the population. Rare outcomes are difficult to predict because you have few examples for the model to learn from. They provide a baseline for model performance: if 95% of the population are not at risk of the outcome, the simplest (but an unhelpful) model would predict that no cases were at risk of the outcome and be accurate 95% of the time.
- **Accuracy:** it is instructive to compare the accuracy score to the proportion of the population not at risk of the outcome. The accuracy score should be greater than a baseline model predicting no cases at risk as described above.
- **Average precision score on unseen data:** this is our main metric of interest because it captures well the tradeoff between a precise model (few false alarms) and one with high recall (few children missed) on datasets with a high proportion of the population not at risk from the outcome. Testing how well the model does on unseen data allows us to get a sense of how well the model generalises to new cases.
- **'Area under the curve' (AUC) on unseen data:** we report this because it is a common metric used by data teams building models in this and analogous sectors, and so it is useful for comparison purposes. We do not count it as our main metric because the AUC rewards the correct identification of cases not at risk of the outcome, a relatively easy task given that these cases make up the vast majority of the population.
- **Mean and standard deviation of the average precision during cross validation:** these statistics give a sense of how well the model performs when exposed to different datasets - whether the patterns it has learnt generalise well.
- Whether the mean performance metrics during validation or performance metrics on the holdout data exceed the threshold we set for 'success' in the research protocol, above 0.65.

Accuracy, average precision and AUC are all measured on a scale of 0-1 with 0 being the worst possible model and 1 being the best. The comparing cross-validation table (which compares models restricted to learning from earlier cases and those not) and the comparing data included table (which compares models using only structured data, and both structured and text data) present the same information but facilitate different comparisons: the former showing whether restricting the model to learning from earlier cases harms its performance and the

latter showing whether adding text data improves the model. We provide some commentary below the tables for each outcome predicted.

We then discuss the patterns emerging about the type when considering the scores from all the outcomes predicted together.

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)?

Description

Outcome and population Table 1: outcome, population, sample size, years of data available and best algorithm to predict escalation to CPP or CLA after 'No further action'

Outcome	Case escalates to a Child Protection Plan (CPP) or the child being Looked After (CLA) in 12 months
Population	Case designed 'No further action' (NFA) within previous 12 months
Sample size	6544 cases
Years	April 2016 - June 2019
Best algorithm	Gradient boosting

Overall results

Comparing Cross-validation Table 1: Predicting escalation to CPP or CLA after 'No further action': comparison of performance metrics for models learning from all cases or restricted to learning only from earlier cases

Metrics	Data included	Learning from all cases	Learning only from earlier cases	Validation technique with best model performance	Greater than 'success' threshold (0.65) - learning from all cases	Greater than 'success' threshold (0.65) - learning only from earlier cases
Proportion not at risk of outcome	Structured and text data	0.94	0.94			
Proportion not at risk of outcome	Structured data only	0.94	0.94			
Accuracy	Structured and text data	0.94	0.94	Same performance	Exceeds success threshold	Exceeds success threshold
Accuracy	Structured data only	0.94	0.93	Learning from all cases	Exceeds success threshold	Exceeds success threshold

Average precision	Structured and text data	0.54	0.54	Same performance	Does not exceeds success threshold	Does not exceeds success threshold
Average precision	Structured data only	0.51	0.47	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Structured and text data	0.79	0.5	Learning from all cases	Exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Structured data only	0.75	0.44	Learning from all cases	Exceeds success threshold	Does not exceeds success threshold
Std average precision (training)	Structured and text data	0.07	0.12	Learning from all cases		
Std average precision (training)	Structured data only	0.04	0.06	Learning from all cases		
AUC	Structured and text data	0.94	0.95	Learning only from earlier cases	Exceeds success threshold	Exceeds success threshold
AUC	Structured data only	0.95	0.95	Same performance	Exceeds success threshold	Exceeds success threshold

Comparing data included Table 1: Predicting escalation to CPP or CLA after 'No further action': comparison of performance metrics for models including just structured data and structured and text data

Metrics	Cross Validation Method	Structured data only	Structured and text data	Data included with best model performance	Greater than 'success' threshold (0.65) - structured data only	Greater than 'success' threshold (0.65) - structured and text data
Proportion not at risk of outcome	Learning from all cases	0.94	0.94			
Proportion not at risk of outcome	Learning only from earlier cases	0.94	0.94			
Accuracy	Learning from all cases	0.94	0.94	Same performance	Exceeds success threshold	Exceeds success threshold
Accuracy	Learning only from earlier cases	0.93	0.94	Structured and text data	Exceeds success threshold	Exceeds success threshold

Average precision	Learning from all cases	0.51	0.54	Structured and text data	Does not exceeds success threshold	Does not exceeds success threshold
Average precision	Learning only from earlier cases	0.47	0.54	Structured and text data	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Learning from all cases	0.75	0.79	Structured and text data	Exceeds success threshold	Exceeds success threshold
Mean average precision (training)	Learning only from earlier cases	0.44	0.50	Structured and text data	Does not exceeds success threshold	Does not exceeds success threshold
Std average precision (training)	Learning from all cases	0.04	0.07	Structured data only		
Std average precision (training)	Learning only from earlier cases	0.06	0.12	Structured data only		
AUC	Learning from all cases	0.95	0.94	Structured data only	Exceeds success threshold	Exceeds success threshold
AUC	Learning only from earlier cases	0.95	0.95	Same performance	Exceeds success threshold	Exceeds success threshold

The average precision score exceeds the threshold for a 'successful' model under the criteria defined in the research protocol for two out of the four models during validation but none of the four models on the holdout dataset. All four of the models exceed the threshold for AUC. However, as mentioned, please note that subsequent analysis indicated likely information leakage, and these scores are likely to be artificially inflated and so we do not treat these findings of the scores exceeding the threshold as evidence for machine learning working children's social care.

Do the models perform as well when restricting them to learn from just earlier cases?

Average precision is worse for models restricted to learning from just earlier cases in validation and on holdout data when including just the structured data. Average precision is worse for models restricted to learning from just earlier cases in validation but the same on holdout data when including structured and text data. The standard deviation of the average precision is higher for models restricted to learning from just earlier cases.

Does adding text improve the results?

Adding text adds marginal improvement to model score for models learning from all cases and models restricted to learning from just earlier cases both during validation and on holdout data; however the standard deviation is higher with the text added, suggesting that the models using text data may be overfitting.

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

Description

Outcome and population Table 2: outcome, population, sample size, years of data available and best algorithm to predict escalation to CPP or CLA after contact

Outcome	Case escalates to a Child Protection Plan (CPP) or the child being Looked After (CLA) within 6 to 12 months
Population	All children and young people at the point of referral
Sample size	17560 cases
Years	April 2016 - June 2019
Best algorithm	Gradient boosting

Overall results

Comparing Cross-validation Table 2: Predicting escalation to CPP or CLA after contact: comparison of performance metrics for models learning from all cases or restricted to learning only from earlier cases

Metrics	Data included	Learning from all cases	Learning only from earlier cases	Validation technique with best model performance	Greater than 'success' threshold (0.65) - learning from all cases	Greater than 'success' threshold (0.65) - learning only from earlier cases
Proportion not at risk of outcome	Structured and text data	0.98	0.98			
Proportion not at risk of outcome	Structured data only	0.98	0.98			
Accuracy	Structured and text data	0.97	0.97	Same performance	Exceeds success threshold	Exceeds success threshold
Accuracy	Structured data only	0.97	0.97	Same performance	Exceeds success threshold	Exceeds success threshold
Average precision	Structured and text data	0.16	0.12	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold

Average precision	Structured data only	0.17	0.14	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Structured and text data	0.68	0.37	Learning from all cases	Exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Structured data only	0.70	0.38	Learning from all cases	Exceeds success threshold	Does not exceeds success threshold
Std average precision (training)	Structured and text data	0.08	0.04	Learning only from earlier cases		
Std average precision (training)	Structured data only	0.07	0.02	Learning only from earlier cases		
AUC	Structured and text data	0.91	0.90	Learning from all cases	Exceeds success threshold	Exceeds success threshold
AUC	Structured data only	0.92	0.90	Learning from all cases	Exceeds success threshold	Exceeds success threshold

Comparing data included Table 2: Predicting escalation to CPP or CLA after contact: comparison of performance metrics for models including just structured data and structured and text data

Metrics	Cross Validation Method	Structured data only	Structured and text data	Data included with best model performance	Greater than 'success' threshold (0.65) - structured data only	Greater than 'success' threshold (0.65) - structured and text data
Proportion not at risk of outcome	Learning from all cases	0.98	0.98			
Proportion not at risk of outcome	Learning only from earlier cases	0.98	0.98			
Accuracy	Learning from all cases	0.97	0.97	Same performance	Exceeds success threshold	Exceeds success threshold
Accuracy	Learning only from earlier cases	0.97	0.97	Same performance	Exceeds success threshold	Exceeds success threshold
Average precision	Learning from all cases	0.17	0.16	Structured data only	Does not exceeds success threshold	Does not exceeds success threshold

Average precision	Learning only from earlier cases	0.14	0.12	Structured data only	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Learning from all cases	0.70	0.68	Structured data only	Exceeds success threshold	Exceeds success threshold
Mean average precision (training)	Learning only from earlier cases	0.38	0.37	Structured data only	Does not exceeds success threshold	Does not exceeds success threshold
Std average precision (training)	Learning from all cases	0.07	0.08	Structured data only		
Std average precision (training)	Learning only from earlier cases	0.02	0.04	Structured data only		
AUC	Learning from all cases	0.92	0.91	Structured data only	Exceeds success threshold	Exceeds success threshold
AUC	Learning only from earlier cases	0.90	0.90	Same performance	Exceeds success threshold	Exceeds success threshold

The average precision score exceeds the threshold for a 'successful' model under the criteria defined in the research protocol for two out of the four models during validation but none of the four models on the holdout dataset. All four of the models exceed the threshold for AUC. However, please note that subsequent analysis indicated likely information leakage, and these scores are likely to be artificially inflated and so we do not treat these findings of the scores exceeding the threshold as evidence for machine learning working children's social care.

Do the models perform as well when restricting them to learn from just earlier cases?

Average precision in validation is approximately half for models restricted to learning from just earlier cases when including just the structured data or both the structured and text data but is similar (but still worse) on the holdout data. However, models restricted to learning from just earlier cases have lower standard deviation of average precision in validation.

Does adding text improve the results?

Adding text marginally worsens the model score for both models trained on earlier cases and all cases both during validation and on holdout data. The standard deviation during validation is also marginally higher when adding in text data for models under both types of validation.

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'?

Description

Outcome and population Table 3: outcome, population, sample size, years of data available and best algorithm to predict open case after 'No further action'

Outcome	Case open to children's social care
Population	Children / young people at the point of referral
Sample size	6450 cases (model not restricted to only learning from earlier cases) 6885 cases (model not restricted to only learning from earlier cases)
Years	April 2015 - July 2019
Best algorithm	Gradient boosting

Note: the sample sizes differ by the type of cross validation because observations are dropped according to different criteria to handle the different restrictions

Overall results

Comparing Cross-validation Table 3: Predicting open case after 'No further action': comparison of performance metrics for models learning from all cases or restricted to learning only from earlier cases

Metrics	Data included	Learning from all cases	Learning only from earlier cases	Validation technique with best model performance	Greater than 'success' threshold (0.65) - learning from all cases	Greater than 'success' threshold (0.65) - learning only from earlier cases
Proportion not at risk of outcome	Structured and text data	0.94	0.98			
Proportion not at risk of outcome	Structured data only	0.94	0.98			
Accuracy	Structured and text data	0.95	0.98	Learning only from earlier cases	Exceeds success threshold	Exceeds success threshold
Accuracy	Structured data only	0.95	0.98	Learning only from earlier cases	Exceeds success threshold	Exceeds success threshold

Average precision	Structured and text data	0.41	0.03	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
Average precision	Structured data only	0.38	0.03	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Structured and text data	0.51	0.17	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Structured data only	0.45	0.18	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
Std average precision (training)	Structured and text data	0.06	0.05	Learning only from earlier cases		
Std average precision (training)	Structured data only	0.05	0.06	Learning from all cases		
AUC	Structured and text data	0.75	0.58	Learning from all cases	Exceeds success threshold	Does not exceeds success threshold
AUC	Structured data only	0.70	0.59	Learning from all cases	Exceeds success threshold	Does not exceeds success threshold

Comparing data included Table 3: Predicting open case after 'No further action': comparison of performance metrics for models including just structured data and structured and text data

Metrics	Cross Validation Method	Structured data only	Structured and text data	Data included with best model performance	Greater than 'success' threshold (0.65) - structured data only	Greater than 'success' threshold (0.65) - structured and text data
Proportion not at risk of outcome	Learning from all cases	0.94	0.94			
Proportion not at risk of outcome	Learning only from earlier cases	0.98	0.98			
Accuracy	Learning from all cases	0.95	0.95	Same performance	Exceeds success threshold	Exceeds success threshold
Accuracy	Learning only from earlier cases	0.98	0.98	Same performance	Exceeds success threshold	Exceeds success threshold

Average precision	Learning from all cases	0.38	0.41	Structured and text data	Does not exceeds success threshold	Does not exceeds success threshold
Average precision	Learning only from earlier cases	0.03	0.03	Same performance	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Learning from all cases	0.45	0.51	Structured and text data	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Learning only from earlier cases	0.18	0.17	Structured data only	Does not exceeds success threshold	Does not exceeds success threshold
Std average precision (training)	Learning from all cases	0.05	0.06	Structured data only		
Std average precision (training)	Learning only from earlier cases	0.06	0.05	Structured and text data		
AUC	Learning from all cases	0.70	0.75	Structured and text data	Exceeds success threshold	Exceeds success threshold
AUC	Learning only from earlier cases	0.59	0.58	Structured data only	Does not exceeds success threshold	Does not exceeds success threshold

None of the average precisions of the four models exceed the threshold for a 'successful' model under the criteria defined in the research protocol in validation or on holdout data. Two of the four models exceed the threshold for AUC.

Do the models perform as well when restricting them to learn from just earlier cases?

Average model performance is substantially worse for models restricted to learning from just earlier cases when including just the structured data or both the structured and text data both in validation and on the holdout data. The standard deviation of average precision during validation is similar for models validated using either methodology. The model performance drops off more substantially when predicting on holdout data for the models restricted to learning from just earlier cases.

