PROJECT TITLE	Pilots of predictive analytics in children's social care
LEAD ORGANISATION	What Works for Children's Social Care (WWCSC)
PRINCIPAL INVESTIGATOR	Michael Sanders
PROTOCOL AUTHORS	Vicky Clayton, Daniel Bogiatzis Gibbons, Michael Sanders
TRIAL DESIGN	Analysis of pre-existing administrative data using machine learning models.

Executive Summary

Background

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 and a small but increasing number of local authorities in the UK, predictive models are starting to be used to assist social workers by predicting outcomes relating to child protection. In their favour, they may potentially be helpful as decision aids to social workers when undertaking assessments, or identifying the cases most at risk to team managers. However, there are unsolved questions about the accuracy of the models and for which situations it is ethical and acceptable to use them.

To try to answer whether predictive models should be used in children's social care in a holistic way, What Works for Children's Social Care:

- is working with six local authorities on a technical feasibility pilot to assess the accuracy of the models,
- is conducting ongoing research into the acceptability of the use of predictive models in children's social care and
- has commissioned an independent ethical review.

The findings from this research could help local authorities understand whether it's worthwhile investing in developing these types of models and associated tools to assist social workers in practice.

Research Aims

This protocol outlines chiefly the technical feasibility pilots. The aim of the technical work is to assess the accuracy of these models for a range of different research questions across local authorities of a range of sizes, practice models and case management systems to understand which outcomes the models might predict well, as well as some of the practical obstacles (ease of data extraction from the case management system, quality of data etc). A key research question is whether the inclusion of referral and assessment reports substantially improves the accuracy of the models.

Research Design

The research involves building predictive models and evaluating their accuracy. The accuracy of the model tells us how many errors the models make, which is important to know if we are testing whether they would be useful to assist decision-making. In order to test whether referral and assessment reports improve the models, we conduct "topic modelling", a natural language processing (NLP) technique to find human-recognisable themes, and feed the topics into the models. We anticipate that being able to incorporate the information from referral and assessment reports is 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. We also carry out additional analysis to understand whether the model is biased against particular sub-groups and explore different ways of explaining the outputs of the model. We assess the accuracy of the models by predicting the outcomes for children and young people from unseen historical data. We are not testing the models in practice by doing any "live" predictions. No decisions regarding any individual cases will be taken by, or as a result of, the model.

Outcome Measures

We are trying to predict:

- Are children / young people being re-referred within 12 months of having their case being designated "no further action" (NFA) or their case closing?
- Are children / young people who have been referred to early help being stepped up to a Child in Need (CIN) or Child Protection (CP) plan within a 12 month period?
- Following assessment are children / young people being stepped up to a Child Protection (CP) plan or being looked after?
- Are children / young people being re-registered as CIN within 12 months of their case being designated "no further action" NFA or their case closing?
- Are children / young people registered on a Child in Need (CIN) plan being escalated to a Child Protection (CP) or being looked after within 4 years of initial registration?

Analyses

For each prediction question, we try a range of binary classification algorithms (decision tree, logistic regression, and gradient boosting) which differ in complexity and cost to develop and maintain (if a local authority were to develop it themselves). For topic modelling, we consider two algorithms: Latent Dirichlet Allocation (LDA) and Structured Topic Model (STM). LDA is probably the most common type, and STM allows us to incorporate other information about the case. We validate the models using cross-validation and optimise for the area under the curve (AUC) of the precision-recall graph which allows us to balance different types of errors. We evaluate the AUC precision-recall of the final model chosen on unseen data.

Correspondence

If you'd like to get in touch about the project, please email: wwcprogrammes@nesta.org.uk

Contents

Introduction **4**

Outcomes and Data **13**

Model Specification 24

Reporting **27**

Risks **32**

Timeline **33**

Ethics **34**

Data Protection **39**

Introduction

Motivation

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 the UK, predictive models are starting to be used to assist social workers by predicting outcomes relating to child protection. In their favour, they may potentially be useful to signpost early help or act as a decision aid when undertaking assessments and the downstream impact of outcomes for the child / young person and family, and the associated savings on more intensive support / intervention. But could be damaging if they are not accurate and / or able to be deployed appropriately by practitioners. There are unsolved questions about the accuracy[2] of the models and for which situations it is ethical and acceptable to use them. Answers to such questions would help us understand the appropriateness of using these models in children's social care and would assist local authorities in their decision-making about whether to invest in developing these types of models (and associated tools to be used to assist social workers in practice) and when not to.

The goal of this research strand is thus to try to answer some of those questions. This protocol outlines the technical piece of this work but the WWCSC has also commissioned an independent review of the ethics of using predictive analytics in children's social care,[3], and also are conducting work to assess the acceptability of the use of these types of models to the sector.

Specifically in this project we aim to assess the accuracy of these models for a range of different research questions across local authorities of a range of sizes, practice models and case management systems to understand which outcomes the models might be accurately predict and some of the practical obstacles (ease of data extraction from the case management system, quality of data etc). This allows us to assess both when the models are missing cases they should identify as at risk, and also when they are identifying cases at risk that are actually low risk. We will also carry out additional analysis to understand whether the model is biased against particular sub-groups and explore different ways of explaining the outputs of the model. We assess the accuracy of the models by predicting on unseen historical data on the outcomes for children and young people. We are **not** testing the models as an "intervention" i.e. we are not testing whether the models work in practice by doing any "live" predictions on cases where social workers are currently making decisions. Although how the models would be used in practice is an important question, it is

[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/).

[2] Except in the section on performance metrics, we shall be using the term "accuracy" in the lay sense and not the technical definition. This is because the more general term "performance metrics" used in data science to describe the performance of the model is often used in the local authority context to talk about the performance of the authority, and we wish to avoid any potential misunderstanding that the performance of the model is being used to assess the performance of the local authority.

[3] What Works for Children's Social Care. (2019). REQUEST FOR QUOTE for an ethics review of machine learning in children's social care https://s29720.pcdn.co/wp-content/uploads/Machine_Learning_in_children___s_social_care_ethics_review.pdf

a question that comes after establishing the basics of whether the model is accurate on historical unseen data.

The WWCSC is in an unusual position to do this work as it is an independent research organisation which does not have a profit incentive related to developing a future tool. This allows us to approach the question with a healthy dose of scepticism and provide reliable evidence on whether the use of such models is helpful to the sector. We commit to publishing the results regardless of the outcome. The three strands (ethics, acceptability and technical feasibility) will inform a series of public reports which to enable an informed conversation in the sector about the use of predictive models.

Partners

Local Authorities

We issued an open call[4] for local authorities to work with us on the project. We are working with six local authorities across a wide range of geographies and sizes. The local authorities are providing the data and guiding the outcomes to predict as well as extracting the data and hosting us to analyse the data.

What Works for Children's Social Care (WWCSC)

WWCSC is cleaning the data, building the models and writing the final report as well as managing the project overall.

Approach

Building responsible models

We believe that building responsible (reliable, effective user-centered) models[5] is important in general but in the sector of children's social care, it is particularly important because the decisions that social workers are making (and which the models would be assisting in if they were to be used in practice) carry so much weight in a child's and family's life. Because of this, we have tried to think carefully about the potential pitfalls of this research. We outline broad principles here and explain our technical mitigations in the appropriate technical section.

I) **Assisting social workers**: for absolute clarity on our position, predictive models are no replacement for high quality social work. Conducting assessments of future risk (the only part of social work that we think predictive models could help with) is merely a part of the complex process of a social worker building a relationship with the family and creating safety for the child.

   Additionally, the models rely on data collected by social workers and so would be no good without social workers gathering the information and carefully weighing it in their reports. We are guided by the principle of the models assisting the social worker to make decisions and not burdening them in the process. We chose questions which are likely to be practically helpful to feed into key decision points.

[4] 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/

[5] See Google, Responsible AI Practices. https://ai.google/education/responsible-ai-practices

For this project, we a) design the outputs of the models in response to feedback from social workers as to what helps them most in their decision-making; b) don't ask social workers to collect additional information to act as input to the model beyond what they are already required to collect. Whilst we do not have control over how local authorities might use similar models in practice, we will address the potential for the models to become a compliance tool and how this might be avoided should they be deemed to be technically feasible, ethical and acceptable.

2) **Assist local authorities:** we have restricted ourselves to using standard issue local authority computers during the heavy data processing phase and freely accessible software to replicate as closely as possible the resources available to local authorities. We attempt to provide both a plain language explanation in this protocol and the research report, as well as the technical details (and code used to build the model) necessary for a local authority to replicate the research[6] on their own (cleaned and processed) data so that it accessible and useful to decision-makers, practitioners and analysts in local authorities.
3) **Transparency**: this includes making the output of the models interpretable and providing the reasons why the model has made that prediction. This also includes being transparent in our research through publishing our analysis plan in this trial protocol and also providing the code used to build the models under a copyleft licence so that it can be inspected and improved upon.
4) **Fairness**: we commit to using processes to attempt to understand and document bias present in the model. Whilst it is out of scope to produce a tool to assist decision-making in practice (during assessments), we shall provide a discussion of the fair handling of mispredictions in the final report, again in the case where the use of these types of models is deemed to be technically feasible, ethical and acceptable.

In scope

Kinds of research questions

The types of models we're exploring predict a binary (yes/no) outcome[7] for an individual e.g. "will child A be referred to children's services within 12 months of their case being closed?"

The questions focus 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.

Because of the data protection setup (see "Data protection" below), we keep local authorities' data separate and build a model for each local authority using their own data, meaning that the number of cases each model can learn from is small relative to the complexity of the prediction we're asking the model to make.

[6] This relies on local authorities having analysts with some (R) coding proficiency. We recognise that this will not always be accessible to local authorities but this setup is the most cost effective whilst still using the necessary tools.

[7] We focus on outcomes which are binary because 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.

This constraint focuses our attention on questions early in a child's journey through children's social care where there are a larger number of cases although we do explore a few predictions involving outcomes later along the journey to understand the constraints of these techniques.

Model outputs

The research questions are chosen to augment social workers' appraisal of levels of risk in the future, for example, predicting a step up from a child being on a CIN plan to a CP (child protection) plan indicates the level of risk may be expected to increase in the future, or that there is a current risk that the model assesses as more problematic than the social worker's evaluation. This may confirm a social worker's decision to step up, or provide an additional lens to understand the case if the social worker intended to not escalate the case.

As a result, the model outputs a probability: for example an 86 percent predicted probability that child A will be re-referred within twelve months, and also an indication of how much each variable -- e.g. mentions of parental alcoholism in the case notes -- contributes to the prediction.

The research also explores how best to represent this output in a helpful way to social workers.

Data included in models

We only use 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. The decision will most likely be within an assessment, either at the beginning of working with a family or as part of ongoing assessment work .

Since a feasible use case of these models is giving social workers additional information at the time they are making a decision about a case, it is important that the data we feed the model accurately represents the data that would be available to the social worker at the time they are making that decision.

Even though it is out of scope to build tools to be used in practice during this research (see "Out of scope" below), it is important that the models reflect the situation in a real setting so that the accuracy is not artificially inflated by the model in this technical feasibility study having prescience (knowledge about an event before it occurs).

Since we follow the principle that the model should be transparent to the social worker, using data from multi-agency partners prior to a section 47 enquiry would not be appropriate as social workers do not have access this information without consent. If the model processed this data it would not be able to give the social worker transparent explanations for why it had made a prediction without breaching data protection law.

