Article

Ethical considerations in research when building predictive risk modelling in child and family welfare

by

Anne Marie Villumsen
Senior Researcher
VIVE – The Dannish Centre for Social Sciences Research
Denmark
E-mail: amav@vive.dk

Michael Rosholm
Professor
Department of Economics and Business Economics, Aarhus University
Denmark
E-mail: rom@econ.au.dk

Simon Tranberg Bodilsen
Postdoc
Department of Economics and Business Economics, Aarhus University
Denmark
E-mail: sibo@econ.au.dk

Sanne Dalgaard Toft
Project Manager
Department of Economics and Business Economics, Aarhus University
Denmark
E-mail: sanne@econ.au.dk

Line Svolgaard Berg
Associate Professor
VIA University College
Denmark
E-mail: libe@via.dk

Liesanth Yde Nirmalarajan
Research Assistant
Department of Sociology and Social Work, Aalborg University
Denmark
E-mail: lin@socsci.aau.dk

Keywords:
ethics, child and family welfare, child protection, predictive risk modelling, machine learning, decision-making and notifications

DOI: https://doi.org/10.31265/jcsw.v19i1.619

Abstract

This article presents and discusses ethical issues and implications in research when building a pre-dictive risk model for potential use in Danish child and family welfare. The idea is to build a pre-dictive risk model in order to study whether such a model can be valuable to child and family wel-fare services in the assessment of risk – aimed specifically at the decision-making process regard-ing notifications.

Based on a framework developed especially for this field, we present and discuss ethical consider-ations, reflections and actions in relation to four main ethical principles: non-maleficence, auton-omy, justice and explicability. We hope that our reflections on these ethical challenges can inspire research – and potentially also the field of practice when taking a deep dive into the difficult field of digitalization in social work.

Keywords: ethics, child and family welfare, child protection, predictive risk modelling, machine learning, decision-making and notifications

Introduction

Digitalization, and more specifically Predictive Risk Modelling (PRM) based on machine-learning, is emerging in the very sensitive field of social work with children and families at risk (see e.g. Gillingham, 2016; Chouldechova et al., 2018; Cuccaro-Alamin et al., 2017; Taylor, 2020; Vaithianathan et al., 2013). The overarching ethical dilemma regarding the use of PRM is the duality of potential risks and benefits. First, the many potential risks associated with the use of PRM include the risk of perpetuating systemic biases and discrimination, in which the use of a poorly built model has the power to disproportionally impact individuals from marginalized groups based on, for example, their gender, race or socioeconomic status (Ada Lovelace Institute, 2022). There are also great risks in using poor-quality outcomes, and of reducing critical human and relational factors in decision-making practices (Devlieghere, Gillingham, & Roose, 2022). Second, the potential benefits are also highlighted in the literature. These potentials are primarily imaginary in the sense that they have not yet been thoroughly tested in the field of social work with children and families in adversity (see e.g. Lehtiniemi, 2024; Kawakamiet al., 2022; Cheng et al., 2022). PRM potentially offers an opportunity to exploit data to secure the rights and improve the safety and wellbeing of children and families, for example, by exploiting historical/statistical evidence to ensure that the right children receive care by improving the efficiency and effectiveness of services provided to children and their families, and by providing an empirical foundation for systematic case-based judgement (thus increasing the fairness of the process by reducing pre-existing caseworker biases) (Leslie et al., 2020). In line with this, Coulthard et al. (2020) argue that in relation to child protection decisions, big data approaches are an improvement ethics-wise. PRM may help speed up decisions, reduce errors and increase accuracy and consistency in social workers’ response to similar risk situations, thereby potentially reducing social worker bias and between-group inequalities (Coulthard et al., 2020; Taylor, 2020; Søbjerg et al., 2020). These are all potential risks and benefits which have not been researched in detail in relation to child and family welfare (see e.g. Lehtiniemi, 2024; Kawakamiet al., 2022; Cheng et al., 2022). Hence, we do not know which of the potential risks or benefits could actually be realized if PRMs are implemented, which makes it difficult to weigh the advantages and disadvantages against each other.