Does adding text improve the results?

Adding text improves the model performance for models not restricted to just learning from earlier cases but adds no improvement for models or slightly worsens the score for models restricted to learning from earlier cases on holdout data and in validation respectively. The standard deviation in validation is very similar irrespective of whether text data is included or not.

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?

Description

Outcome and population Table 4: outcome, population, sample size, years of data available and best algorithm to predict escalation to CPP, CLA, RO or SGO after open case

Outcome	Case escalates to the child being on a Child Protection Plan, being Looked After, being adopted, Residence Order, Special Guardianship Order) between 3 months and 2 years of the referral start date
Population	Children / young people open to children's social care (all assessments within 2 weeks minus 1 day of the referral date)
Sample size	9877 cases (model not restricted to learning from earlier cases) 10625 cases (model restricted to learning from earlier cases)
Years	April 2015 - July 2019
Best algorithm	Gradient boosting

Note: the sample sizes differ by the type of cross validation because observations are dropped according to different criteria to handle the different restrictions

Overall results

Comparing Cross-validation Table 4: Predicting escalation to CPP, CLA, RO or SGO after open case: comparison of performance metrics for models learning from all cases or restricted to learning only from earlier cases

Metrics	Data included	Learning from all cases	Learning only from earlier cases	Validation technique with best model performance	Greater than 'success' threshold (0.65) - learning from all cases	Greater than 'success' threshold (0.65) - learning only from earlier cases
Proportion not at risk of outcome	Structured and text data	0.98	0.99			
Proportion not at risk of outcome	Structured data only	0.98	0.99			
Accuracy	Structured and text data	0.98	0.98	Same performance	Exceeds success threshold	Exceeds success threshold

Accuracy	Structured data only	0.98	0.99	Learning only from earlier cases	Exceeds success threshold	Exceeds success threshold
Average precision	Structured and text data	0.42	0.09	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
Average precision	Structured data only	0.36	0.23	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Structured and text data	0.55	0.38	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Structured data only	0.44	0.34	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
Std average precision (training)	Structured and text data	0.06	0.10	Learning from all cases		
Std average precision (training)	Structured data only	0.13	0.08	Learning only from earlier cases		
AUC	Structured and text data	0.90	0.88	Learning from all cases	Exceeds success threshold	Exceeds success threshold
AUC	Structured data only	0.90	0.87	Learning from all cases	Exceeds success threshold	Exceeds success threshold

Comparing data included Table 4: Predicting escalation to CPP, CLA, RO or SGO after open case: comparison of performance metrics for models including just structured data and structured and text data

Metrics	Cross Validation Method	Structured data only	Structured and text data	Data included with best model performance	Greater than 'success' threshold (0.65) - structured data only	Greater than 'success' threshold (0.65) - structured and text data
Proportion not at risk of outcome	Learning from all cases	0.98	0.98			
Proportion not at risk of outcome	Learning only from earlier cases	0.99	0.99			
Accuracy	Learning from all cases	0.98	0.98	Same performance	Exceeds success threshold	Exceeds success threshold

Accuracy	Learning only from earlier cases	0.99	0.98	Structured data only	Exceeds success threshold	Exceeds success threshold
Average precision	Learning from all cases	0.36	0.42	Structured and text data	Does not exceeds success threshold	Does not exceeds success threshold
Average precision	Learning only from earlier cases	0.23	0.09	Structured data only	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Learning from all cases	0.44	0.55	Structured and text data	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Learning only from earlier cases	0.34	0.38	Structured and text data	Does not exceeds success threshold	Does not exceeds success threshold
Std average precision (training)	Learning from all cases	0.13	0.06	Structured and text data		
Std average precision (training)	Learning only from earlier cases	0.08	0.1	Structured data only		
AUC	Learning from all cases	0.90	0.90	Same performance	Exceeds success threshold	Exceeds success threshold
AUC	Learning only from earlier cases	0.87	0.88	Structured and text data	Exceeds success threshold	Exceeds success threshold

None of the four models exceed the threshold for a 'successful' model under the criteria defined in the research protocol during validation or on holdout data. The AUC exceeds the threshold on all four models.

Do the models perform as well when restricting them to learn from just earlier cases?

Average model performance is worse for models restricted to learning from earlier cases than those not when including just the structured data or both the structured and text data both in validation and on the holdout data. Model performance drops off substantially when predicting on holdout data for the model restricted to learning from earlier cases and containing both structured and text data. The standard deviation of the average precision score during validation is considerably lower for the model restricted to learning from earlier cases when including just the structured data.

Does adding text improve the results?

Adding text improves the model performance for both models trained using both types of validation during validation and on holdout data with the exception of models trained using structured and text data from earlier cases on holdout data. The standard deviation of the average precision score is slightly higher for the model restricted to learning from earlier cases using both structured and text data. For the model restricted to learning from earlier cases, the high standard deviation during validation and the lower performance on the holdout data would suggest a preference for using just the structured data in this case.

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?

Description

Comparing Cross-validation Table 5: Predicting escalation to CPP or CLA after contact: comparison of performance metrics for models learning from all cases or restricted to learning only from earlier cases

Outcome	Case escalates to a Child Protection Plan (CPP) or the child being Looked After (CLA) within 6 to 12 months
Population	Children or young people at the point of contact
Sample size	21120 cases (model not restricted to only learning from earlier cases) 24259 cases (model restricted to only learning from earlier cases)
Years	April 2014 - July 2019
Best algorithm	Gradient boosting for all models except the model using only structured data and restricted to learning from earlier cases

Note: the sample sizes differ by the type of cross validation because observations are dropped according to different criteria to handle the different restrictions

Overall results

Comparing Cross-validation Table 5: Predicting escalation to CPP or CLA after contact: comparison of performance metrics for models learning from all cases or restricted to learning only from earlier cases

Metrics	Data included	Learning from all cases	Learning only from earlier cases	Validation technique with best model performance	Greater than 'success' threshold (0.65) - learning from all cases	Greater than 'success' threshold (0.65) - Learning only from earlier cases
Proportion not at risk of outcome	Structured and text data	0.95	0.95			
Proportion not at risk of outcome	Structured data only	0.95	0.95			
Accuracy	Structured and text data	0.95	0.95	Same performance	Exceeds success threshold	Exceeds success threshold
Accuracy	Structured data only	0.95	0.95	Same performance	Exceeds success threshold	Exceeds success threshold

Average precision	Structured and text data	0.18	0.16	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
Average precision	Structured data only	0.18	0.07	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Structured and text data	0.33	0.11	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Structured data only	0.18	0.07	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
Std average precision (training)	Structured and text data	0.02	0.04	Learning from all cases		
Std average precision (training)	Structured data only	0.03	0.01	Learning only from earlier cases		
AUC	Structured and text data	0.77	0.76	Learning from all cases	Exceeds success threshold	Exceeds success threshold
AUC	Structured data only	0.78	0.59	Learning from all cases	Exceeds success threshold	Does not exceeds success threshold

Comparing data included Table 5: Predicting escalation to CPP or CLA after contact: comparison of performance metrics for models including just structured data and structured and text data

Metrics	Cross Validation Method	Structured data only	Structured and text data	Data included with best model performance	Greater than 'success' threshold (0.65) - structured data only	Greater than 'success' threshold (0.65) - structured and text data
Proportion not at risk of outcome	Learning from all cases	0.95	0.95			
Proportion not at risk of outcome	Learning only from earlier cases	0.95	0.95			
Accuracy	Learning from all cases	0.95	0.95	Same performance	Exceeds success threshold	Exceeds success threshold
Accuracy	Learning only from earlier cases	0.95	0.95	Same performance	Exceeds success threshold	Exceeds success threshold

Average precision	Learning from all cases	0.18	0.18	Same performance	Does not exceeds success threshold	Does not exceeds success threshold
Average precision	Learning only from earlier cases	0.07	0.16	Structured and text data	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Learning from all cases	0.18	0.33	Structured and text data	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Learning only from earlier cases	0.07	0.11	Structured and text data	Does not exceeds success threshold	Does not exceeds success threshold
Std average precision (training)	Learning from all cases	0.03	0.02	Structured and text data		
Std average precision (training)	Learning only from earlier cases	0.01	0.04	Structured data only		
AUC	Learning from all cases	0.78	0.77	Structured data only	Exceeds success threshold	Exceeds success threshold
AUC	Learning only from earlier cases	0.59	0.76	Structured and text data	Does not exceeds success threshold	Exceeds success threshold

None of the four models' average precision scores exceed the threshold for success as defined in the research protocol during validation or on holdout data. The AUC of three of the four models exceed the threshold.

Do the models perform as well when restricting them to learn from just earlier cases?

The average precision score calculated during validation is worse for the model taking into account time but the model performs similarly on holdout data when including structured and text data. The standard deviation is lower for the model learning from earlier cases when just including structured data.

Does adding text improve the results?

During validation, the average precision scores are better when including both structured and text data; however the standard deviation is higher also for the model including both structured and text data and learning only from earlier cases. When testing on holdout data, adding text data improves the average precision score of the model restricted to learning from earlier cases (the performance is the same for models not restricted to learning from earlier cases).

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

Description

Outcome and population Table 6: outcome, population, sample size, years of data available and best algorithm to predict referral after finishing early help

Outcome	Referral to children's social care within 12 months
Population	Children or young people who have successfully finished early help
Sample size	714 cases (model not restricted to only learning from earlier cases) 772 cases (model restricted to only learning from earlier cases)
Years	November 2014 - July 2019
Best algorithm	Logistic regression for the model using structured data only and learning from all cases, and the model using both structured and text data and learning from only earlier cases; gradient boosting for the model using both structured and text data and learning from all cases, and the model using structured data only and restricted to learning from only earlier cases

Note: the sample sizes differ by the type of cross validation because observations are dropped according to different criteria to handle the different restrictions

Overall results

Comparing Cross-validation Table 6: Predicting referral after finishing early help: comparison of performance metrics for models learning from all cases or restricted to learning only from earlier cases

Metrics	Data included	Learning from all cases	Learning only from earlier cases	Validation technique with best model performance	Greater than 'success' threshold (0.65) - learning from all cases	Greater than 'success' threshold (0.65) - Learning only from earlier cases
Proportion not at risk of outcome	Structured and text data	0.84	0.85			
Proportion not at risk of outcome	Structured data only	0.84	0.85			
Accuracy	Structured and text data	0.8	0.83	Learning only from earlier cases	Exceeds success threshold	Exceeds success threshold

Accuracy	Structured data only	0.8	0.84	Learning only from earlier cases	Exceeds success threshold	Exceeds success threshold
Average precision	Structured and text data	0.38	0.47	Learning only from earlier cases	Does not exceeds success threshold	Does not exceeds success threshold
Average precision	Structured data only	0.33	0.54	Learning only from earlier cases	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Structured and text data	0.59	0.53	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Structured data only	0.49	0.53	Learning only from earlier cases	Does not exceeds success threshold	Does not exceeds success threshold
Std average precision (training)	Structured and text data	0.03	0.16	Learning from all cases		
Std average precision (training)	Structured data only	0.02	0.15	Learning from all cases		
AUC	Structured and text data	0.77	0.77	Same performance	Exceeds success threshold	Exceeds success threshold
AUC	Structured data only	0.70	0.85	Learning only from earlier cases	Exceeds success threshold	Exceeds success threshold

Comparing data included Table 6: Predicting referral after finishing early help: comparison of performance metrics for models including just structured data and structured and text data

Metrics	Cross Validation Method	Structured data only	Structured and text data	Data included with best model performance	Greater than 'success' threshold (0.65) - structured data only	Greater than 'success' threshold (0.65) - structured and text data
Proportion not at risk of outcome	Learning from all cases	0.84	0.84			
Proportion not at risk of outcome	Learning only from earlier cases	0.85	0.85			
Accuracy	Learning from all cases	0.80	0.80	Same performance	Exceeds success threshold	Exceeds success threshold

Accuracy	Learning only from earlier cases	0.84	0.83	Structured data only	Exceeds success threshold	Exceeds success threshold
Average precision	Learning from all cases	0.33	0.38	Structured and text data	Does not exceeds success threshold	Does not exceeds success threshold
Average precision	Learning only from earlier cases	0.54	0.47	Structured data only	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Learning from all cases	0.49	0.59	Structured and text data	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Learning only from earlier cases	0.53	0.53	Same performance	Does not exceeds success threshold	Does not exceeds success threshold
Std average precision (training)	Learning from all cases	0.02	0.03	Structured data only		
Std average precision (training)	Learning only from earlier cases	0.15	0.16	Structured data only		
AUC	Learning from all cases	0.70	0.77	Structured and text data	Exceeds success threshold	Exceeds success threshold
AUC	Learning only from earlier cases	0.85	0.77	Structured data only	Exceeds success threshold	Exceeds success threshold

None of the average precision scores for the four models exceeds the threshold for success as defined in the research protocol during validation or on holdout data. The AUCs of all four models exceed the threshold.

Do the models perform as well when restricting them to learn from just earlier cases?

Unusually, the average precision of the models restricted to learning from earlier cases exceeds the average precision of the models not restricted on holdout data and for when including just the structured data during validation. However, the standard deviation of the average precision score during validation is quite high, suggesting that the patterns learnt by the model restricted to learning from just earlier cases may struggle to generalise.

Does adding text improve the results?

Adding text data improved the model performance for the model not restricted to learning from earlier cases in both validation and when testing on holdout data but worsens it for the model restricted to learning from earlier cases when predicting on holdout data. The standard deviation is approximately the same when including just structured data and both structured and text data.

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?