We include as inputs to the model: structured data e.g. age, gender etc and text data i.e. data in referral and assessment reports. We do not use all the data accessible to the social worker: some of the information about the case will inevitably be held in the social worker's head and not be recorded in the case management system. Additionally, we deliberately exclude data recorded in diagrams or forms because of the difficulty of automatically processing such data, and structures that will consist mainly of names e.g. genograms because we anticipate that the improvement to the model would not overcome the trade-off with the difficulty of anonymisation (text which is mostly names is unlikely to add much to the model so the case against including such text does not have to be particularly strong). We do not include sibling or other relative's data as inputs for the individual's predictions. Figure 1 shows the thought process for data fields to include.

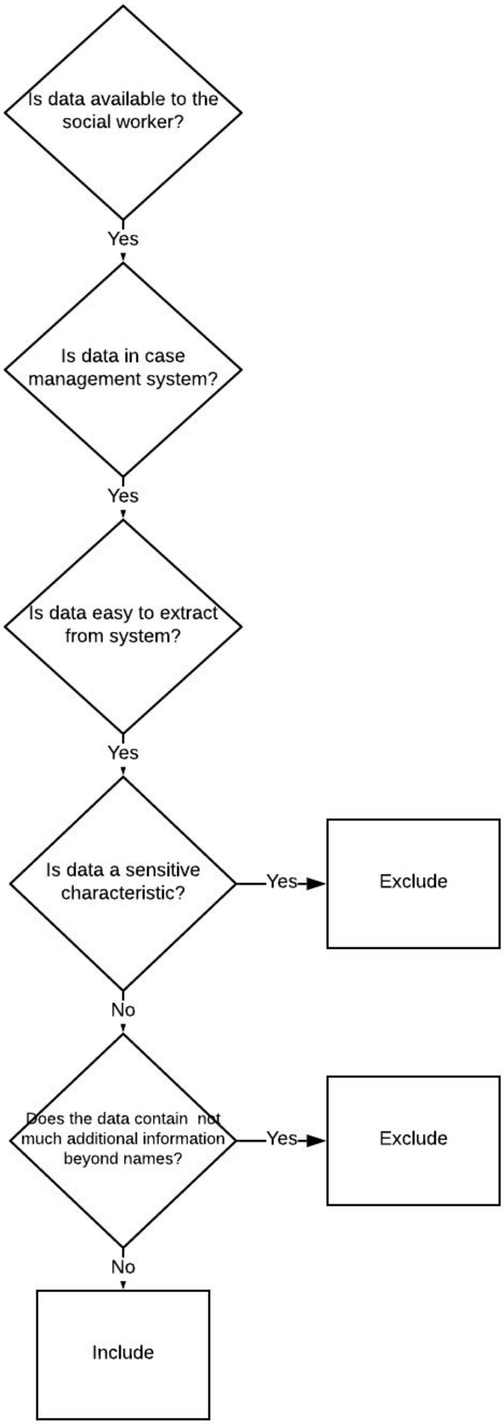
Is data available to the social worker?
Yes
Is data in case management system?
Yes
Is data easy to extract from system?
Yes
Is data a sensitive characteristic?
Yes
Exclude
No
Does the data contain not much additional information beyond names?
Yes
Exclude
No
Include

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

Out of scope

We exclude certain applications of these techniques.

We are not building a decision tool

Most importantly, for this project we are building models for the purpose of assessing the technical feasibility of applying these models to questions in children's social care and we are **not** building a tool for local authorities to use in any "live" way in practice, not even in a trial / test scenario. The accuracy of the models will be tested on historical unseen data. This also has the advantage of giving an insight into the accuracy of the models now rather than having to wait for the child or young person's outcomes to play out.

The usual workflow for building a tool to be used in practice is to build the model using historical data, assess if it adequately meets the need it's designed for without having adverse consequences and then develop a deployment pipeline to allow prediction on "live" data.

We purposely are excluding this last step - developing the deployment pipeline because a) this allows us to remain open to the conclusion that these models are not appropriate in children's social care without any motivated reasoning to reach the favourable conclusion needed for the development of tools; b) our role as a Centre is to enhance the evidence base around practice.

Having said that, we are aware that local authorities are for the most part interested in the technical feasibility of these models for considering their use in tools to assist social workers and managers. For this reason, when making choices about the analysis we have tried to consider what the downstream effects would be were the models to be developed for use in a tool.

Where we discuss the effects of the possible use of these models in a tool, this just indicates our intention for the discussion to be practically useful to local authorities.

We are not doing research on children in the general population

There have been concerns raised during public debate on the issue relating to some existing applications of these techniques when it focuses on flagging "risky" children and families in the general population who have not already been referred to and assessed by children's social services.

Due to the known low acceptability in the sector of that type of application, for the purpose of this research we restrict ourselves to using data and making predictions about children and families who have already been referred to children's social care and been assessed on at least one occasion.

There are good practical reasons for doing so also: without any data from children's social care about the child or family, the model would need data from other multi-agency partners. Bringing together the data systems from such multi-agency partners is a separate and considerable technical challenge with significant data protection implications.

Even for children about whom someone has raised a concern to social services (in some local authorities this is known as a “contact”), there is very little data to use as an input to the model if no assessment follows and so the model is unlikely to be helpful unless an assessment has also been carried out.

Cost effectiveness analysis

We note that the development of tools based on predictive models is often framed with the aim to save on costs,[8] and that local authorities’ decision about whether to invest in models and associated tools is likely to be related to whether they expect the tools to save them money. We do not make any formal cost effectiveness analysis given that the benefits and costs are likely to vary quite a lot depending on local authority contingent factors which affect how they would be developed and used in practice. For example on the benefits side, identifying children at risk but not having the capacity to intervene earlier or not having effective interventions means that the benefits of even the best possible tool would not be realised. On the cost side, the cost of development and maintenance of the models is likely to depend on the local authorities’ in-house capacity. It is worth noting, however, that we expect cost savings from such tools are likely to come from reducing the need for downstream more expensive intervention rather than reducing cost at the point of assessment.

Process

We conduct the analysis at each local authority sequentially (rather than in parallel). We imagine that there will be considerable learning in terms of the technical approach as we go along - we shall publish addenda to the protocols in the case of any major revisions and note any deviances from the protocol in the final report. Local authorities are of sufficient diversity that we think it’s worthwhile attempting the research with all local authority partners even if the first model performances are relatively poor. We also note that we have a limited amount of time at each local authority - we endeavour to complete as much as possible of the planned analysis but will need to be pragmatic about what is achievable in the time available.

Research Aims

Our aim is to offer preliminary answers to questions that local authorities would have if they were considering investing in developing such models (presumably for the purpose of developing tools to assist social workers in their decision making and / or managers). These are a mixture of practical and technical questions.

Practical Questions

- 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?
- What skills and hardware do you need to carry out this type of analysis?
- What is the level of anonymisation of text data achievable by automated means?

Technical Questions

[8] Dencik, L., Hintz, A., Redden, J. & Warne, H. (2018, December). Data scores as governance: Investigating uses of citizen scoring in public services. Project Report. Data Justice Lab, Cardiff University, UK. p.35

We are predicting a range of outcomes within the children's social care journey (step up, step down, case closed). For each outcome, we would like to answer:

- RQ1: What is the performance of the models using structured data (i.e. data that would be recorded in a statutory return like risk factors)?
- RQ2: What is the performance of the models using structured and text data from assessment and referral reports?
- RQ3: What is the performance of the models on different subgroups of interest?
- 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)?
- RQ5: What is the semantic coherence and the exclusivity of words of the topics?

If there is sufficient time and where it is relevant to the local authority:

- RQ6: What is the performance and predicted performance of the models on different sample sizes (i.e. for different sized local authorities)?
- RQ7: 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)?

Research questions 1-2 give an indication of whether the models would be sufficiently accurate to be used in practice and also the value added by including text from referral and assessment reports. Whether including text adds much predictive power is useful to know because it allows us to understand the tradeoff with the disadvantages of using text, namely that it is much more difficult to anonymise than structured data. Including text data presents a challenge ensuring privacy which requires more processing and checks, and crucially restricts the possibility of combining datasets from multiple local authorities and having access to a larger sample size.

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 deem it would be unfair to discriminate. In some areas, it may be difficult to get definitive answers to this question due to the low numbers of cases in particular groups, however we will endeavour to do so.

Research question 4 informs us about the appropriate level of precision 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 precise given the quantity and quality of the data, and the complexity of the question. Presenting an overly precise 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.

Research question 5 allows us to investigate the quality of the topic models built on the assessment and referral reports. To both improve the performance of the models and make the inputs to the predictive model more interpretable, we attempt to draw out the themes or "topics" of the documents. Topic modelling is an automated process of finding words that occur together in the documents. The technique tries to summarise the document by identifying topics rather than lists of words or phrases so they are more comprehensible to humans. Semantic coherence is maximised when the most probable words in a topic frequently co-occur together. Mimno et al. (2011)[9] show that semantic coherence correlates well with human judgment of topic quality. However, Roberts et al. (2014)[10] noted that it is possible to achieve high semantic coherence by having a few topics dominated by very common words. This way of gaming semantic coherence does not fit with the desire to capture the meaning of the texts. For this reason, we also report the exclusivity of words to topics - the intuition behind this is that exclusive words are likely to give a better indication of what *distinguishes* the topic from other topics. Whilst it is advantageous to check the topics against human judgement, having automatic quantitative metrics is helpful because it allows us to iteratively adjust the topic models without requiring constant input from social workers and / or others familiar with the data (the topic modelling may go through tens of iterations which would unreasonably interrupt local authority staff members' workflow).

Research question 6 attempts to answer the question of how much the model accuracy would improve if we added more data. In general, more data improves the accuracy of the model (up to a point and with some other caveats around current model accuracy and data quality). This is helpful to know because it gives an indication of whether it may be worthwhile for local authorities for whom the accuracy of the model is currently too low to wait until they have more years of data or attempt to overcome the barriers of partnering with other local authorities to increase the sample size contemporaneously.

Research question 7 allows us to test whether using data from before a major change may decrease the accuracy of the model. This may be the case because the model may pick up different signals from the data before and after (although this is likely to depend on the quantity of data already being used). It is helpful to know for local authorities who have gone through recent changes or for local authorities looking to increase the sample size by using data before major changes if this does influence the model accuracy[11].

Questions about utility

- RQ9: Do social workers find the outputs of the model a useful addition to tools and information they already have access to?

It is possible that the models perform well but are not helpful to social workers. This would be the case if the predictions and the variables which contribute most to the predictions are not intelligible to social workers; or if the model is not sufficiently complex to add anything to the social workers' existing evaluation of the level of risk. Because we would like this research to be as practically useful to local authorities as possible, we would like to test some ways of representing the outputs of the models to small groups of social workers and receive qualitative feedback on them.

Outcomes and Data

Outcomes

[9] Mimno D, Wallach HM, Talley E, Leenders M, McCallum A (2011). "Optimizing Semantic Coherence in Topic Models." In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, pp. 262–272. Association for Computational Linguistics, Stroudsburg, PA, USA. ISBN 978-1-937284-11-4.

[10] Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, Gadarian SK, Albertson B, Rand DG (2014). "Structural Topic Models for Open-Ended Survey Responses." American Journal of Political Science, 58(4), 1064–1082.

[11] We imagine that including data in a different case management system is likely to influence how accurate the model is but will not investigate this because of the additional time commitment of cleaning multiple datasets.