Many of these risks and potentials are primarily connected to *the use of PRM* in child and family welfare. However, an important step that comes **before** these very interesting and valid discussions is the actual construction of the PRM model. Some ethical considerations are similar regarding both use and model-building, although there are differences. For instance, a theme like beneficence, in the sense that the PRM should prove beneficial to humanity, is primarily an ethical issue when using a PRM, and therefore primarily an important matter when testing a PRM. Nevertheless, the model-building step is a step that is not always transparent and therefore difficult to discuss ethically - probably because much of the model building in many areas is done by tech companies often lacking transparency for the outside world, and thereby difficulties in raising relevant questions about the ethical considerations and creating accountability when it comes to the constructions of these models (Munn, 2022). Research is therefore needed to outline complex ethical issues regarding the type of algorithm, data quality and the choices made in the construction of a PRM (Keddell, 2023).

The aim of this article is to present and discuss ethical issues and considerations in relation to the process of building a PRM for potential use in Danish child and family welfare decision-making - using our own research as a case. We focus not only on *what* ethical considerations might be relevant, but also on *how* we have applied these considerations to the existing case (Morley et al., 2020). Our hope is that our ethical considerations and actions can inspire other researchers within the field to address and reflect upon ethics when conducting research within this field.

We are aware that separating the build of a PRM from the use of a PRM is a somewhat arbitrary distinction, but because this is a complex and sensitive area (children and families at risk in a statutory context), we will try to separate the two in order to create a learning opportunity. However, during our discussions there will be considerations in relations to the test or use of the PRM when these are relevant to the model building. Nevertheless, our focus is on model building. Our primary argument is that there is great need for two things: 1) Both more qualitative and quantitative research when trying to develop (and test) PRM in social work with

children and families at risk, and 2) a close collaboration between model builders, social work practice and children and families at risk when constructing, testing and potentially implementing a PRM so that any potential harm is solved.
First, we present the Danish context, the research case and the methodological approach to building the PRM. Next, we present and discuss ethical considerations related to building the model.

The context: Danish Child and Family Welfare Services

In Denmark, the social services system is labelled child and family welfare. It is a universal and child-centred system that targets preservation of the family (Pösö & Hestbæk, 2013). The goal of the system is to prevent harm and determine risk in the life of the child - leading to the threshold for intervention being low compared to a child protection system (Gilbert & Skivenes, 2011; Gilbert, 2007; Kriz & Skivenes, 2013). Most cases begin with a referral, typically sent by either the parents themselves or the professionals surrounding the child such as a schoolteacher. When receiving a referral, the municipality has to investigate within 24 hours, and make an assessment on acute danger for the child in question and act immediately. If the child is not in acute danger, the municipality is obliged to make a broader risk assessment (§136 Act of the Child, 2024; Søbjerg, Nirmalajaran, & Villumsen, 2020). In this way, the decision-making process is defined by the system separating acute danger from a broader risk assessment.

The case: Predictive Risk Modelling in child and family welfare

The research project aims to develop and test whether PRM can support decision-making in social work practice within the context of Danish child and family welfare. First, a predictive risk model built to assist decision-making regarding children at risk was developed – as described in the following section.[1] Then, in pilot test no. 1, social workers tried using the prototype model on notifications, in which a decision had already been made, as well as being interviewed about their decision-making practices (Villumsen & Søbjerg, 2020). Next, the model and the interface were improved based on results from pilot no. 1 and pilot no. 2, and an RCT study was planned but never conducted due to legality issues (see footnote no. 9). So far, the

[1] For an elaborate and detailed description of the entire process, see Rosholm et al. (2024).

PRM is only part of this research project, and therefore not part of any public or private case handling.

The overall idea is to make the information already available to social workers more accessible, and to assist social workers in processing this information when assessing the risk concerning a certain child (Cuccaro-Alamin et al., 2017; Taylor, 2020; Vaithianathan et al., 2013). The model has been developed as a tool that may support human knowledge-making and decision-making (Lehtiniemi, 2024), together with other types of knowledge used in social work practice, such as theory, research, regulations by the law and service users' experiences. The model is meant to offer an assessment of the risk of a child being neglected or maltreated, based on historical data, when a notification is received by the municipality. It provides the social worker with an estimated risk of the child based on an integer score ranging from 1 to 10. This risk score is accompanied by an information sheet that lists all the inputs in the model. The model is intended to be visible during meetings between social workers and families. The model does not recommend decisions, as a predictive risk model can never stand alone; any statistical approach within social work needs to be integrated with, and subordinate, to human competencies regarding decision-making (Taylor, 2020).