Description

Outcome and population Table 7: outcome, population, sample size, years of data available and best algorithm to predict escalation to CPP after assessment

Outcome	Case escalates to a Child Protection Plan (CPP) 1-12 months after the assessment authorisation date
Population	Children or young people at the point of assessment
Sample size	2365 cases
Years	March 2012 - March 2019
Best algorithm	Gradient boosting for models not restricted to learning from earlier cases; decision tree for the model using just structured data and restricted to learning from earlier cases; logistic regression for the model using both structured and text data and restricted to learning from earlier cases

Overall results

Comparing Cross-validation Table 7: Predicting escalation to CPP after assessment: comparison of performance metrics for models learning from all cases or restricted to learning only from earlier cases

Metrics	Data included	Learning from all cases	Learning only from earlier cases	Validation technique with best model performance	Greater than 'success' threshold (0.65) - learning from all cases	Greater than 'success' threshold (0.65) - Learning only from earlier cases
Proportion not at risk of outcome	Structured and text data	0.93	0.93			
Proportion not at risk of outcome	Structured data only	0.93	0.93			
Accuracy	Structured and text data	0.92	0.88	Learning from all cases	Exceeds success threshold	Exceeds success threshold
Accuracy	Structured data only	0.9	0.89	Learning from all cases	Exceeds success threshold	Exceeds success threshold

Average precision	Structured and text data	0.10	0.09	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
Average precision	Structured data only	0.11	0.09	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Structured and text data	0.68	0.10	Learning from all cases	Exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Structured data only	0.24	0.11	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
Std average precision (training)	Structured and text data	0.05	0.02	Learning only from earlier cases		
Std average precision (training)	Structured data only	0.15	0.00	Learning only from earlier cases		
AUC	Structured and text data	0.60	0.45	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
AUC	Structured data only	0.67	0.52	Learning from all cases	Exceeds success threshold	Does not exceeds success threshold

Comparing data included Table 7: Predicting escalation to CPP after assessment: comparison of performance metrics for models including just structured data and structured and text data

Metrics	Cross Validation Method	Structured data only	Structured and text data	Data included with best model performance	Greater than 'success' threshold (0.65) - structured data only	Greater than 'success' threshold (0.65) - structured and text data
Proportion not at risk of outcome	Learning from all cases	0.93	0.93			
Proportion not at risk of outcome	Learning only from earlier cases	0.93	0.93			
Accuracy	Learning from all cases	0.90	0.92	Structured and text data	Exceeds success threshold	Exceeds success threshold
Accuracy	Learning only from earlier cases	0.89	0.88	Structured data only	Exceeds success threshold	Exceeds success threshold

Average precision	Learning from all cases	0.11	0.10	Structured data only	Does not exceeds success threshold	Does not exceeds success threshold
Average precision	Learning only from earlier cases	0.09	0.09	Same performance	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Learning from all cases	0.24	0.68	Structured and text data	Does not exceeds success threshold	Exceeds success threshold
Mean average precision (training)	Learning only from earlier cases	0.11	0.10	Structured data only	Does not exceeds success threshold	Does not exceeds success threshold
Std average precision (training)	Learning from all cases	0.15	0.05	Structured and text data		
Std average precision (training)	Learning only from earlier cases	0.00	0.02	Structured data only		
AUC	Learning from all cases	0.67	0.6	Structured data only	Exceeds success threshold	Does not exceeds success threshold
AUC	Learning only from earlier cases	0.52	0.45	Structured data only	Does not exceeds success threshold	Does not exceeds success threshold

None of the four models had average precision scores which exceeded the threshold when calculated when predicting on holdout data. One of the four models' mean average precision exceeds the threshold during validation. One of the AUCs for the four models exceeded the success threshold specified in the research protocol.

Do the models perform as well when restricting them to learn from just earlier cases?

The average precision scores calculated during validation are substantially worse for models restricted to learning from earlier cases - in particular the model using structured and text data worsens substantially when just learning from earlier cases - but the models perform similarly on holdout data. but the standard deviation is lower for models restricted to learning from earlier cases.

Does adding text improve the results?

Adding text substantially improves the average precision scores during validation for the model not restricted to learning from just earlier cases and is about the same for the model restricted to learning from earlier cases. The models including structured and text data perform about the same on holdout data.

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

Description

Outcome and population Table 8: outcome, population, sample size, years of data available and best algorithm to predict escalation to CLA after assessment

Outcome	Case escalates to the child being Looked After 1-12 months after the assessment authorisation date
Population	Children or young people at the point of assessment
Sample size	2365 cases
Years	March 2012 - March 2019
Best algorithm	Decision tree for models using just structured data; gradient boosting for models using both structured and text data

Overall results

Comparing Cross-validation Table 8: Predicting escalation to CLA after assessment: comparison of performance metrics for models learning from all cases or restricted to learning only from earlier cases

Metrics	Data included	Learning from all cases	Learning only from earlier cases	Validation technique with best model performance	Greater than 'success' threshold (0.65) - learning from all cases	Greater than 'success' threshold (0.65) - Learning only from earlier cases
Proportion not at risk of outcome	Structured and text data	0.96	0.96			
Proportion not at risk of outcome	Structured data only	0.96	0.96			
Accuracy	Structured and text data	0.96	0.95	Learning from all cases	Exceeds success threshold	Exceeds success threshold
Accuracy	Structured data only	0.96	0.96	Same performance	Exceeds success threshold	Exceeds success threshold
Average precision	Structured and text data	0.07	0.06	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold

Average precision	Structured data only	0.10	0.06	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Structured and text data	0.21	0.27	Learning only from earlier cases	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Structured data only	0.24	0.15	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
Std average precision (training)	Structured and text data	0.06	0.10	Learning from all cases		
Std average precision (training)	Structured data only	0.02	0.08	Learning from all cases		
AUC	Structured and text data	0.56	0.54	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold
AUC	Structured data only	0.63	0.56	Learning from all cases	Does not exceeds success threshold	Does not exceeds success threshold

Comparing data included Table 8: Predicting escalation to CLA after assessment: comparison of performance metrics for models including just structured data and structured and text data

Metrics	Cross Validation Method	Structured data only	Structured and text data	Data included with best model performance	Greater than 'success' threshold (0.65) - structured data only	Greater than 'success' threshold (0.65) - structured and text data
Proportion not at risk of outcome	Learning from all cases	0.96	0.96			
Proportion not at risk of outcome	Learning only from earlier cases	0.96	0.96			
Accuracy	Learning from all cases	0.96	0.96	Same performance	Exceeds success threshold	Exceeds success threshold
Accuracy	Learning only from earlier cases	0.96	0.95	Structured data only	Exceeds success threshold	Exceeds success threshold
Average precision	Learning from all cases	0.10	0.07	Structured data only	Does not exceeds success threshold	Does not exceeds success threshold

Average precision	Learning only from earlier cases	0.06	0.06	Same performance	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Learning from all cases	0.24	0.21	Structured data only	Does not exceeds success threshold	Does not exceeds success threshold
Mean average precision (training)	Learning only from earlier cases	0.15	0.27	Structured and text data	Does not exceeds success threshold	Does not exceeds success threshold
Std average precision (training)	Learning from all cases	0.02	0.06	Structured data only		
Std average precision (training)	Learning only from earlier cases	0.08	0.1	Structured data only		
AUC	Learning from all cases	0.63	0.56	Structured data only	Does not exceeds success threshold	Does not exceeds success threshold
AUC	Learning only from earlier cases	0.56	0.54	Structured data only	Does not exceeds success threshold	Does not exceeds success threshold

None of the average precision scores calculated during validation and on holdout data for the four models exceed the 'success' threshold. None of the AUCs of the models exceed the threshold either.

Do the models perform as well when restricting them to learn from just earlier cases?

During validation, the model which can learn from all the cases has a higher mean average precision score than the model restricted to learning from earlier cases for the model including just structured data but it is lower for the model including the structured and text data. The models which can learn from all cases irrespective of time perform better on the holdout data when including both the structured data only, and the structured and text data.

Does adding text improve the results?

Adding text improves model performance for the model restricted to only learning from previous cases during validation. Adding text data worsens the average precision score on holdout data for the model not restricted to learning from earlier cases and does not improve performance for the model restricted to learning from earlier cases. Furthermore, the standard deviation of the average precision score during validation of the model including structured and text data is three times that of the model including just structured data for the model not restricted to learning from earlier cases.

Discussion

Please note that we did not prespecify how we would aggregate the results from the 32 models. We hence describe the choice of analysis in some depth below.

Do the models 'work'?

Overall, none of the 32 models exceeded the threshold for success as measured by average precision on holdout data. Given that none of the models exceed the threshold for success

we prespecified, we conclude that the machine learning applied to a children's social work context does not 'work' in the way we have applied it here: using some variation of Annex A datasets and reports / assessments, and testing a range of algorithms (decision tree, logistic regression and gradient boosting algorithms) to predict outcomes mostly relating to escalation of cases. This setup describes the most likely data available to a local authority without additional complex data sharing arrangements with multi-agencies and the most likely interest in outcomes (given that the outcomes were initially selected by the local authority partners).

As well as not exceeding the success threshold, the models have a high variance in their performance on different datasets. This can be seen in the high standard deviation of the average precision score (especially relative to its mean) in validation (see Table 1). The models are neither good nor consistent (they have 'high bias' and 'high variance').

Table 1: Mean and standard deviation of average precision, averaged over all models

	Mean average precision	Standard deviation of average precision
Local authorities 1-4	0.24	0.18
Local authorities 2-4	0.21	0.16

Note: we report the aggregate statistics with and without LA1 because information leakage identified after the analysis was complete is likely to have artificially inflated the scores.

The models tend to overfit to the patterns identified in the training data and these patterns do not generalise to unseen data. This can be seen in the substantial drop in model performance from validation to predicting on unseen data: the average difference between the mean average precision during validation and the average precision on holdout data is 0.12. The learning curves in Appendix 1 (under 'Learning Curves') drawn for 24 out of the 32 models show a large gap between the lines representing the scores calculated on the training and holdout data in 15 out of the 24 graphs, which is indicative of overfitting (learning patterns in the training data which don't generalise). As shown in Table 2 below, the overfitting tends to be worse for gradient boosting (which was the best algorithm for 25 out of the 32 models), when including both structured and text data, and when the model is not restricted to learning from earlier cases.

Table 2: Difference between the mean average precision in validation and the average precision calculated on holdout data, averaged by the algorithm, the data included and the cross validation methodology

Comparison	Difference averaged by:	Difference between average precision in validation and on holdout data
Algorithm	Decision Tree	0.08
Algorithm	Gradient Boosting	0.14
Algorithm	Logistic Regression	0.06
Data included	Structured and text data	0.16
Data included	Structured data only	0.08
Cross Validation Method	Learning from all cases	0.16
Cross Validation Method	Learning only from earlier cases	0.08

The choice of model performance metric can mask a considerable error rate

Whilst we prespecified that our criterion for success would be based on the average precision score, we also report the area under the curve (AUC) because it is a common choice of metric. As can be seen in Table 3, fourteen models out of 24 exceeded the threshold measured by AUC. The same models perform much better when measured by the AUC than the average precision. This is because the outcomes we've focused on predicting only have a small percentage of the children / young people at risk of that outcome (there is considerable 'class imbalance') and the AUC rewards the model for correctly identifying the children / young people not at risk. The accuracy scores are high for similar reasons. Although the models seem to perform reasonably by the AUC and accuracy metric, this masks considerable false positive and false negative rates. Figure 6 visually illustrates how different metrics measuring the performance of the same model can give a very different picture of how well the model predicts outcomes.

Table 3: The number of models out of 24 that exceed the success threshold of 0.65 by metric

Metrics	Does not exceeds success threshold	Exceeds success threshold
AUC	10	14
Accuracy	0	24
Average precision	24	0
F score (beta = 0.1)	22	2
Maximum lower bound (training)	23	1
Mean average precision (training)	23	1
Precision	22	2
Recall	23	1

Note: LA2-LA4 only - LA1 is excluded because of information leakage leading to artificially high scores

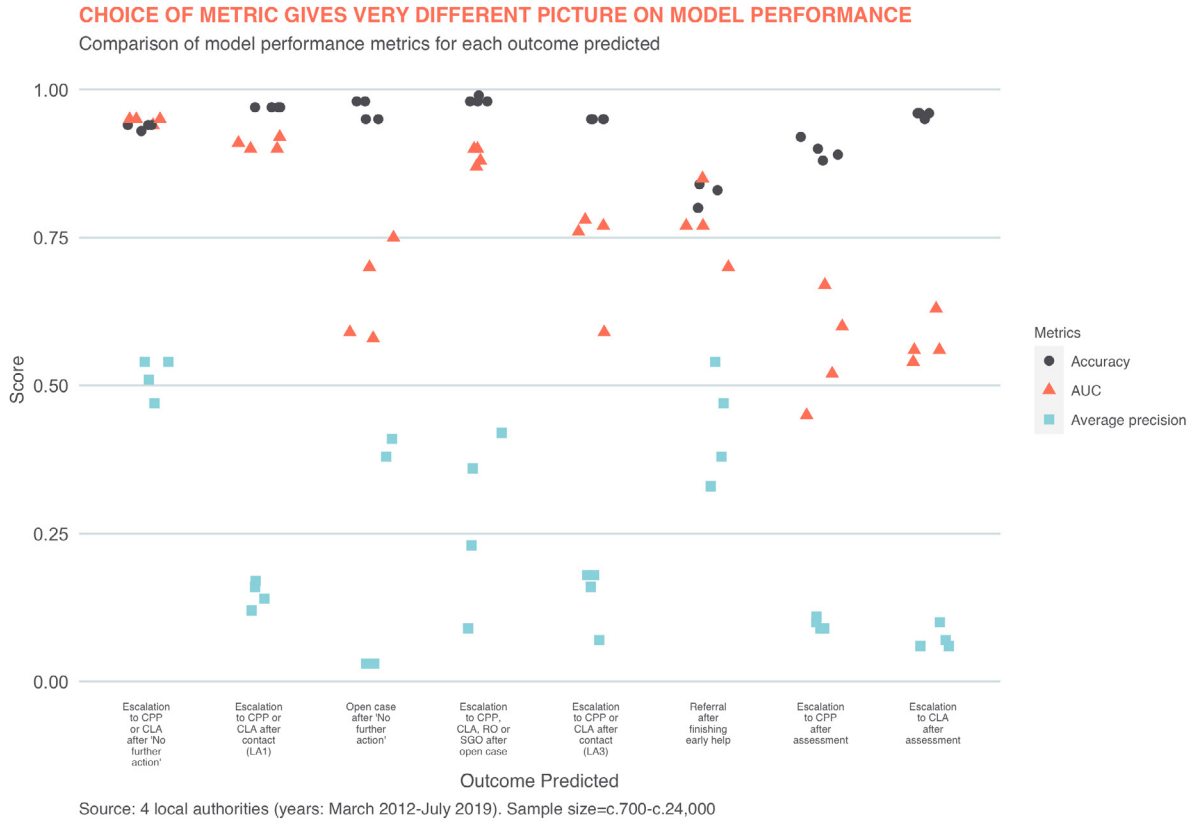


Figure 6: Choice of metric gives very different picture on model performance

We also report the f-score (with beta = 0.1), precision and recall for each model and we report the mean average precision in validation for the model which performed best under our decision rule (the highest lower bound). The f-score is the weighted harmonic mean of model's precision and recall - a beta of 0.1 means that it is heavily weighted towards precision. We chose to report the f-score based on initial results indicating a high proportion of false positives (low precision). These scores give an indication of how the model is making the tradeoff between high precision (few false alarms) and high recall (few children at risk missed). Please see the set of 'Model Performance Metrics' in Appendix 1 if you would like to see these additional details.