The outcomes investigated are specific to each local authority according to what the local authority was specifically interested in focusing on. In general, they refer to transitions in a child's journey through social care (referral, re-referral, steps up, steps down, case closed)[12]:

- Are children / young people being re-referred within 12 months of having their case being designated "no further action" (NFA) or their case closing?
- Are children / young people who have been referred to early help being stepped up to a Child in Need (CIN) or Child Protection (CP) plan within a 12 month period?
- Following assessment are children / young people being stepped up to a Child Protection (CP) plan or being looked after within 18 months?
- Are children / young people being re-registered as CIN within 12 months of their case being designated "no further action" NFA or their case closing?
- Are children / young people registered on a Child in Need (CIN) plan being escalated to a Child Protection (CP) or being looked after within 4 years of initial registration?

Structured Data

Local authorities have different case management systems and different procedures for how to record data. Structured data required in the statistical return to the Department for Education is likely to follow the same definitions and categorisation for demographic variables but this is not necessarily the case for status and pathway variables which may differ by organisation.

Where data is available (and where it wouldn't cause the model to be prescient i.e. have knowledge of events before they take place), we seek to include:

- Current and historical statuses (e.g. CIN plan, CP plan etc) to create the outcome variable where relevant (further lags can be included as independent variables)
- Pathway variables describing what actions were taken within the social work system (e.g. NFA, referral to early help etc) to create the outcome variable where relevant (further lags can be included as independent variables)
- Duration of open case
- Duration of episode
- Month and year of start and end of episode
- Source of referral
- Primary need identified at assessment
- Factors identified at the end of assessment
- Category of child protection registration (where relevant)
- School attendance and exclusions
- Age
- Gender
- Disability
- Length of time of case in the system

[12] For more information on the child's journey through social care, please see the flowcharts in Working together to safeguard children: July 2018. Chapter 1. (https://www.workingtogetheronline.co.uk/chapters/chapter_one.html).

We exclude social worker ID. Although there is some evidence that there is variability in social worker decision-making (even with standardised test case vignettes)[13] and so we expect social worker ID to be a useful feature in prediction, we're 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 avoids the risk of this research being used as a historical performance tool.

Text Data

Where the documents are available (and where it wouldn't cause the model to be prescient i.e. have knowledge of events before they take place), we seek to include:

- Referral/allocation to children's social care (at the front door)
- Assessments (initial and ongoing assessments)
- Initial / Review Child Protection Conference Report (from a multi-agency meeting held to decide whether a child needs a child protection plan)
- Case notes aka observations or case summary records
- Referrals for services

We treat each text field in each document (or each document) separately (depending on the length of the text fields). We include the number of documents per child as a feature and also construct per text field per document where possible:

- Total length of text in the text field;
- "Sentiment" of the text field: this captures markers such as how positive / negative the text is, whether it is associated with a particular feeling (e.g. trust, anger), the intent, and the subjectivity of the text (whether it is more factual, or opinion-based, and how certain the author is);
- "Politeness" of the text field captures markers such as whether the notes give reasons, hedge or qualify their statements,
- "Concreteness" of the text field captures markers such as how abstract the text is e.g. whether the social worker creates psychological distance between themselves and the child / young person and family;
- Counts of social worker elicited risk and protective.

We write "sentiment", "politeness" and "concreteness" in quotation marks because they have specific meanings in the machine learning literature, and we use them in that context rather than their everyday meanings. It is worth noting that "sentiment", "politeness" and "concreteness" packages have been built with other domains in mind - if these features do not add much predictive value, they will be excluded from the final model.

Handling Data

Languages

[13] 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

We conduct the analysis in two commonly used, open-source languages, R (v. 3.6) and Python (v. 3.7). Both languages are free and have supportive communities developing them. Among others, we'll be using the following packages:

- R: stargazer, data.table, tidyverse, mlr, NLP, OpenNLP, stm, lda, topicmodels, iml, lime, aquitas, rerf, ggplot2, xlsx, car, rms, lda, topicmodels, pROC, LDAvis, stmBrowser, wordcloud, caret, ALEPlot, gender, textclean, NLP, magrittr, spacyr, politeness, tidytext, tm, tm.plugin.sentiment
- Python: aequitas, alepython, bs4, codecs, ftfy, fuzzywuzzy, gensim, Imblearn, lime, matplotlib, nltk, numpy, pandas, pickle, pyLDAvis, spaCy, Skater, pyimp, scipy, sklearn, tensorflow, tensorflow-hub, wordcloud, Xgboost

Please note that this is a non-exhaustive list. Since we will publish the code under a copy-left licence, the final list of packages will be available with the code.

Pre-processing

Handling of Duplicates

We will check for duplicate entries. We will work closely with colleagues at each local authority to understand what the duplicates mean.

From initial conversations, it is more likely that referral or assessment reports or case notes will be duplicated across siblings than whole records being duplicated. For cross-validation purposes, we will ensure that siblings are included in the same fold[14].

If there are exact duplicate entries, we will drop all duplicates. If there are duplicate entries for the same individual over the same time period, we will take the entry which is more complete.

Handling of Missing Data

If the outcome variable is missing, we will drop the row. If more than 30% of the values of a structured data column are missing, we will consider dropping the column, and otherwise we will null impute. Null imputation is preferred because missingness in administrative data may be a signal in itself,[15] for example, the social worker is not able to collect information from a non-cooperative family. If referral or assessment reports are missing and their missingness is not due to faulty data extraction from the case management system, we include the text as a document of length 0. Again, this missingness may be indicative in and of itself.

Structured Data cleaning

To validate data quality, we will conduct checks on the following:

1. data-type constraints (e.g. ages should be numbers),
2. range constraints for numeric data (e.g. ages should be between 0 and 18 years);

[14] For clarity, this stops the model from performing well just because it has seen very similar data on the individual's sibling. It does not mean that the sibling's data is being used as an input to the model for an individual.

[15] García-Laencina, P.J., Sancho-Gómez, JL. & Figueiras-Vidal, A.R. Pattern classification with missing data: a review. Neural Comput & Applic (2010) 19: 263.

3. set-membership constraints for categorical data (e.g. limited to list of source of referrals as in Department for Education statistical returns);
4. regular expression / formatting patterns (e.g. dates)
5. cross-field validation (e.g. start dates start before end dates).

We will in particular check for patterns in outcome variables which indicate that the data is being entered in a different way to how we're defining a value e.g. 'case closed' is being used to indicate that the social worker has arranged the next action rather than the case actually being closed.

For distance-based or regularisation algorithms, we will scale the variables (demean and standardise).

Pseudo-anonymisation of Text

Structured data is relatively easy to pseudo-anonymise by excluding instant identifiers (names, dates of birth, addresses etc) and meaningful identifiers (variables that would allow identification if linked with another dataset e.g. a unique pupil number) as data fields.

Text data provides a more significant challenge in isolating and deleting the names from the documents. Social workers are likely to include the name and nicknames of the child, and may include the names of the child's relatives and professionals from other multi-agency partners in assessment records.

We outline the technical methods for pseudo-anonymisation, and process safeguards we shall follow to remove personally identifying information.

We use the term "pseudo-anonymise" instead of "anonymise" because research has demonstrated that it extremely difficult to fully anonymise data[16]. Discussions with local authorities concluded that pseudo-anonymising - where an individual cannot be re-identified except with the addition of extra data - is sufficient given the process safeguards we have in place.

Technical methods:

- We request a list of names and nicknames / alternative names for children / young people, family members and professionals in the dataset (where available).
- We tag the parts of speech in the documents. This identifies words which are likely to be names, addresses, phone numbers and email addresses on the basis of the structure of the sentence so that they can be removed.
- We use public lists of common names to further identify any names to be removed.
- We remove this identifying information found and any words sufficiently similar to take into account spelling mistakes[17].

Process safeguards:

[16] For an early illustration of this, see: Sweeney, Latanya. 2000. Simple Demographics Often Identify People Uniquely. Data Privacy Working Paper 3. Pittsburgh: Carnegie Mellon University.

[17] What counts as "sufficiently similar" is mostly a matter of trial and error.

- No non-anonymised data leaves local authority IT systems thus reducing the need for anonymisation to reduce the risk of identification in an unintentional release of data.
- The researchers conducting the modelling will sign confidentiality agreements to not disclose information about cases.
- The researchers will check topic models and other model outputs of any examples used in the social worker workshops and the report for disclosure of information that would allow the identification of the individual (either through statistical disclosure or the text).

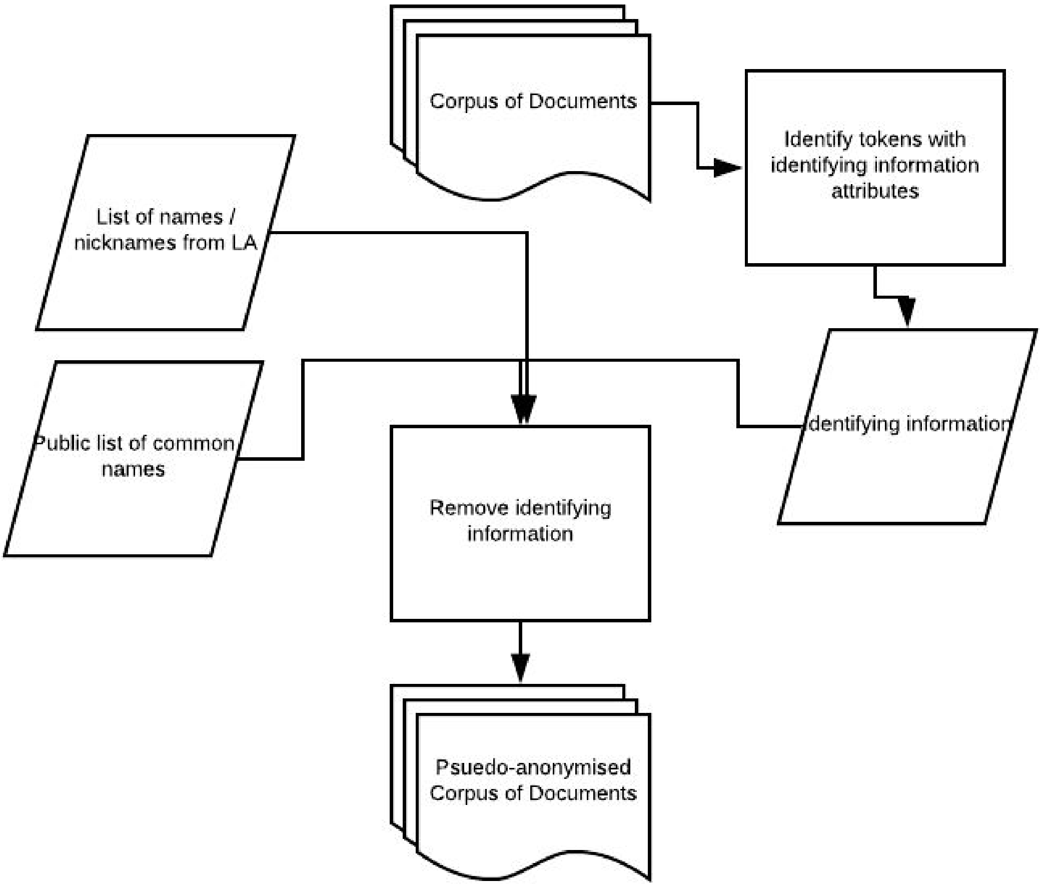
Corpus of Documents
Identify tokens with identifying information attributes
List of names / nicknames from LA
Public list of common names
Identifying information
Remove identifying information
Psuedo-anonymised Corpus of Documents

Figure 2: Pseudo-anonymisation of data technical workflow

Linguistic Markers

We identify linguistic markers, such as politeness, sentiment and concreteness for each text field / document (depending on the length of the text field) prior to pre-processing because the identification of these markers makes use of the ways in which the content is described.

Text Pre-Processing

The predictive models require numerical inputs and so we convert the anonymised documents to vectors of numbers where each number represents a value associated with tokens (words or a small number of consecutive words).