As mentioned, the model is intended to play a supportive role in professional judgement; the social worker will at any given point always have more knowledge about the child or family than the model. It is an obvious limitation that PRM produces probable, but not certain knowledge on a particular child (Mittelstadt et al., 2016). Therefore, the model cannot replace the decision-making process of a social worker; it serves as a model to support decisions while still allowing room for professional discretion and maintaining core values in social work (Keddell, 2023; Cheng et al., 2022).

Developing the PRM: Method

When constructing our predictive risk model, we used comprehensive Danish administrative data. Our model is based on data from children for whom a notification of concern was sent to the child and family welfare service in 2016 or 2017. Using subsequent removal and placement in out-of-home care as a proxy for child maltreatment, we estimated four different models that vary in their degree of complexity to predict child maltreatment.

The models we considered are the logistic regression model, a logistic regression model combined with least absolute shrinkage and selection operator (LASSO), the random forest model and, finally, the XGBoost model.[2] The two former models belong to the class of generalized linear models which are often popular model choices due to their simplicity and interpretability. The latter two methods are ensemble methods, in which the outputs of many classification trees are aggregated in order to produce a model output. As such, trees are highly interpretable; however, this is not the case when multiple trees are combined. Hence, the predictions based on the random forest and the XGBoost models are not as easy to explain as the two other considered algorithms. On the other hand, these methods are typically found to deliver highly accurate predictions in many different contexts. A typical explanation for this phenomenon is the fact that these models are good at capturing non-linearities and interaction effects in high-dimensional datasets, which fits well with the non-linearities and interaction of different circumstances in the lives and development of children at risk (Villumsen, 2017). Based on this approach, we chose XGBoost as our preferred model due to its good predictive power.[3]

In the ML literature, a wide range of accuracy metrics exist that can be used to evaluate the performance of a ML model. We have chosen to use the area under the receiver operating characteristics curve (AUC) as the evaluation metric. The AUC is used as the main evaluation metric, in which according to sweets (1998), a score above 80% indicates a good predictive performance. The AUC ranges between 50% to 100%, in which a value of 100% indicates that the model can perfectly predict which children will be placed in the future. More generally, the AUC score can be regarded as the model's ability to discriminate between positive instances (i.e. children experiencing out-of-home placement) and negative instances (i.e. children

[2] For a detailed description of the four models and the exact implementation details, we refer to Rosholm et al. (2024).

[3] It is important to stress that the model is an algorithm developed *ex ante,* and predictions do not change over time. The model does not train while in use.

not experiencing out-of-home placement). The AUC scores range from 83.93% to 85.91%, depending on the choice of ML model.

Our final model XGBoost has good predictive power, see Table 1, as it can distinguish between maltreatment and non-maltreatment, with a probability of approximately 86% (area under the curve (AUC) ≈ 86%). More precisely, there is an 86% probability that the model will assign a higher risk assessment to a child who is placed outside the home after the municipality has received a notification relative to a child who is not placed.

Method	AUC (%)	95% confidence interval
	Outcome: Out-of-home placement within 120 days	
Logistic regression	83.93	83.06-84.79
Logistic regression w. LASSO	83.53	82.66-84.39
Random forest	85.69	84.87-86.52
XGBoost	85.91	85.0986.73

Table 1: This table summarizes the predictive performance of the considered ML models. The models were trained on a data set comprising 120,395 notifications (representing 63,303 unique children), and were evaluated on a sample of 52,649 notifications (representing 27,341 unique children) received during the period from April 2016 to December 2017.

To further evaluate the model's predictive qualities, we compared its predictions to a wide range of adverse child outcomes that indicate maltreatment (e.g. whether a child has excessive school absence, or has suffered a fracture or tooth injury). This validation exercise showed a monotonic positive relationship between the risk scores based on the model and the alternative maltreatment outcomes, thereby suggesting that the model is predictive of child maltreatment in a broad sense. Taken together, these pieces of evidence suggest that the combined use of statistical methods and administrative data identify children at risk with good precision (Rosholm, et al., 2024).