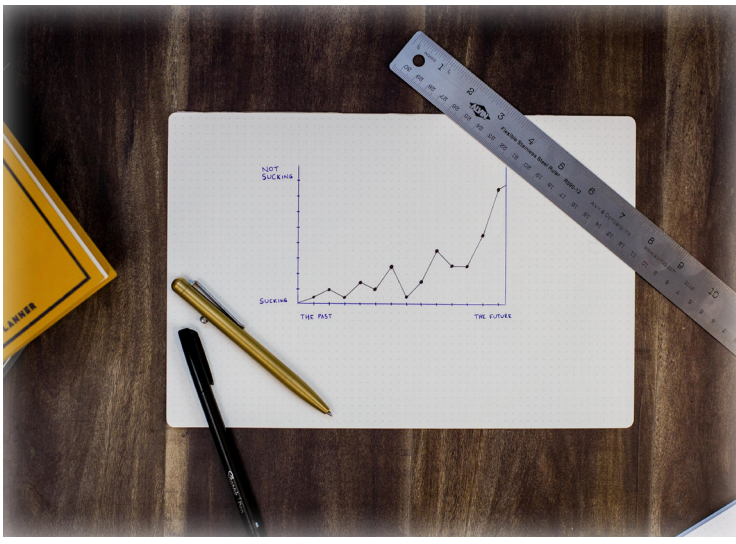
We also report confidence intervals for the metrics on the best performing model for LA2, LA3 and LA4 (analysis at LA1 was complete when we conducted this additional analysis). The confidence intervals give an idea of the range of the metric. The confidence intervals were estimated using bootstrapping with 500 fits of the data on cases sampled randomly with replacement. 'Performance Metrics with confidence intervals for best performing algorithm' in Appendix 1 report the confidence intervals.

Based on the considerable difference between the results depending on how the models' performance is measured, we suggest that class imbalance (what proportion of the population is at risk of the outcome) - common in the context of children's social care - should be taken into account when considering the choice of metric by which to measure performance in the context of the use of machine learning in children's social care.

How well do the models perform on 1000 cases?

The metrics discussed above give an indication of the types of errors the models are making. To make this more concrete, we convert these metrics into how many cases the model would correctly identify as at risk in 1000 cases. (We had originally stated that we would look at 100 cases but because there were very small numbers of positive cases and we felt that the performance of the models would be better represented by reporting the number of children at risk missed and false alarms in 1000 cases.) When looking at the breakdown of types of errors averaged over all the models (see Table 4), 17 / 29 (58%) of cases identified as at risk are false alarms. If the model were to be used in practice, this means that when the model flags a child as at risk, it is a false alarm in 58% of instances. The average model also fails to identify 46 / 58 (79%) children who are at risk. These metrics are of cause for concern.

This balance between the errors is not inevitable and changed considerably as we tweaked the models to reduce one of the types of errors. One way to improve the model is to change the threshold for the probability of the outcome to be classified as 'at risk'. Setting the threshold higher ensures that only cases with high probability are classified as at risk, ensuring that the model is quite confident about the cases it identifies as at risk but with the tradeoff that it is more likely to miss children / young people at risk. This can be seen in the precision-recall curves in Appendix 1 ('Precision Recall Curves'). In the majority of the models, identifying more children at risk (increasing recall, moving to the right on the x-axis) comes with the tradeoff of many more false alarms (a sharp decrease in precision on the y-axis).



If the cases were ranked by the probability predicted by the model from highest risk to lowest risk, a social work manager reviewing only the top 10% would be reviewing the 40% of cases with the highest probability of risk. This is disappointing given the level of class imbalance. To illustrate this, consider a model predicting on an outcome which occurs for 5% of the population: a perfect model would identify 100% of the cases at risk of that outcome in the top 5% of cases when the cases were ordered from those most at risk to those least at risk. A less than perfect model will misclassify more and more cases as you go down the ranked list. Given that we are looking over a larger percentage of the data than the percentage of cases at risk of that outcome, a good model should identify the majority of cases within the top 10%, and certainty within the top 25% (the percentage of cases we originally intended to inspect).

Table 4: Percentage of risk cases in top 10%; percentage of safe cases in bottom 10%; average number of true positives, true negatives, false positives and false negatives in 1000 cases

Metric	N
% of risky cases in top 10%	40
% of safe cases in bottom 10%	10
Number of children at risk missed ('false negatives') in 1000 cases	46
Number of false alarms ('false positives') in 1000 cases	17
Number of children correctly identified as not at risk ('true negatives') in 1000 cases	925
Number of children correctly identified as at risk ('true positives') in 1000 cases	12

Please see 'Intuitive Metrics' in Appendix 1 for these numbers by model.

Does restricting the models to learning from just earlier cases impact on model performance?

Restricting the models to learning from just earlier cases is a more accurate test of how well the models would perform if deployed in practice where the model would have access to the earlier cases and use currently available data on a case to predict the future of the case. In 13 of the 16 head to head comparisons of models restricted to learning from earlier cases and those not, restricting the model to learn from earlier cases worsened the performance of the model whilst one pair of models performs the same (see Table 5).

Table 5: Best performing model (measured by average precision on holdout data) by cross validation method

Cross validation technique	Number of times the technique is the best performing
Learning from all cases	13
Learning only from earlier cases	2
Same performance	1

The average precision scores for the four models predicting each of the eight outcomes are displayed in Figure 7. Except for the outcome 'referral after early help', the models not restricted to learning from earlier cases (identified by the darker coloured bars) have higher scores. Figure 8 illustrates the mean average precision score by whether the model is restricted to learning from just earlier cases to show the overall trend instead of pairwise comparisons for each outcome predicted.

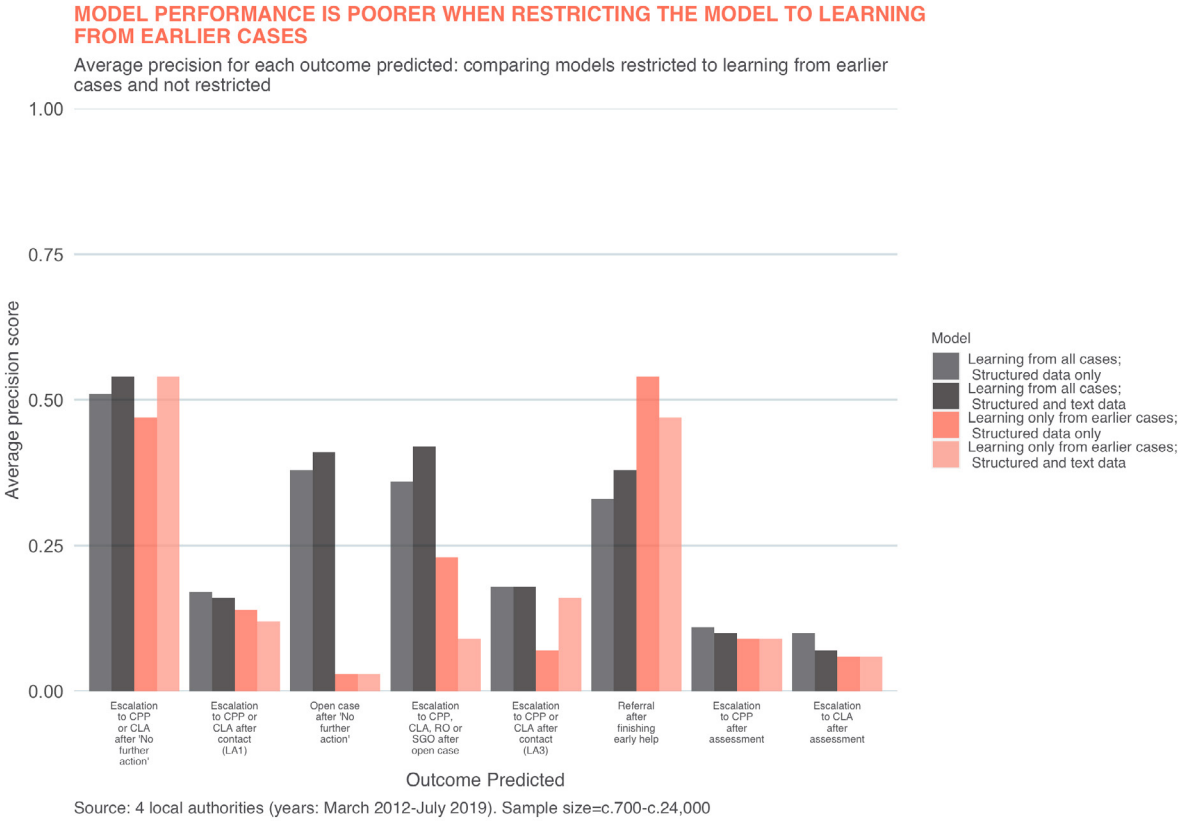


Figure 7: Model performance is poorer when restricting the model to learning from earlier cases

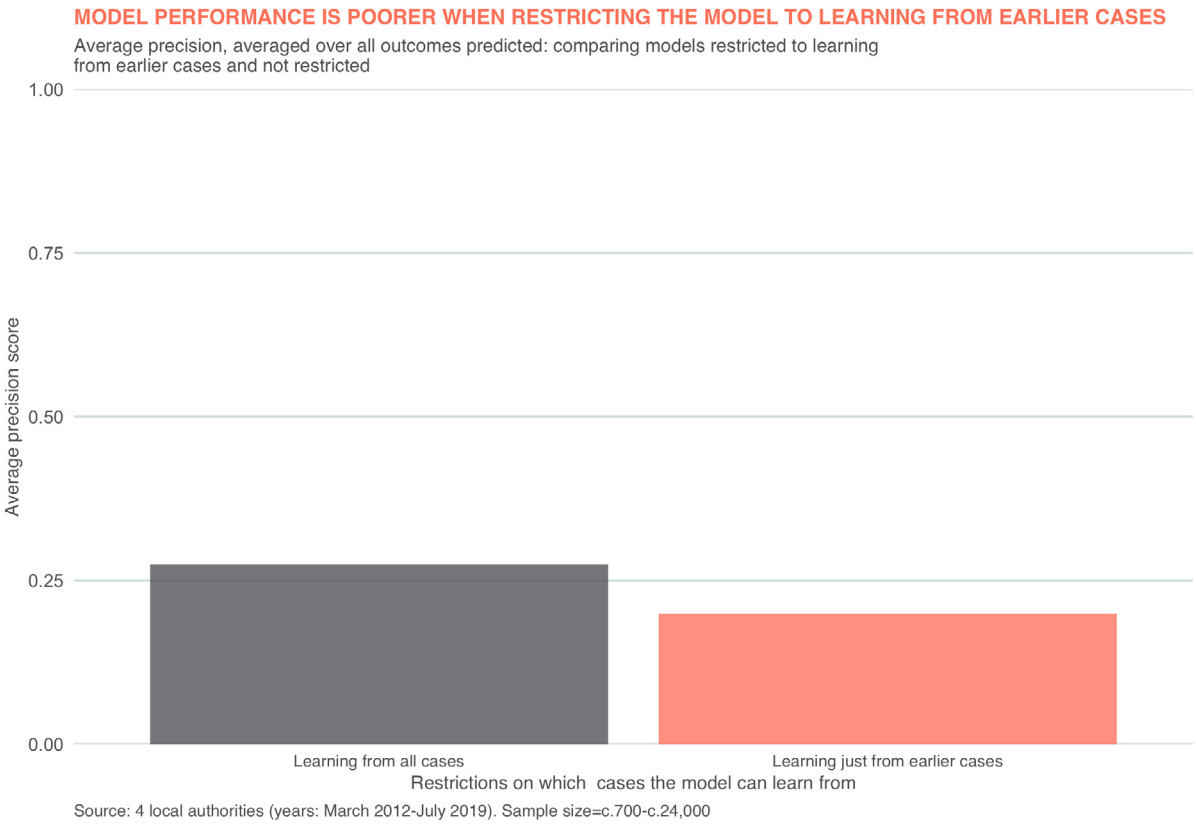


Figure 8: Model performance is poorer when restricting the model to learning from earlier cases (aggregated)

We conducted a Mann Whitney U test to compare the scores of models restricted to learning from earlier cases and those not (metrics: AUC, accuracy, average precision, F score (beta = 0.1), mean average precision (training), precision, recall). The null hypothesis of an equal distribution of scores was rejected at the 5% significance level (Mann–Whitney U = 10145.5, n1 = n2 = 128, P = 0.00 two-tailed). This offers moderate evidence that the models are learning additional information from future cases. We state only moderate evidence because it is inherent to the design of cross-validators restricted by the time structure of the data that the folds increase in sample size as the number of validations increases, and because the models restricted to only learning from earlier cases were validated with fewer cross-validation splits. All other things being equal and up to a point, we would expect models trained on larger sample sizes and models validated with more cross-validation splits to perform better on unseen data.

To explore possible explanations for this difference in performance between models restricted to learning from earlier cases and those not, we conduct t-tests to compare the class balance between the train and test folds within cross validation (i.e. whether the proportion of cases at risk of the outcome was the same), and F-tests for joint orthogonality to test whether the distribution of the input data is the same in the train and test fold.

Table 6: Average p-values from T-tests to test class balance between train and test fold and F-test to test whether the distribution of the input data is the same in the train and test fold averaged by type of cross validation

	Average p-value from F-tests	Average p-value from T-tests
Learning from all cases	0.50	0.74
Learning only from earlier cases	0.00	0.36

Please see the 'Tests for difference between cross validation splits' set of tables in the Appendix for a breakdown by model.