For the tokens to be fed directly into topic models (rather than embeddings), we complete a number of pre-processing steps to eliminate unnecessary complexity:

1. We check for and correct any unrecognised character encoding which may result from data extraction, we remove punctuation and change all the words to lowercase.
2. We remove “stop words” e.g. “the”, “when”: words which commonly occur but do not add much meaning - there are standardised lists of stop words which we shall augment with social work specific common words.
3. We lemmatise all words (remove the conjugation so that we are left with the dictionary root of the word), e.g. walked becomes walk.
4. Finally, there may be some phrases consisting of multiple words which mean something specific when the words are considered together e.g. “domestic abuse” - we detect such phrases and consider them as single tokens.

These steps mean that: 1) we get the simplest version of the word: we would like the algorithm to treat the words the same irrespective of pluralisation or capitalisation as it carries the same meaning, 2) we allow phrases to retain their meaning and 3) we drop words that are likely to have no predictive power.

We do not use these standard preprocessing steps prior to the transfer learning because the process is more capable of handling the additional context. Instead, we check the intersection between the vocabulary of the pre-learnt embeddings and the dictionary built on the text data at hand. We inspect the out of vocabulary tokens and preprocess using the steps outlined above as necessary.

The pre-processing steps are the same for the training and holdout data except that in the holdout data, the bags of words are built to align with the vectors in the training data. This means that tokens which appear in the holdout data but not the training data are dropped, and the tokens which appear in the training data but not the holdout data are assigned a count of 0.

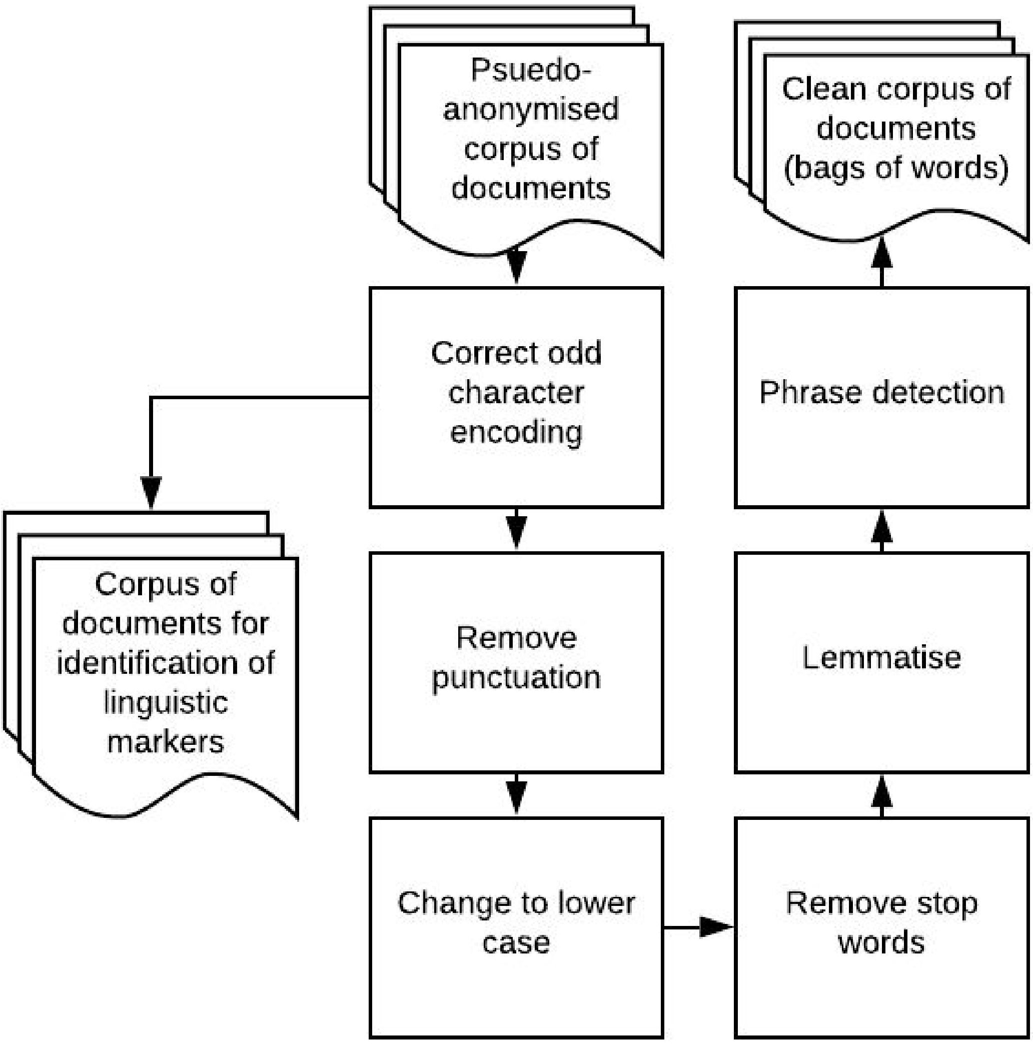
Psuedo-anonymised corpus of documents
Correct odd character encoding
Corpus of documents for identification of linguistic markers
Remove punctuation
Change to lower case
Remove stop words
Lemmatise
Phrase detection
Clean corpus of documents (bags of words)

Figure 3: Text pre-processing steps (not embeddings*)*

Text Processing

The pre-processing steps outlined above leave us with each document being a clean "bag of words" (and multiword phrases).

Transfer Learning

We primarily treat documents as "bags of words" (or tokens). Bags of words are frequently used as the input to topic models. However, the algorithm has no ability to understand the meaning or context of those words, which means that similar words e.g. "harm" and "abuse" are treated distinctly even though a human reading those words would understand the similarity of meaning. We can improve on a bag of words model by replacing the words with "embeddings", vectors of words which co-occur frequently with the word of interest in a separate large corpus. The most famous toy example being the word "queen" being represented by the embedding "king + woman - man". Using pre-trained embeddings allows the algorithm 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 is known as “transfer learning”.

The pre-trained embeddings are likely to be out of the box from Tensorflow (an open source machine learning library) and trained on a standard and easily available corpus. This choice reflects the low likelihood of being able to access a large corpus of documents more similar to social care referrals and assessments given the sensitivity of such documents. It also reflects the desire to use standard tools so that barriers to local authorities using such models is low. At first, it seems counter-intuitive that we can give the model a warm start by using patterns learnt on other text, particularly something as different to an assessment report as Google news; however, there are lots of general patterns the model can learn from the additional dataset which relate to the structure and semantics of the English language before learning the specifics of social work vocabulary.

Since the corpus of a local authority’s documents on children in social care is small, we wish to retain most of the learning of the general patterns from the large corpus. We therefore use the pre-trained embeddings as a feature extractor and feed these features into the topic models. This allows us to keep the patterns about the meanings of words learnt from the large corpus. This works better the more similar the corpuses are - we will monitor the intersection of vocabulary between the pre-trained embeddings and our corpus of documents.

We are likely to try the following embedding frameworks:

- Google’s universal sentence encoder[18] which is trained with a deep averaging network (DAN) encoder.
- Google’s universal sentence encoder lite[19] which is trained with a transformer encoder but is specifically designed for cases when computation resource is limited e.g. on-device inference.

Both are trained and optimised for greater-than-word length text, which is important in so far as we consider the meaning of the documents to come across in sentences and paragraphs and not merely individual words. Both takes in as input variable length English text and produces an output of a 512-dimensional vector. They are trained unsupervised on a variety of web sources (Wikipedia, web news, web question-answer pages and discussion forums) and augmented with training on supervised data from the Stanford Natural Language Inference (SNLI) corpus[20]. Both are designed for when computation resource is limited[21].

Topic Modelling

[18] See Tensorflow Hub, universal sentence encode version 2, https://tfhub.dev/google/universal-sentence-encoder/2 for more details. See also: Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, Ray Kurzweil. Universal Sentence Encoder. arXiv:1803.11175, 2018. https://arxiv.org/pdf/1803.11175.pdf

[19] See Tensorflow Hub, universal sentence encoder lite version 2, https://tfhub.dev/google/universal-sentence-encoder-lite/2 for more details. See also: Daniel Cer et al. Universal Sentence Encoder. arXiv:1803.11175, 2018. https://arxiv.org/pdf/1803.11175.pdf

[20] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In Proceedings of EMNLP.

[21] The DAN encoder has time and space complexity are O(n) in sentence length compared with O(n2) for the transformer encoder.

To both improve the performance of the models and make the inputs to the predictive model more understandable to practitioners, we attempt to draw out the themes or "topics" of the documents. Topic modelling is an automated process of finding words that occur together in the documents. The technique tries to summarise the document by identifying topics rather than lists of words or phrases so they are more comprehensible to humans. We try both topic modelling over all the documents for a child and for each document / text field (the latter depending on the length of each text field).

We consider two types of topic models (Latent Dirichlet Allocation (LDA) and Structured Topic Models (STM))[22] which both imagine that documents are generated by the author picking a set of topics and then for each topic picking a set of words. The topic models then fit the data to find the most likely values for the parameters of the model. STM 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.

Dimensionality Reduction as an alternative to topic modelling

Although topic modelling has the advantage of being somewhat interpretable, the topics may not be very insightful given the constrained topic space overall (children's social care) and within each individual text field (the question asked). We therefore also try an alternative way of reducing dimensionality - latent semantic indexing - over a tf-idf matrix and structured data. We use the words / phrases from the tf-idf matrix to help us label the latent constructs so that some interpretability is maintained.

[22] We initialise STM with spectral initialisation, based on the "stm" package creators recommending it for consistent results (Roberts M, Stewart B, Tingley D, 2016. "Navigating the local modes of big data: The case of topic models." In Computational Social Science: Discovery and Prediction. Cambridge University Press, New York.).

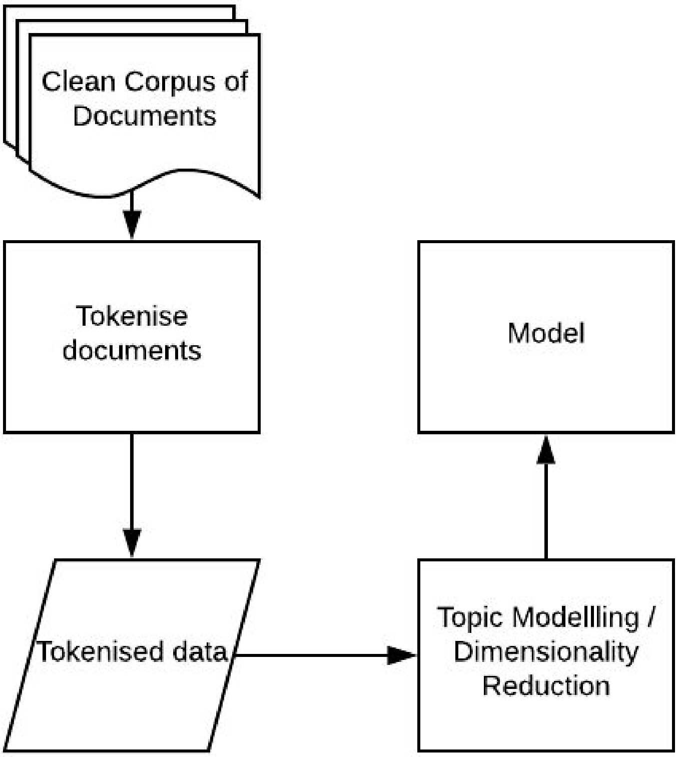
Clean Corpus of Documents
Tokenise documents
Tokenised data
Topic Modellling / Dimensionality Reduction
Model