We also tested the model with regard to its ability to reduce errors in decision-making, and showed that the predictions hold a strong potential to reduce such errors. Indeed, we found that 60% of the notifications did not give rise to any concern during the investigation period, but that subsequently led to an out-of-home placement, belonged to the top two decile of the predicted risk distribution. We also showed that our predictions could potentially help reduce social worker biases; even though we found no differences in the manner in which social workers treat children with a similar risk of maltreatment but a different ethnic background and gender, we found that they tend to treat children from diverse socioeconomic backgrounds differently in the sense that children from better socioeconomic conditions were less likely to be placed. These results suggest that child and family welfare identification (or speed of identification) of severe risk of maltreatment cases may indeed be improved, and can potentially be improved by incorporating PRM into the decision-making process by increasing caseworkers' awareness of potential biases, and pointing to potentially high risk cases. However, working with PRM in decision-making does not have the potential to control uncertainties in social work.

Discussion

To work systematically, we chose to frame our reflections and discussions based on the framework put forward by Morley et al. (2020), combined with central documents on machine learning (ML) in social work with children and families (Leslie et al., 2020), in addition to ethical guidelines for trustworthy AI (High-Level Expert Group, 2019[4]).[5] Now, frameworks like these can be criticized for contributing to creating a gap between principles and practice, because when applied by, for instance, tech companies, they can potentially become a set of 'rules' that can be 'resolved' with simple measures without being clear about who gets to decide - for instance - what is fair and for whom (Munn, 2022).

In this case, we argue that a framework is useful because it creates a frame for both development and testing of the PRM in social work practice (Munn, 2022). In our case, we apply it in relations to model building. The particular framework has been

[4] The key requirements derive from fundamental rights stated in the EU treaties and the EU charter, which the High-Level Expert Group in the context of AI systems has narrowed down to principles or values of respect for human autonomy, prevention of harm, fairness and explicability. In our discussion, we have incorporated elements from the High-Level Expert Group that was not already part of the framework. As stated, this is merely to work systematically through ethical issues in relation to the use of ML and PRM.

[5] It can be discussed whether the predictive risk model resembles an artificial intelligence (AI) system, especially in regard to rationality (High-Level Expert Group on AI, 2018). Ethical considerations arise irrespective of whether we consider the predictive risk model to be artificial intelligence or not, and in parts of the literature no distinction is made (see e.g. Morley et al., 2020).

chosen because it is developed especially in relation to ML: it is based on a systematic literature search, and it has an explicit focus on closing the gap between principle and practice. Also, the main strength is that in the creation of this framework, there is a close comparison of ethical principles concerning ML (Morley et al., 2020). From the framework, we have chosen four ethical principles - relevant to model building - to be discussed: Non-Maleficence, Explicability, Autonomy and Justice. In the following, we will present and discuss these principles in relation to our case. However, it is an ongoing issue that when developing and testing these kinds of models, collaboration with social workers and families are paramount, and needs to go hand-in-hand with both qualitative and quantitative research.

Non-maleficence

The principle of non-maleficence is about preventing harm to anyone involved in PRM or AI systems. It concerns privacy, robustness and safety, as well as responsible use (Morley et al., 2020; High-Level Expert Group, 2019; Leslie et al., 2020). In relation to our case, we address data quality and privacy.

Machine learning models are only as good as the data on which they are trained. Hence, **data quality** is a serious concern when building a PRM, as a PRM based on poor data quality could potentially seriously harm the individuals involved (Keddell, 2023). At least two issues arise. First, the precision of the data on which the ML model feeds is important. Second, the outcome on which the model is trained should represent a good measure of what it is supposed to capture. The ethical consideration here is trustworthiness.

We believe that the first issue is largely addressed by relying on register-based data only. In the Danish context, such data is reputed for being accurate and extensive.

The second issue concerning data quality is more problematic. If the outcome on which the model is trained is not a good measure of what it is supposed to capture, we risk attributing potential maltreatment risk to families not at risk, while overlooking families needing help. We aim to detect children at risk of maltreatment, but there is no readily available measure of ‘maltreatment’ in the administrative registers. Hence, when building the model, we have used ‘subsequent placement outside the home’ as a proxy for (severe) maltreatment. This choice is definitely debatable. We have also experimented with using information on less severe interventions (interventions applied in the family) and on future (severe) notifications. We have validated our preferred measure against alternative outcomes that we also believe are indicative of maltreatment, such as hospitalizations due to fractures, mental health diagnoses, mandatory measurements of well-being in school and dental quality measures. The predicted risks on our prediction sample are highly correlated with all these outcomes, also for children who are not subsequently placed outside the home (see Rosholm et al., 2024). We take these results as a strong indication that our risk measure is closely related to maltreatment. Nevertheless, we acknowledge this risk, and we will continue to analyse its validity, and, if possible, improve upon it, as the project evolves.