Whilst the class balance is not significantly differently between the train and test splits (as shown by the high average p-values for the T-tests above); the difference between the train and test fold input data within cross-validation is highly significant for the models restricted to learning just from earlier cases but the input data in the train and test folds for the models not restricted to learning from earlier cases is not significantly different. This suggests that the distribution of the data is changing over time and it is not suitable to go back further in time in the quest for a larger training dataset. This is an important finding because often the solution proposed to low model performance is 'more data'. However, it is not clear how to facilitate that in light of these findings. Due to the responsibility of children's social care sitting at a local government level, there are considerable challenges in combining data from multiple local authorities: both on a practical level (data governance and combining data from different case management systems which record the information about children and young people in different ways) and on an analytical level (data from local authorities with different practice models, interpretation of thresholds and populations is likely to add considerable noise as well as signal). For these reasons, it is difficult - and we suspect not helpful - to combine data from multiple local authorities. With the input data looking significantly different over time, training the model using cases from earlier in time does not look like a way to solve low model performance either.

Is the model generalising the experience of a small number of families?

With relatively small datasets for some outcomes, high class imbalance for all outcomes and need to group siblings in the same cross validation folds, we were concerned that the model may be learning from a small number of families and generalising their experiences. To investigate this, we counted the number of times a sibling group appeared in a cross validation fold (this may have been repeated observations of the same individuals or different siblings from the same sibling group). On average, the largest sibling

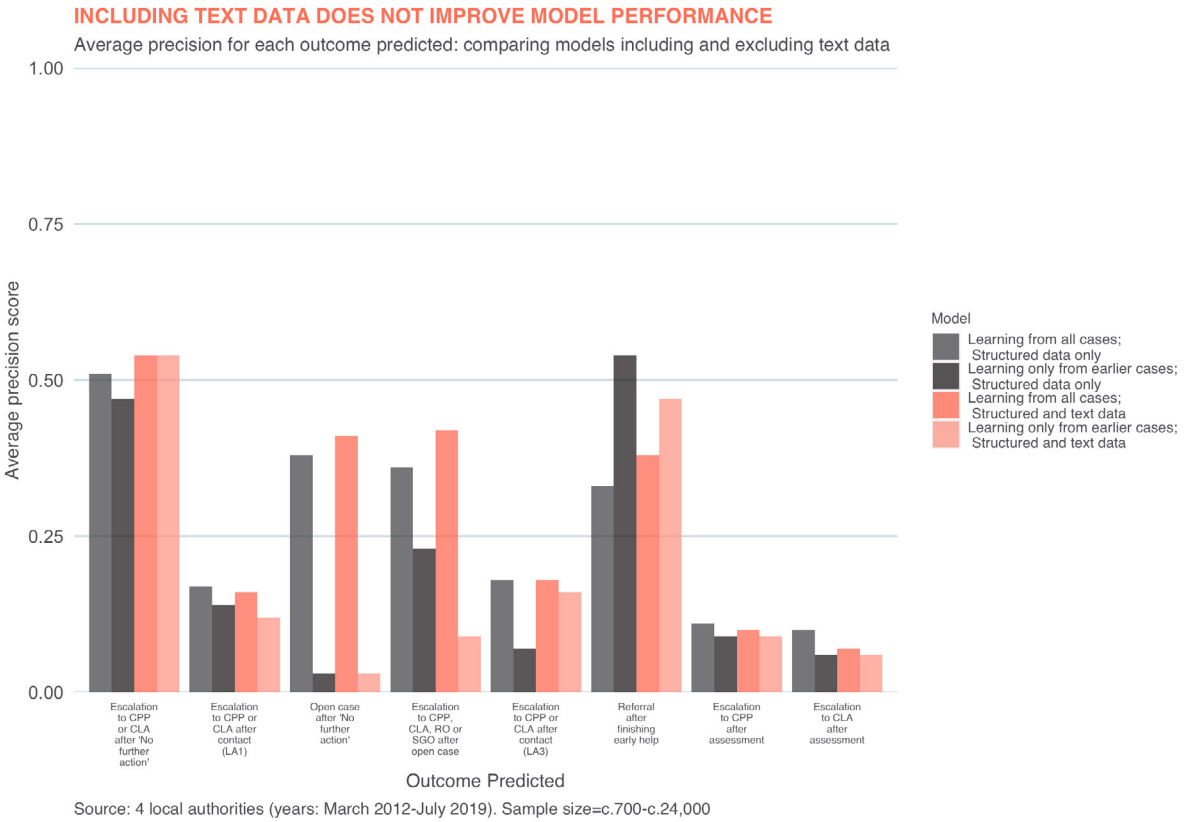
group represents 1.05% of the population in the fold. This is a reasonably small percentage but is substantial given that siblings tend to have the same outcomes and the percentage of children and young people at risk in the data is between 2% and 6% for seven of the eight outcomes being predicted. This may explain why we see a considerable decrease in average precision from validation to predicting on holdout data - the patterns learnt from the small number of sibling groups during training are not generalising to the individuals observed in the holdout dataset. Please see the 'Siblings within cross validation folds' set of tables in Appendix 1 for a breakdown by model.

Does adding text data improve the model performance?

The 16 head to head comparisons of models including just structured data or including structured and text data were evenly split between the type of data included with 6 models with just structured data performing better, 6 models including structured and text data performing better and 4 with the same performance (see Table 7). Figure 9 shows the average precision score for each of the four models predicting each of the eight outcomes, and Figure 10 shows the mean average precision scores by whether they include text data in addition to the structured data.

Table 7: Best performing model (measured by average precision on holdout data) by data included

Data included in model	Number of times the model with this data included is the best performing
Structured and text data	6
Structured data only	6
Same performance	4



Source: 4 local authorities (years: March 2012-July 2019). Sample size=c.700-c.24,000

Figure 9: Including text data does not improve model performance

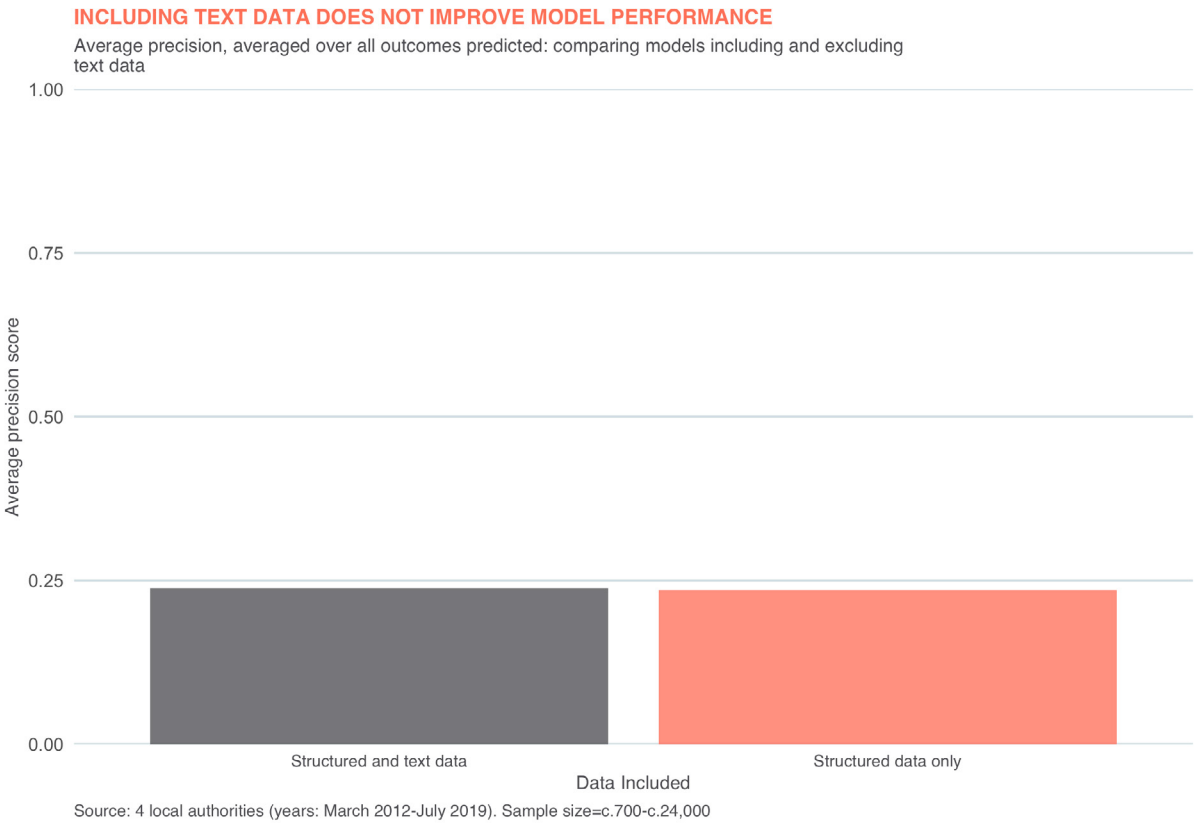


Figure 10: Including text data does not improve model performance (aggregated)

A Mann Whitney U test comparing the scores of models on the same metrics as the test above including just structured data or structured and text data failed to reject the null hypothesis of an equal distribution of scores at the 5% significance level (Mann–Whitney U = 7596, n1 = n2 = 128, P = 0.31 two-tailed).

The models including both structured and text data also tend to overfit more as can be seen in the average drop in average precision from the mean average precision during cross-validation to the average precision as calculated when the model predicts on the holdout data: the drop is 0.16 for models including structured and text data whilst it is 0.08 for models including only structured data.

We were interested to understand whether text data could be used to support the interpretation of the outputs of the model even if it does not add much predictive performance. However, the word clouds (included in Appendix 1 under 'Word Clouds') are not particularly informative: the highly frequent words are as expected and do not help distinguish between cases. It appears that topic modelling may not add much in the way of interpretation when the topic space is so constrained. It may be that other natural language processing (NLP) methodologies such as text summarisation improve the model performance and be more interpretable but are also likely to require larger datasets because they rely on deep learning methodologies.

Given that these results suggest that including text data does not offer additional value, we suggest that the additional burden of extraction from the system and the data governance arrangements is not worthwhile.

What can be done to improve the models?

In summary, the performance of all the models is below the threshold for success we prespecified in the research protocol. Restricting the models to learning only from earlier cases tends to decrease the performance of the model. This is

an important finding as this is a more accurate representation of how the model would perform if it were to be used in practice and demonstrates the importance of transparency about whether the model is restricted to learning from just earlier cases when reporting model performance.

This finding also suggests that trying to improve the model using data on cases from further back in time is unlikely to be helpful - and hence creating larger datasets which cover a long time period is not a solution to low performance. Furthermore, trying to improve the model by adding text data does not look fruitful, especially when considering the additional data governance hurdles. As discussed above, trying to improve the model performance by combining data from different local authorities is both practically difficult and unlikely to be helpful. Local authorities and their data are very different, and these differences are likely to be greater than any medium term change within a single authority. Hence, if using historic data from one local authority does not produce good predictions, using historic data from a *different local authority* is unlikely to produce better predictions. It may be that adding richer data about each individual child / young person / family (rather than adding more observations) could provide a richer picture of the child's case and improve model performance. However, making this data available requires considerable investment in multi-agency data sharing and data infrastructure. Given that the use of machine learning in children's social care context does not show promise, we would not advise making these considerable investments in order to further pilot machine learning techniques in this context (although of course these investments may be worthwhile in their own right or in the pursuit of other purposes). Furthermore, because the models seem to be overfitting to the data - learning too many of the nuances of the cases - adding richer data risks worsening this instead of improving model performance.

Testing different algorithms is sometimes also proposed as a solution to improve model performance. We tested algorithms with varying levels of complexity (decision trees, regularised logistic regression and gradient boosting algorithms). In 25 out of the 32 models, the most complex algorithm - gradient boosting - performed the best in validation; however, it was liable to overfitting with the decrease in average precision score from the mean in validation to the score on holdout data was 0.14. This suggests that testing more complex algorithms, for example, neural networks, is not likely to be fruitful.

Tuning of the hyperparameters to better facilitate the search for the best parameters for each model is also something which data scientists use to try to improve model performance. Given the extent of the overfitting, it is likely that the hyperparameters could be set to explore simpler models and this may reduce the considerable drop in performance from validation to holdout observed. It is difficult to estimate how likely this is to improve the models. We conducted a randomised search with 50 iterations over a rather wide hyperparameter space. We were short on time to explore the hyperparameter space further (we spent approximately 12 weeks preparing and analysing the data from each local authority); however, we expect such resource constraints to also apply to others attempting the use of machine learning in the context of children's social care.

Why is model performance so low and do we expect better model performance as techniques advance?

It is worthwhile trying to answer these questions in light of the prevalence of machine learning techniques in our lives, the complexity of the tasks machine learning can achieve in some other areas and how quickly such techniques advance. Observing machine learning in other areas, it may appear inevitable that machine learning techniques will be used helpfully in children's social care over the coming years. Our results from these pilots do not directly allow us to answer these questions but help us generate hypotheses about which factors specific to the children's social care context in England may put a ceiling on model performance.

How well a model performs depends on a number of factors. The quality and quantity of the data available play a large role. With regard to quality, the representativeness, relevance and accuracy of the data is important.

The data we used was representative of the population on whom the model would predict. The data was relevant to the extent that we believe that previous involvement in children's social care and demographic factors are predictive of the outcomes of interest. We did not have access to education data which may have been relevant. The feature importance plots ('Feature Importance Plots' in Appendix 1) which show how much each feature contributes towards the prediction show that we still do not know a huge amount about what is relevant to predicting these outcomes. This speaks to the findings of the ethics review by The Alan Turing Institute and the Rees Centre at the University of Oxford that we need to better understand the mechanisms at play in a child's social care journey. Depending on how much strengths-based practice is emphasised in the local authority's practice model, the records, reports and assessments varied in the extent to which they included the strengths of the family, which is additional relevant information.