Figure 4a: Text processing steps (non embeddings)

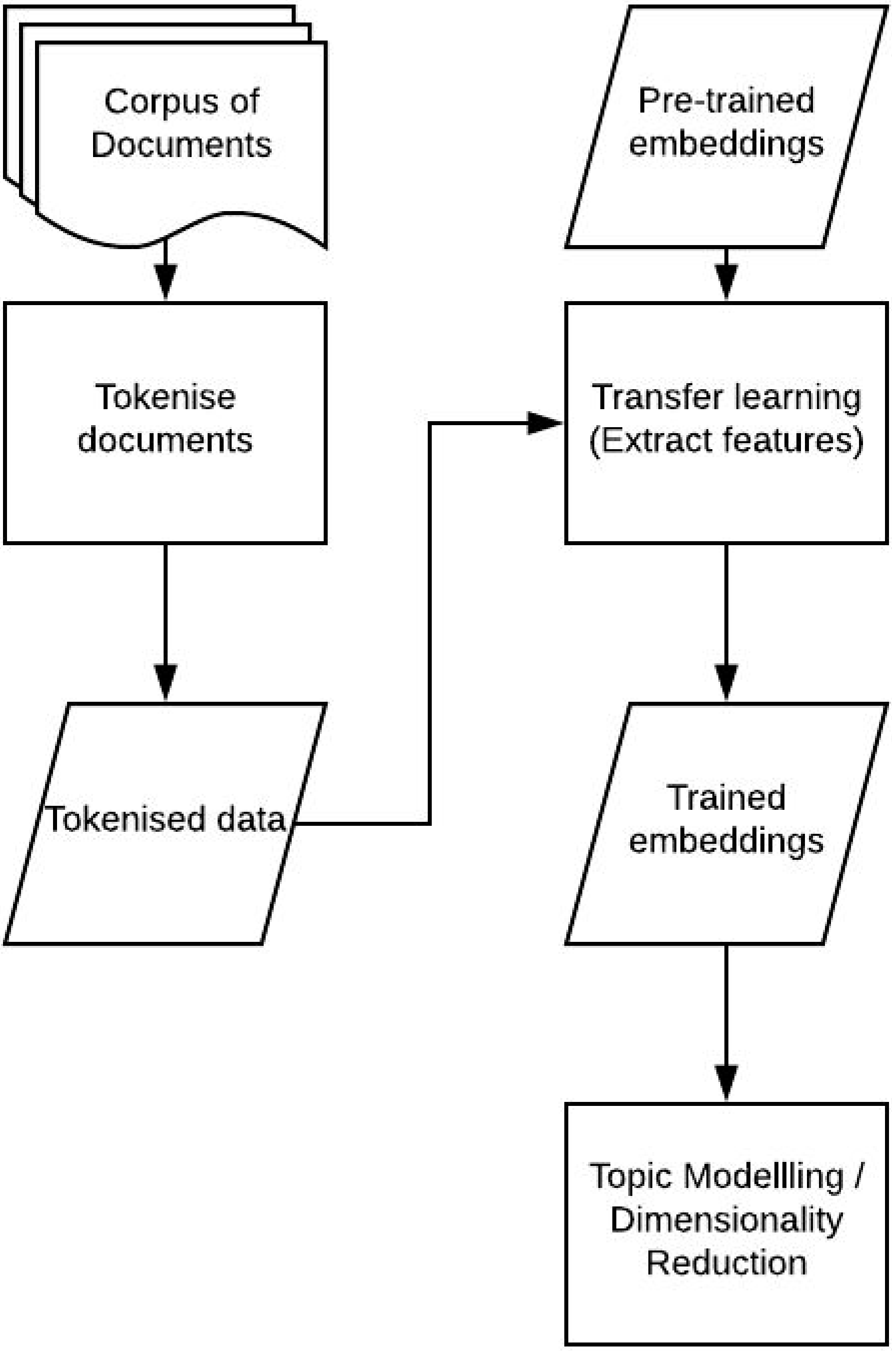

Figure 4b: Text processing steps (embeddings)

Model Specification

Model Types

The prediction questions are all binary classification problems where the positive value is the outcome of interest e.g. re-referral, stepping up. We have chosen a few models to test which have different complexities, are based on different assumptions, and reflect different costs for LAs. Decision trees are highly interpretable but have lower predictive accuracy - we include gradient boosting as a more complex tree-based model likely to have higher predictive accuracy. Unregularised logistic regressions are familiar to many analysts whilst regularised logistic regression provides finer control for the fitting of the model. We exclude K-nearest neighbours given that it is not both transparent and privacy-preserving as investigating why the model made the prediction involves looking at similar cases. We exclude Naive Bayes because there is little room for adjustment or improvement, and SVM because it is difficult to interpret. We exclude neural networks because the training data required for deep learning is more than is available for this project.

For each prediction question, we use (after feeding in the structured data and the topic models)[23]:

1. decision tree
2. logistic regression (un-regularised, regularised);
3. gradient boosting.

Satisfying Metrics

Our optimisation is restricted by the available computer power (on a standard issue machine) and also by the time available: we will have a short period of time at each local authority to check the data extract, anonymise the text data, clean and process the data and build and evaluate the models. Model building is a very iterative process and given the short length of time, we shall need relatively rapid training so that we can tweak our models sufficiently quickly.

Our initial approach will be relatively crude: starting with simple models (with lots of regularisation to restrict the number of features) and building up to greater complexity. We start with a wide but sparse searches of hyperparameter not associated with complexity, evaluated through 3-fold cross-validation. We will take more of a nuanced approach as we progress.

We shall monitor the convergence time and may discontinue the trial of particular models or topic models whose low complexity versions take too long to converge (> 0.5 days). We anticipate that once the model is chosen, the time taken to re-train the model in the future is not likely to be a limiting factor (model degradation is not likely to be drastic over a period of a few days and so training over this period of time would be acceptable). For this reason, models which take longer to train aren't necessarily off the table but we will have had fewer opportunities to optimise them.

[23] Where the models involve random choices, we set seeds so that each model can be fully replicated.

Some time savings can also be achieved through efficient use of hardware e.g. utilising all CPU cores and running no other programmes concurrently. We use stochastic or mini-batch processing where the algorithm allows for it.

Hyperparameter Tuning

For reasons explained above (see “Satisfying Metrics”), we start with simple models to test the pipeline and training time.

We are likely to start with the following initial hyperparameters to create simple models:

- Logistic regression: L1 regularisation with regularisation strength of 1.
- Gradient boosting: 50 trees, maximum depth of 10 and minimum number of samples in a leaf node of 5 and the maximum number of features to consider when looking for the best split being a square root of the features.

We then increase the complexity of the models and tune the hyperparameters by conducting a gridsearch around the hyperparameters where the trade off between bias and variance starts to take effect.

Performance metric for training

The choice of metric to optimise for depends on the kinds of mistakes that one wants to avoid. During training, we maximise the average Area Under the Curve (AUC) of the precision-recall graph in cross-validation and monitor its variance. Precision measures how many are true positives (TP) out of all of those identified as positive. Recall measures how many the algorithm identified as true positives (TP) out of all of the actual positives.

$$Precision = TP/(TP + FP)$$

$$Recall = TP / (TP + FN)$$

One maximises precision by minimising the number of false alarms (false positives, FP) whilst one maximises recall by minimising the number of cases you miss that should actually be a cause for concern (false negatives, FN). There is a tradeoff between the two depending on the threshold of risk: in social worker parlance, by trying to make sure you intervene in every case that needs support, you are likely to also intervene in cases where intervention isn’t necessary.

We chose to measure the area under the curve (AUC) of the precision-recall graph instead of simply monitoring both precision and recall because: AUC measures are helpful in giving a threshold-independent evaluation of the performance of the model, which is important when the output we are interested in is a probability and not a “yes” / “no” prediction. AUC metrics are also scale-invariant so prioritise getting the ranking of cases correct instead of absolute value - again, this is a useful property if one of the ways this model could be useful for providing managers with oversight of cases that are most concerning. We chose the AUC of the precision-recall graph instead of the AUC of the recall-specificity graph (the more commonly used metric) because a) we are likely to have considerable class imbalance which the AUC of the precision / recall graph accounts for; and b) the idea of trading off precision and recall is likely to make more intuition sense to social workers.

Monitoring other metrics during training

Although we need to chose one metric to optimise for during parameter tuning; we also check other metrics on a less regular basis[24] to understand unintended consequences. These other metrics we shall monitor include: the percentage of "risky" cases identified in the top 10% of cases and the reverse for the bottom 10% of cases when the cases are ranked by predicted probabilities. This gives us an idea of how many cases the social worker would have to look through to find the ones the model identifies as at high risk or at very low risk.

Topic modelling performance

The topic models will be fed in to the pipeline having already been optimised to some extent using a combination of grid searching over a hyperparameter space and human judgement. We plot the number of topics against the semantic coherence and use the "elbow criterion" to visually identify the number of topics at which the marginal gain to semantic coherence levels off. We then use a combination of the "stm" and "(py)LDAvis" packages to:

1) Inspect words associated with the topics to understand whether the topics "make sense" i.e. whether they sound internally coherent and distinct.
2) Plot of the topics as circles in the two-dimensional plane whose centers are determined by computing the distance between topics and inspect to understand the prevalence of the topics and how they relate to each other.
3) Where local authorities grant permission to read the raw text, inspect extracts of several documents highly associated with a particular topic to understand whether the allocation of topics to documents is fitting[25].

If there is sufficient time, for the STM model we can also investigate the correlation between the topics and the structured data to understand whether the topics are correlated with values that we expect e.g. a topic discussing themes appropriate to infants correlating highly with an adolescent age group would indicate a poor topic model.

Cross-validation and holdout data

Prior to training we split the dataset of historical cases into training and "holdout" (unseen) data. Holdout data is data in which the outcome has also already happened but is data that the model will not see during the course of training. Once the final model is chosen, the model predicts what happened on this unseen historical data, and we can compare the prediction to the true outcome, which we already know. This is a common workflow in predictive analytics because it allows the model to be tested in a safe environment before having real world consequences.

The datasets are small and so we require a reasonably high percentage of holdout data - we expect this to be 30%. The holdout data shall be stored separately and not consulted until a final model for that prediction question is chosen. We use the training dataset to train and validate through cross-validation. As mentioned above, we build the initial models with 3-fold cross-validation for speed, and then shift to 5-fold cross-validation as we narrow the search

[24] E.g. for the model chosen after each sparse hyperparameter grid search before tweaking the search space or pre-processing or processing steps.

[25] The text data will not be fully anonymised because of the difficulty of removing all names from the documents. If the local authority wishes to inspect each document to check its anonymity before a WWCSC researcher reads the document, we only use this type of evaluation to assess the final choices for models to avoid a too complex workflow.

over hyperparameters. Our workflow consists of checking the bias and variance reported from the cross-validation, and increasing the complexity of the model if the bias is high relative to the variance (i.e. the model hasn't learnt enough from the historical data), and decreasing the complexity of the model if the variance is high to reduce the risk of overfitting to the training data i.e. memorising the nuances of historical cases.

Since we may expect patterns in the data to change over time, we conduct forward-chain cross-validation which only validates the model on data that occurred after the data the model is trained on. If there are duplicates of text across siblings, we shall ensure that siblings are not split between the training folds and the test fold in cross-validation or the holdout dataset (this data leakage would artificially inflate the performance metrics of the model). Where there have been changes to practice or the way data has been recorded (from which the model would learn patterns which are unhelpful for future prediction) and where sample size allows, we monitor the performance of metrics trained on data before the change and validated on data after the change.

Choosing a Model

Machine learning models mostly optimise for some measure of prediction accuracy. In the case of predictions in social work, accuracy is of course also very important but transparency is also important in allowing interrogation of why the model made that prediction. Since the predictions relate to a decision which can be one of the many decisions a social worker makes relating to a child's journey through social care, it is important to consider how the model for a particular decision fits with other models that could be modelling other parts of the journey. We attempt to be transparent by using model-agnostic methods to explain the models; however, we also explore whether it is worthwhile for transparency purposes to choose the same algorithm across all predictions in children's social care. This standardisation may be helpful in communicating clearly the assumptions of that algorithm. This has trade-offs with performance.

Reporting

Model Performance

For each model built for each question, we report:

- The proportion of cases in the positive class i.e. the outcome we are trying to predict (to illustrate the baseline performance that our models should exceed[26])
- The mean and variance of AUC precision / recall from cross-validation on the training data (a high mean indicates that the model is making few mistakes, and a low variance indicates that the performance of the model is less likely to depend on being "lucky" in future data being similar to the data the model is trained on[27])
- The AUC precision / recall on the holdout data (this metric is the ultimate judgement of the models and is the metric against which we compare whether the model is "good enough" for practice).
- The training time and the hardware used.

[26] The simplest prediction you can make is that all future cases will be whichever outcome occurs for the majority of historical cases. For a predictive model to be at all useful, it needs to at least beat this.

[27] Although this should be taken with a pinch of salt given that we are training and validating on this dataset and so the model is likely to learn well how to reduce variance in the dataset

For the model chosen for that prediction at that local authority (the model with the highest AUC precision / recall on holdout data), we also report:

- The number of false negatives and false positives in 100 cases.
- The percentage of "risky" cases identified in the top 10% of cases (and the reverse for the bottom 10% of cases) when the cases are ranked by predicted probabilities.

Unfortunately, it is difficult to estimate what the Bayes optimal error (the irreducible error) for predicting these types of outcomes would be i.e. the best case scenario we can hope for, and hence what a "good" AUC precision / recall looks like. As with almost all predictions that use machine learning, these outcomes are non-deterministic; however, predictive models in children's social care may have particularly high Bayes optimal error rates because they are making predictions that involve the interaction of multiple people and because the lives of families involved in children's social care are frequently complex.

Because of the expected high Bayes optimal error rate, models which exceed AUC precision / recall of 0.65 or being able to identify 75% of cases "at risk" by only inspecting the top 25% of cases that the model considered risky[28]would be deemed a "success". However, we acknowledge that this still represents a considerable error rate and there is a separate question over what is good enough to be used in practice. This distinction is mostly because the costs and benefits of raising a false alarm or missing a case to worry about are different for different stages of the child's journey. For example, the risk of harm to the child is higher when the model and social worker miss a case which should be flagged as a child protection case compared with when they miss a child in need case.

Furthermore, whether the performance of the model is high enough also depends on how the model would be deployed, particularly on the training of frontline staff interpreting the outputs of the models and recourse for correction of errors. Good processes around handling of errors tolerate a higher number of errors.

Testing ways to make the models interpretable

Interpretability is the degree to which a human can understand a prediction made by the model. Machine learning models which are sufficiently complex to learn the nuanced patterns and predict well are often difficult to interpret because there tend to include many features (hundreds to thousands) and include non-linearity and interactions between variables, meaning that changing the value of one feature can affect the outcome through multiple pathways.

Interpretable models allow the algorithm to give reasons for the prediction, which is useful for debugging models in training and more crucially allows the algorithm to be challenged if the prediction goes against expectation. Building models that are interpretable to humans is particularly important in the sector of children's social care because the decisions which the predictive model may contribute to could have a drastic impact on the lives of the child and

[28] 21.9% of cases were re-referred within 12 months in England in 2018 (Department for Education, 26 April 2019, Local authority interactive tool (LAIT)) so a perfect model would identify 100% of cases in the top 21.9%. Once would expect a model with an accuracy of 65% to identify 65% of the cases in the top 21.9% if the errors were uniformly distributed throughout the ranking of cases by probability but the top 21.9% are those that the model has assigned a higher probability of being at risk to and so one would expect a higher accuracy rate for those ranked highest. We thus give a higher percentage accuracy than would be expected when looking at 100% of the cases, and round to looking at 25% of cases.

family, and so it's especially important to understand why the model is making that prediction to the extent that the social worker could explain it to the family, and also to allow opportunity for it to be challenged.

We chose interpretation methods which are algorithm-agnostic to facilitate flexibility in picking the highest performing model without the need to change the interpretation method, and also to facilitate comparisons.

We do not use techniques which illustrate by example from the dataset because of the preference to be privacy-preserving and also for the practical reason that showing a case as an example could be an overload of information. There is a tradeoff between the fidelity of the explanation to the model and the comprehensibility to humans (unless the model is very simple). Following our principle that the models should be useful to social workers, our priority is comprehensibility.

What we lose in this tradeoff is complete attribution which is more appropriate to a court of law. However, the kinds of outcomes we're predicting are not within the family court system but earlier in the child's journey through children's social care so the loss is not felt as keenly.

We are interested in explaining global model behaviour across all cases to gain an insight into general patterns as well as explaining individual predictions, and also explaining the contents of topics. We report the feature importance for the final models chosen. From this, local authorities could implement a simple logistic regression using a top N number of features (where N is a relatively small number). We also report examples of Accumulated Local Effects (ALE) plots for features of interest, and examples of individual predictions and topics. However, what is most interesting about the explanations is how intuitive they are to social workers and so we report on feedback from social workers where it has been possible to conduct a workshop with them.

Global Behaviour

We calculate the importance of each feature, which in the algorithm-agnostic definition is the increase in model error when the feature's information is destroyed. We plot feature importance to give a global explanation for the predictions of the model. Feature importance takes into account all interactions between features so it's possible to see the effect of changing the input value on the output value via all pathways. It is debated whether feature importance should be calculated on training or holdout data - if the model is generalising well, calculating on training or holdout data shouldn't be too different and since we're not optimising for feature importance, using the training data with its larger sample size and its availability throughout training seems sensible. It is worth noting that the importance of a feature has to be interpreted in the context of the model run and the other features included, for example, if two features are highly correlated, removing one feature may increase the importance of the other - in this way it's not to be interpreted as an exact description of the world.

We are also interested in the shape of the relationships between individual features and the predicted outcome. To do so, we use Accumulated Local Effects (ALE) plots which demonstrate how the predictions change in a small “window" of the feature around the value for data instances in that window. ALE plots are unbiased (they still “work” when features are correlated), and the amount of time required to create the ALE plots is typically shorter than partial dependence plots.

Individual Predictions

To explain an individual prediction, we trial using counterfactual explanations, which describe the smallest change to the feature values required to change the prediction. The output presented is the features that would be required to change and not the counterfactual example. Although the example can be synthetic (and thus can maintain privacy), synthetic examples of text are bags of words which aren’t particularly human-friendly, and the example is likely to contain too much information to pick out the contrast with the case of interest.

Counterfactual explanations are interesting to test as an output because they do not require access to the data or the model, only the model’s prediction function, and thus are an option for encouraging transparency amongst companies which do not wish to reveal details of the data or model to protect data privacy and trade secrets.

One possible output of the model is a ranking of cases by highest to lowest risk (instead of providing a predicted probability / probability range of the outcome). This may be interesting from the perspective of supporting a team manager to prioritise which cases to be particularly concerned about. Where possible in a workshop with social workers, we shall investigate whether providing a ranking of cases is intuitive to social workers, and whether model rankings correspond with the social workers’ rankings by calculating Spearman’s rho.

If there is sufficient time, we will also trial explaining individual predictions through local interpretable model-agnostic explanations (LIME), which trains local surrogate models.

Explaining topics

We show word clouds with a small number of the highest weighted words to visually represent the topics. Where topics are based on embeddings we show the individual word instead of the embedding.

Fairness

There are multiple different ways of considering whether a process is fair in everyday life and these translate into different technical definitions, for instance:

1. Equal parity where each group is represented equally in the set of children identified as being at risk. However, this definition does not reflect that there may be risks associated with having a particular characteristic e.g. disabled children may have a higher risk of their needs not being met because they have higher needs than non-disabled children.
2. Proportional parity where each group is represented in the set of children identified as being at risk proportional to their proportion in the dataset. This definition overcomes the challenge of equal parity but is naive about social workers’ historical identification of children at risk being representative of “true” risk and indiscriminatory [29].

[29] This is not to accuse social workers in particular.

To avoid making assumptions about the blindness of social worker decision-making to sensitive characteristics, we instead focus our evaluation of whether a model is fair based on whether the error rates are different by group membership defined by sensitive characteristics.

If the model's error rate is much higher for adolescents, for example, adolescents are more likely to either be mislabelled as high risk which could lead to unnecessary interventions in their lives, or mislabelled as low risk which could lead them to unnecessary harm (were the models to be used in a tool to assist social workers).

The choice of fairness metric also usually depends on whether the intervention to be given post-prediction is punitive or assistive as it changes the kind of error you'd like to minimise.

Given that social work interventions can be considered either punitive or supportive on a case-by-case basis and dependent on perspective we report both the false discovery rate parity (to monitor cases being falsely identified and unfairly "punished") and false omission rate parity (to monitor cases being falsely omitted and unfairly missing out on a benefit they are entitled to).

$$False\ discovery\ rate\ =\ FP\ /\ (FP\ +\ TP)$$

$$False\ omission\ rate\ =\ FN\ /\ (FN\ +\ TN)$$

It is desirable to minimise both those unfairly punished and those unfairly missing out on a benefit they are entitled to. These fairness metrics assume that the output of the model is binary ("at risk" or "not at risk") rather than the more nuanced probability of being at risk. We also report the pinned AUC (area under the curve) which assesses fairness without having to specify a percentage probability at which one considers at child at risk.

Unfortunately we do not have data on all the characteristics we would like to check and so can only check bias for the ones that we do have access to, namely age (depends on categorisation in datasets but likely to follow the Department for Education statistical return categories: under 1 year, 1-4 years, 5-9 years, 10-15 years, 16+ years), gender (likely to follow the Department for Education categorisation - male, female, unknown), disability (recorded or not recorded) and ethnicity (depends on categorisation in datasets but likely to follow the Department for Education statistical return categories: white, mixed, Asian or Asian British, black or black British, other, unknown). We may further aggregate these categories depending on the sample size of each group.

Predictive Inference

For any individual predictions which we give as examples we report the 90% prediction interval which tells you where a value will fall in the future, given enough samples, 90% of the time. For the models overall, we report:

- a histogram of predicted probabilities so it is possible to visually inspect the spread of predictions
- the average width of the 90% prediction intervals

- the prediction interval of the threshold value (the value at which the prediction changes from “not at risk” to “at risk”)

Predictions with overlapping prediction intervals are not statistically different from each other. Very wide prediction intervals indicate that the model is not very useful. The average prediction interval indicates whether it is worthwhile displaying a lower precision version of the output e.g. within the decile.