As the data extracts were mostly Annex A datasets prepared for Ofsted, they had already undergone accuracy checks by the intelligence teams. Missing data was relatively rare. Some values were outside of expected range (e.g. years of birth suggesting an age far beyond which the person would be eligible for support from children's social care) and some derived values suggested that the record on the case management system had not been updated (for example, durations of assessment which suggested that some assessments continue for years). One of the local authorities had transitioned from a previous case management system and some of the reports had been replaced with a warning message about the data migration. The local authorities were all rated Good or Outstanding in their most recent Ofsted inspections and so we imagine that these foibles in data quality are fewer on average than in other local authorities' datasets, and in some cases these inaccuracies are relatively straightforward to handle. How well the data recorded in the case management system reflects the 'true' picture of the family's situation on the ground is something which we could not assess but is worth noting in further detail. As is common to all local authorities, case management systems restrict the way information can be recorded - this standardisation can be helpful but can mean that cases are allocated to categories that aren't really appropriate (or that individual social workers use these categories in slightly different ways) and that additional information that may be relevant may not be recorded (for example, there may be multiple needs identified at assessment but only the primary need recorded). Qualitative research also conducted as part of the research feeding into the ethics review found that reports contain information which is contested by families. We don't expect these factors to be any worse for any of the local authority partners we worked with but it is worth keeping in mind this general context of the data in children's social care when thinking about building predictive models.

The quantity of the data is small relative to the complexity of the outcomes being predicted and also the structure of the dataset. The raw number of cases who are at risk of the outcome is small. Additionally, because the outcomes for siblings tend to be similar, the amount of information the model can learn from observing multiple individuals who are related is less than observing the same number of individuals who are unrelated. Furthermore, the kinds of outcomes which local authorities are likely to want to predict are social worker's decisions - for example, about whether to escalate or deescalate / close a case - and given that decisions made by the same social worker are likely to be more similar to each other than decisions made by different social workers, the amount of information the model can extract about generalisable patterns is further reduced. Given these restrictions, the model attempts to generalise from a small number of siblings and cases decided by the same social worker, and given the complexity and nuance of each case, we start to understand why the model struggles to perform better. Furthermore, feedback on whether the model is right in practice is likely to be slow (6 - 12 months down the line), reducing how quickly the model can learn from how well the model performs in practice.

We now turn to whether these factors which we suspect put a ceiling on the model performance are within the local authority's power to change or whether their impact will lessen over time. As we've seen, the input data changes over time to the extent that gathering more observations by going further back in time or simply sitting back and waiting for more cases to be observed would not substantially aid model performance. As discussed above, combining data from different local authorities comes with its own challenges. We cannot change the similarity of the outcomes for related individuals.

However, we can change the outcomes of interest to outcomes not dependent on social worker decisions. An example of such an outcome would be risk of unwanted pregnancy under the age of 16 or exclusion from school. This would decrease the number of observations which are similar to each other by virtue of having the same social worker (but may increase the similarity of children who are attending the same school) and would thus would allow more information to be learnt from each individual case. Outcomes which a higher proportion of the population are at risk of would also likely improve the model because the model would be able to learn more general patterns from the wider variety of cases. However, of course, the outcome has to be meaningful rather than just predictable.

Machine learning techniques will continue to improve but given that we expect the factors which we suspect put a ceiling on model performance to remain, new techniques would be able to improve the models only to a certain extent.

Technical Research Question 3: What is the performance of the models on different subgroups of interest?

Please note that we did not prespecify how we would aggregate the results from the 32 models. We hence describe the choice of analysis in some depth below.

Research question 3 gives insight into whether the model is fair to children and families irrespective of their sensitive characteristics - protected characteristics (those included in the *Equality Act 2010*) and other characteristics on which we deemed it would be unfair to discriminate. We estimate the performance of the model broken down by group: age group, gender, disability and ethnicity (except for LA3 where a data extract on ethnicity was not available).

We calculated the pinned average precisions, false discovery rates (false alarms) and false omission rates (children at risk missed) for each subgroup (e.g. for age: Under 1 Year, 1 - 4 Years etc) and estimated confidence intervals using bootstrapping. The tables for each outcome predicted are detailed in Appendix 1 under 'Fairness Metrics.' The averages of the scores for each subgroup is included in Table 8 below.

Table 8: Average performance metrics (pinned average precision, false discovery rate and false omission rate) by subgroup membership (age group, gender, disability, ethnicity)

Characteristic	Pinned average precision	False discovery rate	False omission rate
Under 1 Year	0.21	0.74	0.08
1-4 Years	0.12	0.76	0.05
5-9 Years	0.13	0.75	0.05
10-15 Years	0.15	0.72	0.05
16+ Years	0.1	0.77	0.04
Missing age	0.29	0.57	0.07
Female	0.14	0.77	0.05
Male	0.14	0.76	0.05
Unknown, Unborn or Indeterminate	0.14	0.82	0.06
Disabled	0.21	0.68	0.08
Not Disabled	0.15	0.83	0.06
Missing Disability	0.17	0.79	0.06
Asian / Asian British	0.16	0.74	0.03
Black / African / Caribbean / Black British	0.16	0.69	0.06
Mixed Ethnicity	0.12	0.71	0.04
Other Ethnicity	0.11	0.78	0.04
Ethnicity Not Known	0.1	0.79	0.02
White	0.15	0.69	0.05

Out of these metrics, the one which causes most concern is the mean false discovery rate (false alarms). The metric ranges from 0 to 1 with 0 representing a perfect model - 0.75 means that 75% of the cases identified as at risk of the outcome are false alarms. The false omission rate is low (again the metric ranges from 0 to 1 with 0 representing a perfect model) - 0.05 means that 5% of cases identified as not at risk are cases which are actually at risk.

Considering all the models together, do the models tend to have higher error rates on any particular subgroup?

We combine the pinned average precision from the 32 models to test whether the models perform consistently better or worse when predicting on some subgroups. As the tests require an equal number of observations in each comparator, we standardise subgroups across different local authority datasets or exclude the subgroup from the comparison where standardisation of the definition was not possible. We compare the following groups:

- **Age group:** Under 1, 1-4 years, 5-9 years, 10-15 years and 16+ years (Missing age is excluded)
- **Disability:** disabled and non-disabled (Missing disability is excluded)
- **Gender:** male and female (Unborn, Unknown / Indeterminate is excluded)
- **Ethnicity:** Black / Black British / African / Caribbean (combining Black / Black British and Black / African / Caribbean / Black British); unknown ethnicity (combining Unknown, Declined / Missing, Ethnicity not given); Mixed ethnicity (combining Mixed / multiple ethnic groups and Mixed ethnicity), Other ethnicity (combining Other ethnic groups and Other ethnicity); White (Arab, Asian / Asian British / Chinese are excluded)

The choice to exclude some groups does not mean that we think they are unimportant or would not be concerned about misclassification of individuals from these subgroups but simply reflects what was possible to test with the data available (these categories were not available for all of the local authorities). In excluding these subgroups, this means that we would fail to detect differences in scores should they exist. This is of concern to the extent that we think subgroups which tend to be poorly defined are at risk of the model misclassifying them. Beyond an expectation of small sample sizes for some of these subgroups - which may mean that the model has trouble learning generalisable patterns - we have no prior as to whether these subgroups are likely to be misclassified. But we note this as a limitation of the approach.

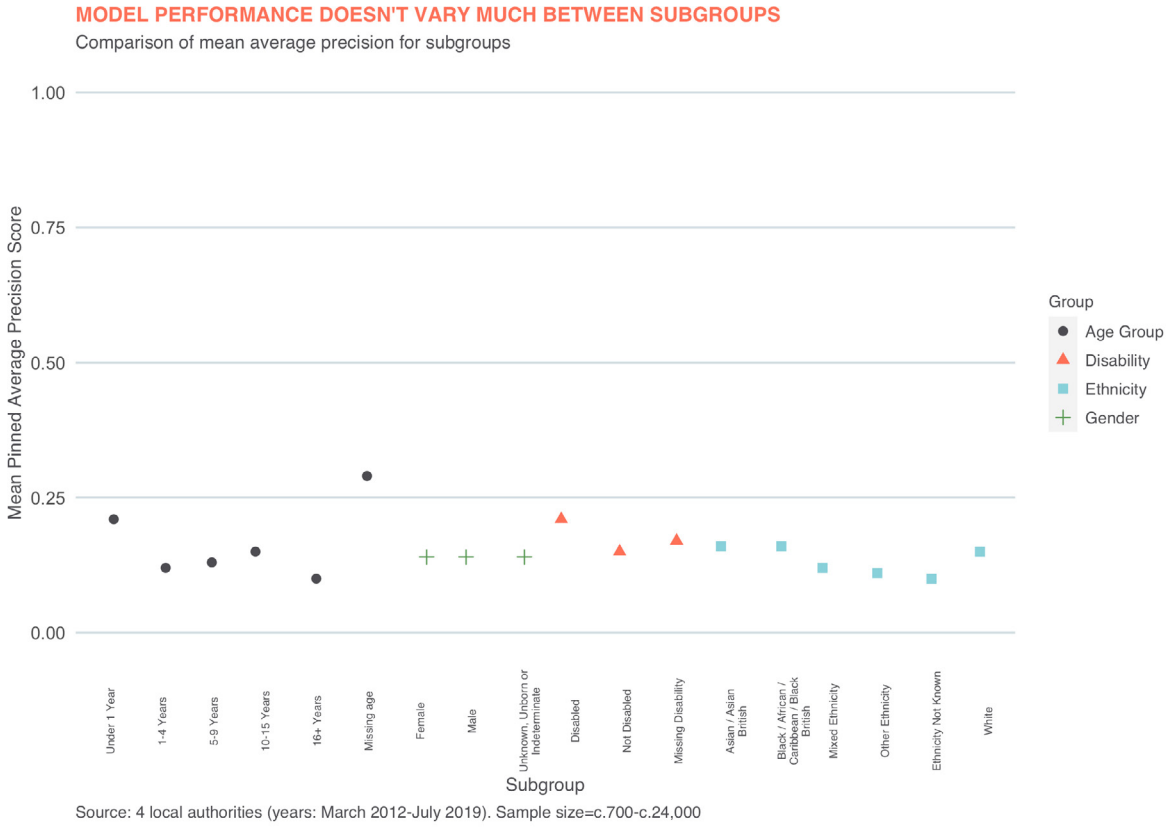
We use non-parametric tests which compare the rankings of the performance metrics of the models: the Mann Whitney U test where the comparison is limited to two subgroups, and the Friedman test for groups with more than two subgroups. We fail to reject the null hypothesis of an equal distribution of pinned average precision score for the following groups:

- **Disability (Mann–Whitney U = 600.5, n1 = n2 = 31, P=0.15, two-tailed)**
- **Gender (Mann–Whitney U = 530.0, n1 = n2 = 32, P=0.81, two-tailed)**

However, we reject the null hypothesis of equal distributions of pinned average precision for age group (Friedman chi squared statistic =46.1, n1 = n2 = 32, P=0.00, two-tailed) and ethnicity (Friedman chi squared statistic =10.7, n1 = n2 = 24, P=0.01, two-tailed). The Nemenyi post-hoc test, which conducts pairwise comparisons of the subgroups, identifies the scores for the 16+ years age group as significantly different at the 5% significance level from those for all other age groups except 1-4 year olds. The scores for under 1 year olds are also significantly different from all other age groups except 10-15 year olds. Conversations with colleagues from the local authority partners suggest that the models performing worse for 16+ year olds may be explained by care plans looking considerably different for 16+ year olds than for other age groups. The subgroup whose ages are not recorded are not included in this comparison because there were only two outcomes predicted for which input data on age was missing, and the Freidman test requires an equal number of observations of each group compared.

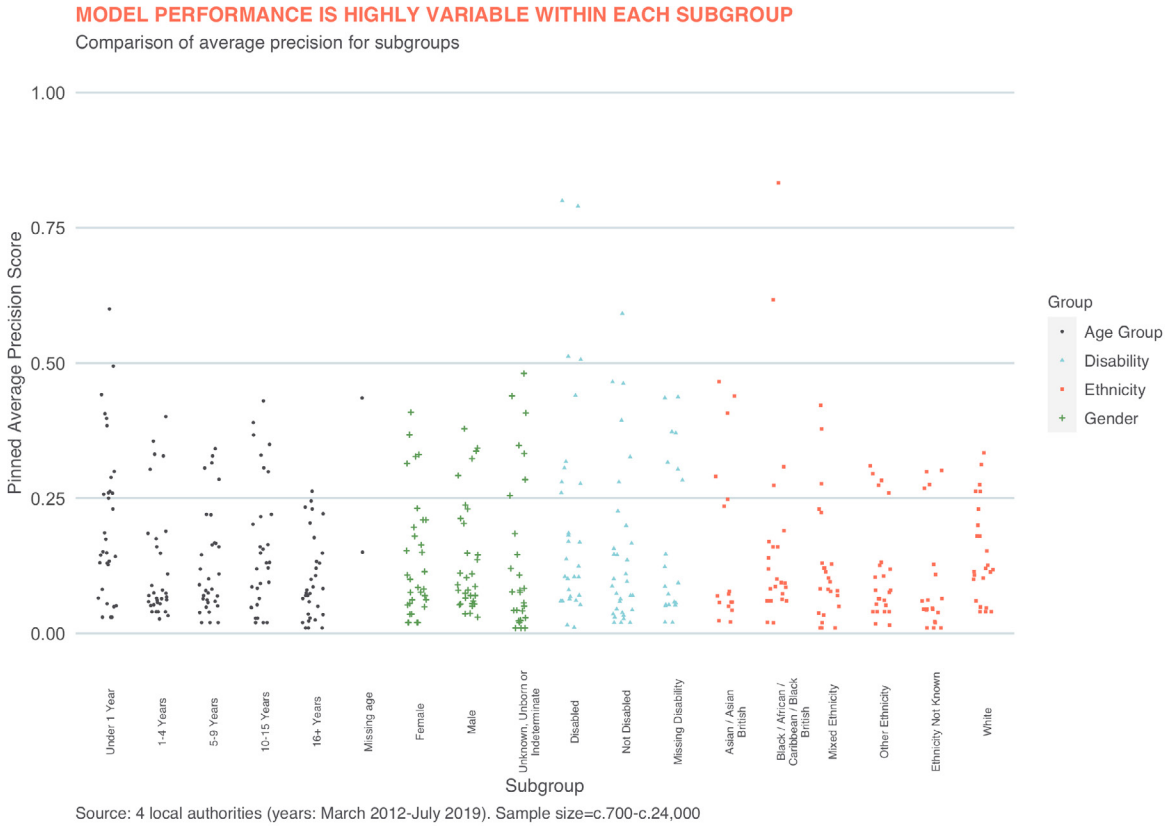
Visually inspecting the average scores by subgroup displayed in Figure 11, the model appears to perform better for those with missing ages; however, this is likely to be a very small group with quite unusual circumstances given that age is required to be recorded and reported in the Ofsted and the Department for Education datasets. Similarly to missing age, the subgroup whose ethnicity is not known is not included in this comparison because this category was not available for all outcomes predicted. However, looking at Figure 11, it appears that the models tend to perform worse for those whose ethnicity is not known.