Risks

We identify risks to the project and outline our mitigation strategies:

Risk	Mitigation
Lengthy project set-up phase reduces the time available for research: the project involves buy-in and contribution from many different teams within a local authority (senior leaders, information governance, IT, frontline social workers) which takes some time to coordinate.	We set aside a substantial amount of time to allow for project set-up.
Data extraction and cleaning takes up a large proportion of the time set aside for research, reducing the time available for the model building at the local authority.	To reduce the amount of time familiarising ourselves with the data, we confirm the data fields to extract with the local authority before our arrival onsite, and check our understanding of the data definitions. To open up time for lengthy data cleaning, we also write the code for model building (which can be pretty standardised) on dummy data prior to arriving on-site at any of the local authorities. This allows us to a) optimise the time spent at the local authority by doing tasks which can only be done onsite, b) test how long the implementations of the algorithms take on similar data prior to arrival so that we can better plan our time at the local authority.
Local authority machines may not be sufficiently powerful for machine learning algorithms to train in sufficient time for us to iteratively develop the model.	To mitigate this, we have chosen fast implementations of the algorithms, we shall test the code on dummy data to get an insight into the training time, and also implement the code on simple models as a first step.
There is a risk that the way in which we implement the project is not in line with the	We instead draw inspiration from existing frameworks for building ethical AI (e.g.

technical strategies that the ethics review recommends. We were interested in doing the technical feasibility pilots and the ethics review concurrently so that we could publish the final report more quickly given that the question of whether to invest in such tools is very much a current question for local authorities. However, this does mean that we do not have the conclusions from the ethics review when planning the technical feasibility pilots.	frameworks by DCMS[30], The Institute for Ethical AI & Machine Learning[31], FAT ML[32]) and make pragmatic choices to adapt them to the children's social care context.

Timeline

Timelines differ slightly for each local authority but we demonstrate a generic version of the timeline (days are per local authority):

Task	Responsible Party	Timeline
Agree data fields to request	WWCSC, local authority	Before agreed date for data processing
Confirm data protection arrangements	local authority	Before agreed date for data processing
Confirm logistics (access to buildings and computers, download software, supervisor)	WWCSC, local authority	Before agreed date for data processing
Extract data	local authority	Before agreed date for data processing
Anonymise data	WWCSC	3 days
Check understanding of data and data cleaning	WWCSC	5 days
Analysis	WWCSC	8 days
Workshop with social workers	WWCSC, local authority	0.5 days
Statistical disclosure checks for Model Outputs (i.e. checking no names are	WWCSC	0.5 days

[30] Department for Digital, Culture, Media and Sport, 2018, 30th August. Data Ethics Framework. https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework

[31] The Institute for Ethical AI & Machine Learning, The Responsible Machine Learning Principles. https://ethical.institute/principles.html

[32] Fairness, Accountability, and Transparency in Machine Learning. Principles for Accountable Algorithms and a Social Impact Statement for Algorithms. https://www.fatml.org/resources/principles-for-accountable-algorithms

included and we comply with rules around small numbers)		
Check Model Outputs with information governance team	WWCSC, local authority	0.5 days
Provide Model Outputs (including 2 page document summarising the results and commented code if desired)	WWCSC	2 days

Ethics

As mentioned in the “Motivation” section, we commissioned a separate independent review of the ethics of using predictive models in children’s social care. Although the review assesses the application of models (work going a step further than we are doing in this research) in addition to the development of the models, the technical work is sufficiently close that we wish to not “mark our own homework”. We also did not have ethics expertise in-house, and are fortunate to have the Rees Centre (part of the Department of Education at the University of Oxford) and The Alan Turing Institute partnering to conduct the review.

Ethics of the Technical Feasibility Pilots

WWCSC is currently setting up an ethics panel. Meanwhile, we sought the review of an independent ethicist who is familiar with machine learning and ethics.

We outline how we’ve tried to mitigate against risk of harm, and maximise the benefits of the research through the lens of the ESRC’s core principles for ethical research[33]:

1. Research should aim to maximise benefit for individuals and society and minimise risk and harm

Maximising benefit

Although the development of tools is outside the scope of this project, the benefit from this research predominantly derives from informing whether tools based on predictive models should be used in practice, and if so for which situations and how to do so in a way which is transparent and fair. The potential benefit from such tools are assisting social workers and team managers in their evaluation of risk for children and families, and the downstream effects on improved safety and wellbeing for children and families as well as the better allocation of resources to those most vulnerable. Should we find that these models are technically feasible, ethical and acceptable, by conducting this research we hope to set the norms for developing these models in ways which are transparent and fair, and empowering local authorities and citizens to be able to challenge the companies developing these tools to do so in an ethical way.

[33] Economic and Social Research Council. Our principles: research ethics committees. https://esrc.ukri.org/funding/guidance-for-applicants/research-ethics/our-principles-research-ethics-committees/

Should we find that these tools are not appropriate for use in children's social care, we hope that local authorities would not invest in the development of such tools unless they have good reasons to believe that their situation deviates from the assumptions we make in our recommendation. In this case where risks are considered to outweigh the benefits of such tools, the benefit of this research would be the prevention of unnecessary expenditure of public funds and the prevention of the harms identified.

We attempt to maximise the benefit by publicly publishing this trial protocol and the final report so that all local authorities can learn from the research and allow it to feed into their decision-making processes on commissioning such tools.

We also will publish the code used to build the models under a copyleft licence (offering people the right to freely distribute copies and modified versions of the code with the stipulation that the same rights be preserved in derivative works created later). We do so in order to reduce the duplication of technical work, and provide absolute clarity on how the principles we are abiding by translating into the practicalities of the code. It also allows for the code to be scrutinised for error and developed further.

Minimising risk and harm

Similarly to the benefits, the vast majority of risks associated with this project relate to the potential downstream impacts of tools based on predictive models. As already mentioned, the building of such tools is out of scope but we attempt to minimise the risk of the development of tools that are likely to cause harm by assessing the technical feasibility of *transparent and fair* models, not just models that maximise predictive accuracy. The independent review of ethics will cover a more extensive exploration of minimising the risks of using tools based on predictive models, which we anticipate to include a combination of technical and process considerations. The technical ones we have considered include:

- reporting the uncertainty associated with the prediction and focusing on performance on ranking rather than absolute prediction values to reduce an overestimation of the certainty provided by the models;
- testing the clarity of explanations for why the model is making that prediction to allow for easier detection of errors by the model and understanding of where disagreements between the social workers' evaluation of risk and the model's evaluation of risk are coming from;
- using embeddings and topics instead of words to reduce the risk of families or social workers gaming the algorithm by mentioning particular words in conversations and reports (although we note that there is a tradeoff between transparency and the opportunity for gaming).

The ethics review may indeed conclude that different or further technical strategies are necessary to minimise the risks identified, or that there are considerations we have missed. We have had to be pragmatic about developing strategies whilst the ethics review is in early stages. We shall integrate the two lines of enquiry in a summary report.

With regard to risks and harms associated with the research itself, we identify the following risks and mitigations:

Risk	Mitigation
Individual children and / or families are identified from their data, and experience potential downstream negative effects.	We mitigate this risk of identification by the researchers during the development of the models through technical and non-technical procedures to pseudo-anonymise the data (see “Pseudo-anonymisation of Text”). Additionally, the researchers signing agreements which protect the confidentiality of the children and families reduce the risk of harm from identification. We mitigate the risk of identification through the reports by conducting statistical disclosure checks to ensure that individuals cannot be identified from aggregate data published in the report.
The models identify that high risk is associated with particular group membership e.g. by ethnicity and individuals who are members of that group experience unfair discrimination.	This is lower risk than in traditional statistical research whose output is generalised patterns *all else being equal* rather than a prediction for an individual (which can better incorporate nuances of counterbalancing protective factors). However, we shall be mindful of making statements about general patterns associated with sensitive characteristics.
Social workers whose decisions in the historical unseen data differ from the model’s predictions experience additional scrutiny and / or other negative impacts on their job.	It is worth noting that a decision taken that was different from the model’s prediction indicates that the model did not predict that decision given the available information and the patterns it has learnt from other cases. The difference may result from a variety of reasons, for example, there is additional pertinent information that hasn’t been recorded in the case management system, or it is an unusual case where there have been few similar observations historically. The model’s prediction is not ground truth - it is built on historical decisions made by social workers. Having said that, there is a possibility that the local authorities could misinterpret any difference between a social workers’ decision and the model’s prediction as the social worker “getting it wrong”. The outputs agreed with the local authority for this project does not include a breakdown by social worker of each individual prediction which differs from the social worker’s decision. How tools built on predictive models are used is out of scope for this project but are an important question for answer should the predictive models be technically feasible.
An external party (company / local authority) builds a tool with the published code and deploys it in a way we would not recommend.	Any party capable of building a deployable version of the models from our code would be capable of building a model themselves: publishing the code simply makes it easier for them to build it but equally makes it easier to build it in a way we would recommend. Publishing the

	code under a copyleft licence stipulates that if a party distributed and modified the code, the same rights to free distribution and modification need to be preserved in derivative work, facilitating transparency. We hope that this use of the copyleft licence and publishing our recommendations would enable such a party to be held to account more easily.

2. the rights and dignity of individuals and groups should be respected

Respect for your private and family life, home and correspondence

Article 8 of the *Human Rights Act 1998* protects the individual's right to respect for their privacy without interference from the state. It is a qualified right which means that local authorities can intervene to protect other people's rights, a qualification which is highly relevant in the safeguarding of children and young people. Children's social care has a duty under Section 47 of the *Children Act 1989* to "make enquiries" where there is "reasonable cause to suspect that a child is suffering, or is likely to suffer, significant harm".

No further data is collected for the purpose of this research. The data used for the research is administrative data collected in the course of social workers carrying out their day-to-day duties, and has either been obtained with the consent of the family and / or it has already been considered that the threshold for gathering data under section 47 enquiries has been met. For this reason, we consider that this processing of the data poses no further risk to the respect of the privacy of the individual.

Protection from unlawful discrimination

The Public Sector Equality Duty outlined in the *Equality Act 2010* requires that public bodies eliminate unlawful discrimination. Models learn the "correct" answer by learning from historical decisions, which unfortunately can be discriminatory on the basis of protected characteristics. Please see the section on "Fairness" to learn more about our handling of suspected bias in historical data.

3. wherever possible, participation should be voluntary and appropriately informed

We have not sought the consent of individuals whose data we are processing for the research. As mentioned above, the data used is data solely collected in the course of social workers carrying out their day-to-day duties. It has either been obtained with the consent of the family and / or it has already been considered that the threshold for gathering data under section 47 enquiries has been met.

4. research should be conducted with integrity and transparency

We begin the section outlining our approach by setting out the principles which guide the research ("Building responsible models"), and have outlined elsewhere our strategies to conduct the research with integrity and transparency (see sections on "Fairness", "Testing ways to make the models transparent").

5. lines of responsibility and accountability should be clearly defined

Within What Works for Children’s Social Care, Michael Sanders (Executive Director) is the principal investigator of the research, and the technical feasibility pilots will be conducted by Vicky Clayton (Data Science Manager) with support from the research team as necessary.

There is a principal contact within each local authority who is coordinating the research internally.

The lines of responsibility and accountability are outlining in a partnership agreement between What Works for Children's Social Care and each local authority. Broadly the roles are split as follows:

What Works for Children's Social Care will:

- Work with the local authority to agree research questions and data requests.
- Conduct the techniques and process safeguards to anonymise the data to the satisfaction of the local authority, or work with their staff to do so.
- Create predictive models to answer agreed outcomes of interest.
- Seek feedback on the topics produced by the model from social workers.
- Provide the local authority with the model outputs in the form of a short report including, for example:
    - A description of how the models work, both overall and for representative individuals.
    - Measures of accuracy for each model - if there is sufficient data, this will be broken down by age, gender, disability and ethnicity to make sure that the models are not biased against minority groups.
    - Visualisations of 5-10 important factors.
    - 1-3 anonymised examples of predictions and topics from referral and assessment reports.
    - The commented code for anonymisation, data cleaning and model building.
- Publish the project outputs to the public domain:
    - A research protocol detailing the methodology and a non-technical summary of the methodology.
    - A final research report and blog detailing the accuracy measures for each model, which features are most important, visualisations for importance features, and synthetic examples of topics.
    - A generic version of the model building code under a copyleft licence (offering people the right to freely distribute copies and modified versions of the code with the stipulation that the same rights be preserved in derivative works created later).