Figure 12 plots all of the models' scores for each subgroup. We can see that there is a large spread of scores within each subgroup.



Source: 4 local authorities (years: March 2012-July 2019). Sample size=c.700-c.24,000

Figure 11: Model performance doesn't vary much between subgroups



Source: 4 local authorities (years: March 2012-July 2019). Sample size=c.700-c.24,000

Figure 12: Model performance is highly variable within each subgroup

For each model, does the model make more errors for some subgroups than others?

The findings described above give a bird's eye view of whether the 32 models tend to have lower scores for particular subgroups. However, we wished to check whether each individual model made more errors for some subgroups than others given that such bias checks would happen for each model were a model being considered for deployment in practice. We use two different methodologies to test whether this was the case: whether the confidence intervals of the scores overlapped, and whether the null hypothesis of no difference in ranking of the bootstrapped scores was rejected in non-parametric tests. We use the two methodologies to evaluate whether the models perform systematically worse for some subgroups than others because the methodologies make different types of error in detecting differences: whether the confidence intervals overlap is more prone to Type 2 errors (missing when there is a significant difference) whilst the non-parametric tests are more prone to Type 1 errors (especially with large sample sizes - the number of bootstrapped scores is 500 - and when comparing groups with varying standard deviations.[19]

Confidence intervals

We looked at whether the confidence intervals of the pinned average precision scores overlapped in pairwise comparisons of subgroups. This methodology is liable to Type 2 errors where we risk failing to detect a significant difference between the scores of two subgroups. Using this methodology, we detected 10% of subgroups which were significantly different from another subgroup in pairwise comparisons. Although there are multiple comparisons made there is no intuitive way to correct for the multiple comparisons using this methodology. These are summarised in Table 9 below.

Table 9: Percentage of pairwise comparisons where the confidence intervals are non-overlapping for each subgroup

Characteristic	Percentage of pairwise comparisons where the confidence intervals are non-overlapping
Under 1 Year	26
1-4 Years	30
5-9 Years	14
10-15 Years	16
16+ Years	69
Missing age	50
Female	10
Male	21
Unknown, Unborn or Indeterminate	21
Disabled	2
Not Disabled	9
Missing Disability	2
Asian / Asian British	8
Black / African / Caribbean / Black British	0
Mixed Ethnicity	2
Other Ethnicity	6
Ethnicity Not Known	12
White	9

Non-parametric tests

We use non-parametric tests which compare the rankings of the performance metrics of each model

19 Zimmerman, D. W. (1999). Type I Error Probabilities of the Wilcoxon-Mann-Whitney Test and Student T Test Altered by Heterogeneous Variances and Equal Sample Sizes. Perceptual and Motor Skills, 88(2), 556–558.

for each subgroup. Bootstrap fitting of the model (sampling the training data with replacement, fitting the model and scoring it by testing how well it predicts on unseen data 500 times) gives us a distribution of scores to conduct the tests on. We use the Mann Whitney U test where the comparison is limited to two subgroups, and the Friedman test for groups with more than two subgroups. For all of the groups for all of the 24 models with the exception of Gender for one of the models predicting escalation to CLA after assessment, the null hypothesis of no difference in ranking is rejected. We use the Hochberg's step up correction for multiple comparisons within each group. Using this methodology 90% of the pairwise comparisons of subgroups' scores were significantly different. We do not include a breakdown of which subgroups have significantly different scores as it is not meaningful to look at the percentage of pairwise comparisons that were statistically different by subgroup for this methodology because for pairwise comparisons with the p-values < 0.001, ordering the subgroups by the p-value to compare to the adapted Hochberg's significance threshold is rather arbitrary.

The results from these two methodologies point in opposite directions: whether the confidence intervals overlap points towards a low proportion of the models performing worse for certain subgroups whilst the non-parametric tests point towards nearly all of the models performing worse for certain subgroups. On the one hand, failing to detect that a model may be biased is problematic in that models may be deployed without due care but on the other hand knowing that 91% of pairwise comparisons are significantly different does not inform us whether we should be concerned about any particular subgroups. Given the very different pictures painted by these two methodologies, we are uncertain as to whether to conclude that these results indicate that the models perform worse for certain subgroups. We strongly recommend data teams conducting extensive sensitivity analysis to understand the impact of errors on different subgroups.

RQ4: Are the probabilities predicted statistically different (i.e. when the model makes a prediction in the form of a probability, how much confidence can we have in it)?

Please note that we did not prespecify how we would aggregate the results from the 32 models. We hence describe the choice of analysis in some depth below.

Research question 4 informs us about the appropriate level of granularity for the model to give as an output. The motivation is that giving a unit percentage probability of risk, e.g. 89% or 74%, is likely to be overly granular given the quantity and quality of the data, and the complexity of the question. Presenting an overly granular output risks social workers over-interpreting and relying too much on the predictions - false certainty is damaging given that the model can make mistakes and given the complexity of the outcomes we are predicting. But presenting an output which is not granular enough is not likely to be helpful e.g. whether the case is 0-50% or 50-100% likely to be at risk of the outcome.

The prediction interval summaries are calculated for models from LA2, LA3 and LA4 due to time constraints on running the bootstrap fitting of the models at LA1. For all models, the average width of the 90% prediction intervals for all models is 0.0126. This means, on average, we can be confident that a future observation will fall within an interval of 0.0126 around the predicted probability outputted by the model, with 90% probability. The average width of the prediction interval at the threshold which maximises the f-beta score (with beta = 0.1) is 0.0561. This is over four times the average prediction interval - that it is wider than average is not too surprising given that the model is likely to have higher uncertainty around borderline cases - and this is still a relatively narrow interval. We choose the threshold which maximises the f-beta score instead of the average precision score as the f-beta score also summarises this tradeoff between precision and recall but allows

the researcher to specify the relative weight of the metrics. We choose to weigh precision as 10 times more important than recall because of concerns over the number of false positives.

Looking at the prediction interval for thresholds from 0.1-1, the intervals are slightly wider for the thresholds at the middle of the range (0.3-0.6). Again, this is expected for similar reasons outlined above in that observations with predicted probabilities around these threshold values are more borderline cases. The widest prediction interval of 0.0735 at the 0.5 threshold suggests that binning the probabilities returned by the model into bins of 10% would be appropriate from a statistical point of view. More research would be required to understand whether an output from the model of, for example, "he model is 20-30% likely to escalate' is salient to social workers.

There appear to be no significant differences in the distribution of the prediction intervals when comparing the data included (Mann–Whitney U = 5551.5, n1 = n2 = 108, P=0.54, two-tailed) or the cross validation methodology (Mann–Whitney U = 5541.5, n1 = n2 = 108, P=0.53, two-tailed).

Table 10: Precision, recall, f-beta score and prediction interval averaged by threshold for all models

Threshold	Precision	Recall	F score (beta = 0.1)	Prediction interval (threshold +/- 0.03)
0.1	0.2371	0.3658	0.2354	0.0426
0.2	0.2738	0.2688	0.2688	0.0587
0.3	0.2883	0.235	0.2800	0.0671
0.4	0.3242	0.2075	0.3071	0.0719
0.5	0.3379	0.1629	0.3208	0.0735
0.6	0.3329	0.1300	0.3108	0.0721
0.7	0.3212	0.1012	0.2975	0.0678
0.8	0.3371	0.085	0.2971	0.0589
0.9	0.2875	0.0483	0.2321	0.0438

Note: LA1 excluded due to lack of time to bootstrap the models

Technical Research Question 5: What is the semantic coherence and the exclusivity of words of the topics?

As explained under 'Topic modelling performance' of the Methodology section, we took a slightly different approach to this question than originally intended. Instead of the semantic coherence and exclusivity of the words of the topics, we report the log likelihood of topic models (please see Appendix 1 under 'Latent Dirichlet Allocation: Log Likelihood of topics').

Where the text data was available for the duration of the analysis (for LA1 and LA4), we included the text as a list of strings within the cross validation to optimise the average precision of the model. The topics thus served as features in the model instead of aids for interpretation of the models' results.

Where it was not possible to maintain access to the text data for the analysis (for LA2 and

LA3, prediction outcomes 3-6), we conducted a separate cross validation for the best topics optimising for log likelihood and then constructed a dataframe of the term frequency inverse document frequency (TFIDF) matrix, topic proportions and text features to append to the structured data in order to create our structured and text data dataset. We searched over the number of features to feed into the TFIDF matrix and the number of components for the topic models. We experimented with varying batch sizes, learning decay values and maximum numbers of iterations but these did not influence the log likelihood much for the values we tried. We plotted elbow plots (number of topics vs negative log likelihood score) and selected the number of topics by looking for the topic number at the 'crease' of the elbow, the number of topics at which there is little marginal improvement in the log likelihood from adding more components.

We evaluated the topic models by 'eyeballing' the top 15 words representing a topic and plotting word clouds to show how the word frequencies varied by topic. The word clouds can be found in Appendix 1 under 'Word Clouds'. In some cases, for example, 'Word cloud 3c1: Predicting open case after 'No further action': Word Cloud - Previous referral records (structured and text data; learning from all cases)', it appears that the topic model is picking up on a different level of risk with topic 0 focusing on advice whilst topic 1 suggests more action is required. However, in most cases the words are too similar to distinguish well between the different topics. It seems that the topic modelling does not lend itself to aiding interpretability when the topic space is quite constrained in contrast to, for example, a problem such as to categorising news articles which could be written on a wide range of topics. We also compared documents which were canonical examples of each topic to evaluate whether the distinctions the topic model made matched human judgement. Due to the length of the documents, this was difficult to confirm that the distinctions from the topic modelling matched human judgement.

Given that the text data did not improve model performance, models including text data tended to overfit more than those just using structured data and the availability of the text did not support the interpretability of the models, we found that text data was not a useful addition.

Technical Research Question 6: What is the performance and predicted performance of the models on different sample sizes (i.e. for different sized local authorities)?

Please note that we did not prespecify how we would aggregate the results from the 32 models. We hence describe the choice of analysis in some depth below.

Whilst it is difficult to identify whether the differing performance metrics for the models at the different local authorities are due to the size of the local authority or due to any of the other myriad of factors which different between local authorities, we can simulate different sample sizes for each predictive question within each local authority. Whilst this doesn't allow non-participating local authorities to look at the results from the most similar local authority in terms of size and infer how well machine learning would work for them; it is a helpful exercise in hypothesising whether waiting to allow data to accrue would improve the scores of these models.

For LA2, LA3 and LA4[20], we trained models feeding them successfully large subsamples of the training data, and plotted learning curves of the average precision score vs the sample size for the models predicting on the training dataset (i.e. the data the model has learnt the patterns from) and predicting on test folds within cross-validation. The learning curves are available in Appendix 1 under 'Learning Curves'. The researchers then classified the learning curves according to whether the score increased with sample size (which would suggest that adding more data would help improve the score) or whether the curve plateaued as the sample size

20 Analysis at LA1 was already complete by the time that we plotted learning curves for LA2, LA3 and LA4.

increased (which would suggest that adding more data would not help improve the score). For some learning curves, the average precision score remains low irrespective of increases in sample size. This is the case for example with 'Figure 3b: Predicting open case after 'No further action': Learning curve showing the average precision score for models fitted on increasing sample sizes of data (structured data only; learning only from earlier cases)'. Although the score does not increase with the sample size on the graph; the large gap between scores calculated on the training dataset and within cross validation is indicative of the model overfitting to the training dataset. This is where the model learns patterns from the dataset but these fail to generalise to other cases. More data can help reduce overfitting by giving more examples for the model to learn generalisable patterns from.

Table 11: Number of models for which the score increases when sample size increases for each local authority

LA	Learning Curve plateaus	No increase in score with sample size but evidence of overfitting	Learning curve not reached plateau
2	3	2	3
3	5	2	1
4	2	4	2
Total	10	8	6

It seems that for 14 out of the 24 models, increasing the sample size of the dataset would improve the model performance (as the learning curve has either failed to increase at all or has not yet reached a plateau). When looking at the breakdown by cross validation methodology (Table 12), more data is most likely to improve the scores for models restricted to learning from earlier cases, the model design we evaluate as being closest to simulating the model performance if it were to be deployed in practice. There appears to be no difference as to whether more data would help when comparing models including just structured or structured and text data (Table 13).

Table 12: Number of models for which score increases when sample size increases by cross validation methodology

Cross-validation	Learning Curve plateaus	No increase in score with sample size but evidence of overfitting	Learning curve not reached plateau
Learning from all cases	7	0	5
Restricted to learning from earlier cases	3	8	1
Total	10	8	6

Table 13: Number of models for which score increases when sample size increases by data included

Cross-validation	Learning Curve plateaus	No increase in score with sample size but evidence of overfitting	Learning curve not reached plateau
Learning from all cases	7	0	5
Restricted to learning from earlier cases	3	8	1
Total	10	8	6

These results offer some support for the hypotheses discussed above under 'Why is model performance so low and do we expect better model performance as techniques advance?'. In particular, the hypothesis that the additional restriction on the model to learn from just earlier cases requires more data for the models to perform well. However, as discussed, it is likely to

be challenging to obtain data which can improve the performance of the model.

Technical Research Question 7: What is the performance of models including and excluding data before major changes (e.g. in practice -- ways of recording, funding i.e. to understand whether patterns learnt on data collected before the change are helpful to predictions after the change)?

For at least three out of the four local authorities, there were no major changes in practice or the recording of the data for the years under consideration and so answering this question was not feasible with the data available. However, given that the models performed below the threshold before taking into account any changes, whether the models perform worse after a change in practice is unlikely to be a material factor in a local authority's decision as to whether to pursue machine learning in children's social care.

Technical Research Question 9[21]: Do social workers find the outputs of the model a useful addition to tools and information they already have access to?

As the performance of the models were below the threshold we deemed as indicating success, we felt that it would be inappropriate to spend social workers' time on giving feedback on how interpretable the models were. This was of particular concern as the analysis for LA2, LA3 and LA4 were finished during the Covid-19 outbreak when social workers' time was especially scarce. Given these additional considerations, whilst how social workers would interact with the outputs of models is an important empirical question, we decided that it would be appropriate to wait until we had more confidence that the models were performing well and giving the social worker the appropriate information before embarking further down this line of enquiry.