The local authority will:

- Work with WWCSC researchers to agree research questions and data request.
- Conduct necessary internal processes to facilitate the research (e.g. a Data Privacy Impact Assessment).

- Extract the data from the case management system or storage repository.
- Provide WWCSC researchers with access to relevant data on local authority systems.
- Provide WWCSC researchers with supervision and use of computers with relevant software for the duration of the Project, and facilitate the bringing in of pre-written code.
- Facilitate access to a small number of social workers (3-5 individuals) to feedback on the topics produced by the models.

6. independence of research should be maintained and where conflicts of interest cannot be avoided they should be made explicit.

As a research organisation, we do not have a profit incentive related to developing a future tool. This allows us to approach the question with a healthy dose of scepticism and take an objective stance on whether the use of such models is helpful to the sector.

We chose to commission an independent review of the ethics of using predictive models in children's social care. Although the review assesses the application of models (work going a step further than we are doing in this research), the technical work is sufficiently close that we wish to not "mark our own homework".

Dan Gibbons (previously Data Science Manager) was seconded from the Behavioural Insights Team to contribute to planning the research. They are part of the team conducting ongoing projects using predictive models with two local authorities' children's social care teams. They were recruited specifically because of this expertise and prior experience. Since the code is being released under a copyleft licence, there is no benefit to the Behavioural Insights Team over and above another organisation with an interest in this field.

Data Protection

The data used for the research is administrative data collected in the course of social workers carrying out their day-to-day duties. No further data is collected for the purpose of this research. Although we will pseudo-anonymise the data by removing instant identifiers, the data would still be considered personal data under GDPR. This section is structured according to the guidance given by the Information Commissioner's Office, which "covers the General Data Protection Regulation (GDPR) as it applies in the UK, tailored by the Data Protection Act 2018"[34].

The data controller for this project is the relevant local authority.

Principles of the GDPR[35]

Principle (a): Lawfulness, fairness and transparency

1. Lawfulness:

[34] Information Commissioner's Office. Guide to the General Data Protection Regulation (GDPR). https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/

[35] Information Commissioner's Office. The principles. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/

The lawful basis for this processing is for local authorities to decide, consistent with their role as data controllers.

The models incorporate GDPR-defined “special category data”[36][37] on disability because it’s an important factor in whether the child’s needs are being met. A lawful basis for processing this special category data also needs to be decided by the local authorities from the conditions outlined under Article 9(2) of the GDPR (which covers the processing of the special category data).

2. Fairness:
ICO’s guidance says fairness means “you should only handle personal data in ways that people would reasonably expect and not use it in ways that have unjustified adverse effects on them”[38]. We consider that families expect social workers to use their information to carry out their statutory duties and additional compatible purposes. We put in place technical safeguards to reduce the risk of adverse effects (see “Ethics of the Technical Feasibility Pilots”).

3. Transparency:
As data controllers, local authorities are conducting their own arrangements.

Principle b): Purpose Limitation

Local authorities’ data will only be used to answer their own questions and contribute to answering the overall research questions for the project (outlined under “Research Questions”). They will not be used for any other purpose.

Principle c): Data Minimisation

We have only requested data that is adequate, relevant and limited to what is necessary for each predictive problem, and to fulfil the purpose of this project overall. Broadly speaking, we can classify the data requested into four groups:

- Outcome measures (i.e. what we are trying to predict)
- Structured data from case management systems as inputs to the model
- Text data from case management systems as inputs to the model
- Sensitive characteristics to test for bias (please see “Fairness”)

Please see “Outcomes and Data” for more information on the specifics.

Principle d): Accuracy

The data will undergo considerable checks to validate the accuracy of the data (see “Structured data cleaning”). These checks identify impossible values (e.g. ages outside of reasonable human lifespan) but cannot identify mistakes which are possible values (e.g. a child being 5 years old instead of 8 years old). Individuals can follow the usual procedures outlined in the local authorities’ privacy notices to rectify their data if they suspect it’s incorrect. The local authority will assist WWCSC research staff to correct the data in the data extract.

[36] Information Commissioner’s Office. Special category data. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/special-category-data/

[37] We are requesting data on ethnicity, which is included in special category data; however, this is for the sole purpose of checking the model for bias and will not be used in the predictive models.

[38] Information Commissioner’s Office. Principle (a): Lawfulness, fairness and transparency. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/lawfulness-fairness-and-transparency/

Principle e): Storage limitation

WWCSC will not store the data at all and we shall access the data on local authority systems. WWCSC researchers will only access to the data for the period of time necessary to do the research and appropriate quality assurance.

Principle f): Integrity and confidentiality (Security)

Please see the security section below.

Principle g): Accountability principle

Please see principle 5) "lines of responsibility and accountability should be clearly defined" of "Ethics of the Technical Feasibility Pilots".

Individuals' rights under the GDPR

We follow the processes and procedures laid out by the relevant local authority. Privacy notices on the local authorities' websites detail how they meet the rights under GDPR and should be the first port of call for each individual local authority. However, we outline here arrangements for upholding the individuals' rights that are specific to this project.

The right to be informed

As data controllers, local authorities are conducting their own arrangements. For some, this type of processing is already covered in their generic privacy notice in facilitating them to fulfil their statutory duty to safeguard and promote the welfare of children, and also to plan the provision of services.

The right to rectification, the right of erasure, the right to restrict processing and the right to object

The local authorities will ensure that the data in the data extract it gives WWCSC researchers access to for the purposes of the project has been rectified if necessary and contains no cases where legitimate requests have previously been made to erase the data, or restrict or object to the processing of the data. If the local authority receives requests relevant to the data during the processing, it will assist WWCSC to rectify or erase the data from the extract as appropriate.

The right to access

Individuals can follow the normal procedures outlined in the local authority's privacy notice to request access to their data as the data extract is simply a subset of that data.

The right to data portability

The right to data portability allows individuals to obtain and reuse their individual data for their own purposes across different services. This is not particularly relevant in the context of children's social care as children and families can't choose to switch to alternative provider in a similar way to changing provider of a consumer service.

Individual's rights in relation to automated decision-making and profiling

Article 22(1) of GDPR states that "The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her." For absolute clarity, we do not consider that this project constitutes "a decision based solely on automated processing" because no decision about the individual is taken. We grant that a main aim of the research is to help local authorities assess whether to invest in developing these types of models and associated tools and that these tools would be used to assist social workers to make decisions which have significant impacts on individuals. However, a crucial distinction is that we do not anticipate ever recommending that the decision be based *solely* on automated processing but instead such tools should at most give the social worker who is making the decision information about the case.

Accountability and Governance under the GDPR

The Data Protection Officer / information governance team of the relevant local authority is accountable for the data protection element of the project. Researchers from WWCSC are responsible for making sure they comply with the local authorities' data protection policies and procedures.

Security

The text data will be accessed only on password-protected local authority computers on local authority systems. WWCSC staff will be "supervised" by a local authority staff member whilst onsite, and what they import and export from the system will be monitored (no transfers of data allowed; transfers of code and log files allowed but checked for disclosure of personal data).

Checks on staff

The data will only be accessed by WWCSC research team members who are ONS Approved Researchers, with DBS checks. Research staff at WWCSC have undergone data protection training and have substantial experience in handling data. The research team continues to review the training needs of the team to ensure WWCSC's approach remains up-to-date.

Appendix

Glossary

Predictive Analytics[39]

Term	Abbreviation	Definition
Predictive analytics	PA	Predictive analytics is a subset of machine learning. It is the scientific study of algorithms and statistical models that computer systems use to predict an outcome for an individual (in contrast to demand forecasting at a local authority level).
Machine learning	ML	Machine learning is a subset of artificial intelligence. It is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference from historical data instead. It is in contrast to 'expert systems' (another type of artificial intelligence that has fallen out of favour) which relies on programmers writing explicit rules e.g. if re-referred, do an assessment.

[39] For a more comprehensive glossary, please see Google's Machine Learning glossary: https://developers.google.com/machine-learning/glossary/

Artificial intelligence	AI	The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages. Currently mostly domain-specific (known as 'narrow AI') e.g. playing chess rather than the kind of intelligence that humans have which they can flexibly apply to lots of different types of problems (known as 'artificial general intelligence').
Feature / variable		We use feature and variable interchangeably. A feature or variable is an individual measurable characteristic of the phenomenon being observed, for example, age, gender.
Prediction bias		A value indicating how far apart the average of predictions is from the average of labels (in this case the social workers' decisions) in the historical data. All models predicting a problem complicated enough to require machine learning have some prediction bias - we do not try to eliminate all bias when training the model because the patterns learnt are less likely to generalise well to future data.
Variance		Variance is an error from sensitivity to small fluctuations in the training set. High variance indicates that (in the attempt to reduce bias) the model has fitted the random noise in the training data, rather than the intended outputs (overfitting).
Bias (fairness)		Risk assessment system can have two types of biases: ● Biased actions or interventions that are not allocated in a way that's representative of the population. ● Biased outcomes through actions or interventions that are a result of your system being wrong about certain groups of people

Children's Social Care

Term	**Abbreviation**	**Definition**
Child Criminal Exploitation	CCE	Including 'county lines' and gangs, when children are used for criminal activity and feel compelled to conform by more powerful people in the gang/chain/peer group.
Children in Need	CIN	Local authorities have a duty to assess and support any children in their area who are 'in need' under Section 17 of the Children Act 1989. Working with families under Sec 17 is by consent only. If necessary, the child can become the subject of a Child in Need Plan to ensure their needs are met.
Child Protection Conference (Initial/Review)	CPC or ICPC/RCPC	Multi-agency meeting held to decide whether a child needs a child protection plan and under what category (emotional, physical, sexual abuse or neglect).
Child Protection Investigation	Sec 47	An investigation to ascertain whether a child is at risk of or has suffered significant harm. This is done under Sec 47 of the Children Act 1989 and this legal term is often used as shorthand. Initiating a Section 47 gives powers to share information without parental consent and compels other

		agencies to share information relevant to safeguarding children.
Child Sexual Exploitation	CSE	A type of Child Sexual Abuse where child and young people are forced or coerced into sexual activity (online or in person), sometimes without realising.
Contacts		Any contact made with children’s social service.
Early Help	EH	Pre-social care, services designed to meet the needs of families where there are lower level support needs (not child protection) to prevent them entering the social care system.
Front door	FD	The first point of contact for referrals into children's social care. Usually there are a team of social workers and manager screening referrals to see which require further action and in what priority order.
Looked After Child	LAC or CiC	When a child is placed somewhere other than with their legal guardian by the local authority. Typically this means in foster care but also includes kinship, respite and residential care.
Looked After Child Review	LACR	Arrangements for the care of looked after children must be regularly reviewed to ensure it is meeting their needs by holding a LAC Review Meetings. They are multi-agency.
Local Authority	LA	There are approximately 150 local authorities in England with social care teams, including county councils, unitary authorities, metropolitan boroughs and London boroughs.
Practice Model		These are different approaches to doing social work; the notion of having a practice model became popular post 2010. They can either be off-the-shelf models like Signs of Safety or something that a local authority designs for themselves, drawing on theoretical ideas such as systemic practice, restorative practice, relationship-based practice etc.
Section 47	Sec 47	Same as a Child Protection Investigation.