However, we still produced several types of visualisations to test the interpretability of the models and have included them in Appendix 1 so that the reader can judge for themselves whether they would find these a useful accompaniment to a model output such as a ranking of cases, a 'red flag' to indicate risk or a classification of the case as 10-20% likely to be at risk. The word clouds (discussed above and found under 'Word Clouds' in Appendix 1) visualise the topics that represent the corpus of documents and in theory could be used to aid visualisation by presenting the topic that best represents the document. However, given the small number of topics, the constrained topic space and that documents are usually some represented by some combination of topics, this way of representing the text in the document feels crude and a poor substitute for reading the text itself.

We also present feature importance plots in Appendix 1 (under 'Feature importance plots') to allow for an understanding of how important the feature is in contributing to the predictive power of the model. For readers familiar with regressions, coefficients in a regression are one way of conceiving of feature importance but given that we also use algorithms not based on regression techniques we used an algorithm agnostic way of calculating importance (permutation importance). The feature importance plots show the features with the highest importance scores at the top and the aim is to give a sense of how useful the feature is in prediction in comparison to other features. However, the plots are limited in that they don't give an indication of the direction and shape of the relationship between the feature and the outcome. For example, looking at the feature importance plot for the model using structured data and learning from all

21 The research protocol outlined research questions 1-7 and research question 9, erroneously omitting a research question numbered 8. For ease of comparability with the research protocol, we maintain the original numbering.

cases to predict an open case after 'No further action' (Feature importance plot 3a), we can tell that the average length of previous assessments is important and which day the referral starts is important but if you told us that information about a case, we would not know whether the child was more or less likely to be at risk from the outcome. For this reason, feature importance plots are often paired with accumulated local effect plots, which show, conditional on a given value, the relative effect of changing the feature on the predicted outcome. The ALE is the difference from the mean prediction when the feature is changed to that particular value. Given that age is a feature with high importance in many of the models and that those familiar with the children's social care context are already likely to have an intuition of how the risk of various outcomes changes with age, we plot the ALE plots for age for the models. The ALE plots seem relatively stable between models, showing similar relationships between age and the outcome being predicted. This stability of the relationships between models is somewhat reassuring.

Although we did not seek social worker feedback on the interpretability of the model outputs which would have required a considerable portion of their time, we sought their general views about the use of machine learning in children's social care through polling the WWCSC social workers panel[22]. 129 of the panel responded. As can be seen in Figure 13, 29% of participants did not know what predictive analytics was whilst 10% of participants thought that predictive analytics had a role to play in decision making in social care, 29% thought it didn't and 32% weren't sure.

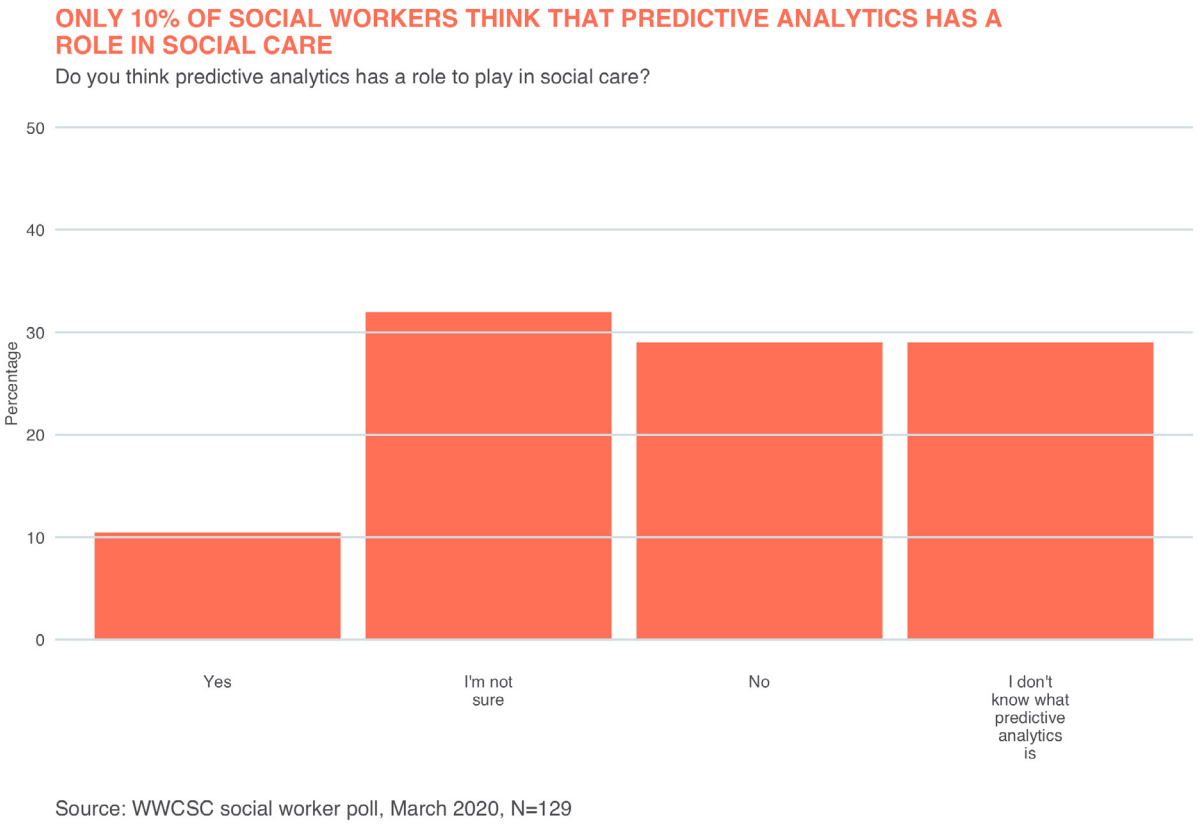


Figure 13: Only 10% of social workers think that predictive analytics has a role in social care

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

We then asked 'If predictive analytics were to be introduced in children's services, which of the following uses might it be acceptable?' As can be seen in Figure 14, 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). 14% of participants thought producing a tool which would help social workers prioritise cases to discuss with their managers would be acceptable, 14% thought that producing a tool that helps decide whether a family meets the threshold for social work intervention would be acceptable and 12% thought that producing a tool which helps to decide whether to close a case would be acceptable.

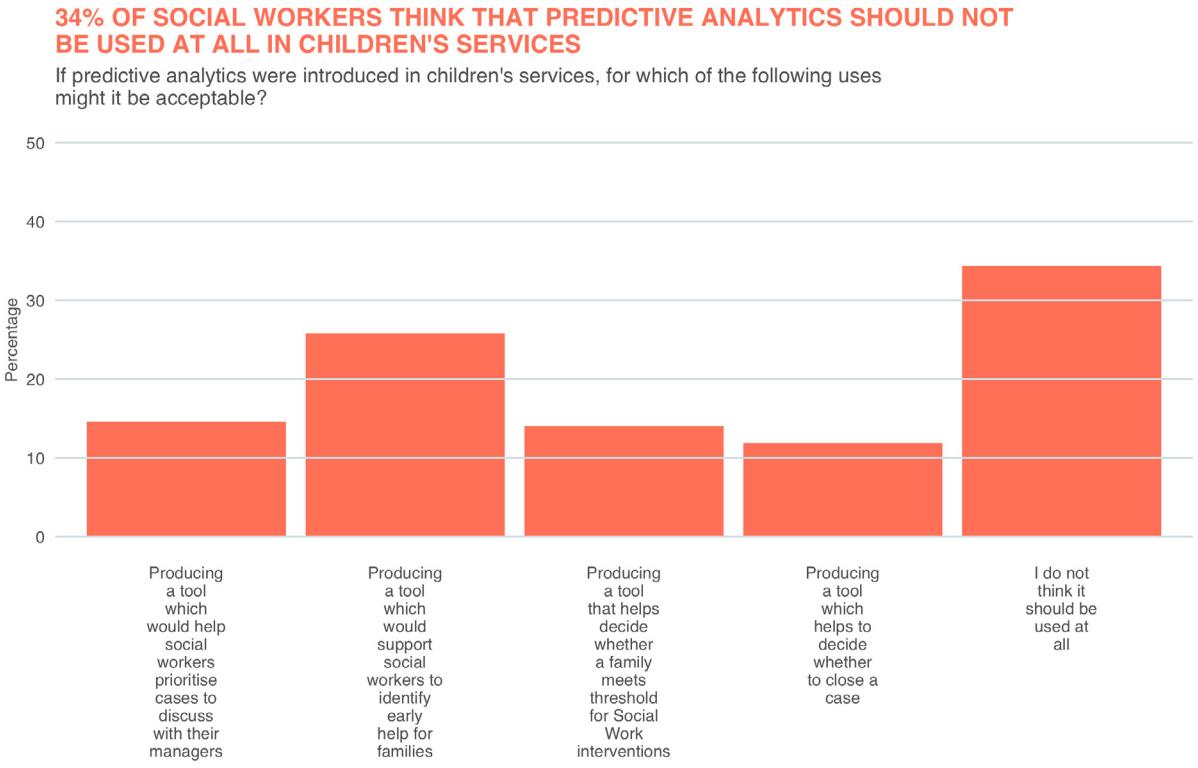


Figure 14: 34% of social workers think that predictive analytics should not be used at all in children's services

The panel was not chosen to be representative of the social work profession in England and the sample size (129) is small relative to the profession (32900[23]) so these results should be taken with a considerable pinch of salt. We also do not know whether social workers would update their views on acceptability if we had found that the models were effective in identifying cases at risk. However, these results give weak evidence of low levels of acceptability of the use of such tools within the profession.

23 The 'headcount' of social workers employed by local authorities and agency social workers in the year ending 30th September 2019. Department for Education. (2020, February 27) Official statistics: Children and family social work workforce in England, year ending 30 September 2019. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/868384/CSWW_2018-19_Text.pdf

DISCUSSION

None of the models performed well enough to exceed our minimum threshold of a 'well performing' model (a threshold which was not set ambitiously). The models tend to miss children at risk of the outcome. When looking at the breakdown of types of errors on 1000 cases, 17 / 29 (58%) of cases identified as at risk are false alarms. If the model were to be used in practice, this means that when the model flags a child as at risk, it is a false alarm in 58% of instances. The average model also fails to identify 46 / 58 (79%) children who are at risk. 'False alarms' may encourage intervention where it is not necessary, and also create an unnecessary burden of additional follow-up and assessments. If the model falsely identified many cases as at risk of the outcome, there is the risk of 'cry wolf' where social workers (rightly so given the error rate) would not trust the model identifying a case at risk.

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. Our understanding is that this focus reflects the focus of projects at local authorities piloting machine learning (which we estimate to be approximately 10% of children's services). 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[24] 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 type of data is different (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.

There are several factors which we suspect make machine learning challenging in the context of children's social care. The outcomes we

24 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

predicted happen to a small percentage of the relevant population and so the model is 'looking for a needle in a haystack.' Because outcomes for siblings tend to be similar (and to a much lesser extent outcomes for cases assessed by the same social worker are more similar than outcomes for cases assessed by a different social worker), this reduces the amount of additional information the model can glean from siblings' cases and cases assessed by the same social worker in the dataset. As we are predicting outcomes on reasonably long time scales (1 month - 12 months), it is important to restrict the model to learning only from cases preceding it so that it does not learn from patterns which emerge in between the point in time at which we are predicting and the outcome being realised. This further reduces the number of cases available for the model to learn from. Given the complexity of social care outcomes and the myriad of pathways a child or young person's journey can take, the models seem to struggle to pick out general patterns amongst the nuance of the small number of particular cases.

Our analysis of how model performance changes with sample size suggests that more observations would be helpful, particularly when the models are restricted to just learning from earlier cases. However, given the way that children's services are set up in England, it is difficult to increase the sample size by collating data from different local authorities given that practice models and case management systems can differ greatly. Even overcoming the practical difficulties of bringing together the data, we think it unlikely that the additional models will be helpful given the different ways of working and contexts of the different local authorities. Our findings also suggest that sourcing more data from case management system archives or waiting for more data to accumulate would not be helpful as the patterns seem to change enough over time that adding this data wouldn't aid the performance of the model. These considerations leave not much space to maneuver to try different types of models or tweaking parameters to improve model performance. It may be that outcomes which look less like a 'needle in a haystack' would be easier to predict. Given that the models tend to overfit and learn non-generalisable patterns, adding richer information about each individual case would not necessarily help; however, if such data is highly predictive, this may be a suitable strategy.

Our analysis of whether the models are biased is unfortunately inconclusive. The findings are very sensitive to the methodology used to test for bias. It is disappointing not to have a clearer picture of bias; however, given the poor performance of the models overall whether the models are biased is unlikely to be of material consequence in local authority decision-making in whether to pursue the development of or procurement of predictive tools. However, we continue to encourage data teams whose models meet a basic efficacy threshold to conduct extensive sensitivity checks to understand the impact of model errors on different subgroups before using any models in practice.

Using results from polling our social worker panel, we found weak evidence that the acceptability of the use of predictive analytics in children's social care is low amongst social workers with only 10% saying that they consider it acceptable. About 14% thought that if predictive analytics were to be used in social care, a use case similar to the one that would be appropriate for the models built - a tool that helps decide whether a family meets the threshold for social work intervention - would be acceptable. Although we did not conduct a cost benefit analysis as part of this work, the costs associated with developing such systems should be considered alongside the purported benefits. As outlined in a report[25] by the Oxford Internet Institute, machine learning is unlikely to save local government at least in the short term because funds need to be available for early intervention for those identified as at risk.

In summary, we do not find evidence that machine learning techniques 'work' well in children's

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

social care. In particular, the desired benefit of being able to identify children and families early in order to better support their needs doesn't seem to play out in this case. This type of pilot research does not speak to the potential risks discussed initially which become a risk when the model is deployed, for example, that predictive tools being used as a crutch and an alternative to professional thought and judgement but such questions assume models of a minimum threshold of efficacy.

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 for the most common purposes in children's social care. For local authorities already piloting machine learning, we encourage them to be transparent about the challenges they experience also. 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.

What Works *for* Children's Social Care

CONTACT

info@whatworks-csc.org.uk

@whatworksCSC

whatworks-csc.org.uk