Another way the use of PRM poses a risk of harm to the individual is by violating its right to privacy. Any AI system must guarantee **privacy** and data protection throughout the system’s entire lifecycle; to secure that the data of individuals are not used in an unlawful manner, that data used are updated and correct, and that only authorized staff can access data. In our case, the information upon which the model is based is already used in the municipalities, when assessing notifications. In this process, social workers and local municipal authorities are obliged to adhere to the GDPR rules and other regulations. In addition, when building the model, we rely on pseudonymized data delivered from municipalities to Statistics Denmark, where they are located on secured servers where only approved researchers may access them. To sum up, when built this way, data quality should be high and easy to verify, and data security should also be high, thereby assuring adequate safety and privacy for children and families at risk.

Explicability

The principle of explicability is about making sure that the predictive risk model is understandable, transparent and accountable (Morley et al., 2020; High-Level Expert Group, 2019; Leslie et al., 2020). In relation to our case, we discuss transparency, understandability and accountability.

It is important that all processes **have a large degree of transparency and is understandable** to increase the trustworthiness of the project as a whole, and the explicability and usefulness of the PRM in particular. It is important that the data sources and the algorithm behind the model are documented extensively.

As mentioned earlier, we considered four different supervised ML models of varying complexity in the developmental phase. As it appears from Table 1 above, we also find that the more complex models deliver more accurate predictions in the present project. However, for these more complex models, it is not straightforward to illustrate how a given input factor contributed to the prediction made by the model. The risk of a non-transparent model is that social workers are more likely to either devalue such a model's risk assessments when they cannot access information about the type of information the model feeds on and the relative importance of these inputs for a given prediction; or that the lack of transparency of the PRM makes it difficult to challenge the prediction. Moreover, such a black-box risk assessment potentially leads to social workers and the involved families feeling less empowered and losing trust in the system. Thus, it is important to consider ways to pry open the black box of PRM, both when building, testing and potentially implementing.

To overcome these issues, the idea is – in a potential test of the model - to provide the social workers with additional information about the model. Each time the model is used to generate a risk assessment, the social worker also receives an information sheet listing all exact values of the model inputs for the present case, as mentioned above. Hence, the social workers are always provided with a detailed description of the information the model relies on, and this enables them to compare the additional knowledge they have about a specific case with the output of the model.

Moreover, to improve transparency, as well as making the model more understandable in a potential test, we would provide so-called *SHAP values* (Lundberg & Lee, 2017) each time the model is used,[6] as *SHAP* makes it is possible

[6] SHAP (SHapley Additive exPlanations) is a method that can explain the prediction of *any* ML model.

to construct a list of model inputs ranked by their importance when generating a given risk assessment. We would use the SHAP values to provide the social workers with information about the model inputs that contributed most to a given prediction in both directions (increasing and reducing the risk relative to a median case). By using SHAP values, it is possible to investigate whether a given factor has an increasing or decreasing influence on the risk assessment provided by the model. We hope that this element will make it more visible and understandable to the user as to how the PRM works and, consequently, easier to challenge the predictions of the predictive model.

The SHAP values should increase both the model's usefulness and its trustworthiness by making it more transparent and understandable. We will continuously investigate how the SHAP values are being used and interpreted by the social workers, and to what extent they find this feature to be valuable. This is important since explainable ML methods such as SHAP are still in their infancy, and evidence is still lacking on how these methods work in practice.

In order to be **accountable**, the model must also be reliable and trustworthy. In the previous section, we argued that the model has a good accuracy based on a sample of historical notifications, and we find evidence that the model is predictive of other adverse outcomes than the one it is designed to predict. To be reliable and trustworthy, however, the model must be reliable for any type of ongoing notification, also when data input is missing. Furthermore, a careful evaluation of the model requires that it is possible to make an exact reconstruction of every prediction made by the model.

With respect to the latter reproducibility requirement, this is automatically fulfilled due to the static nature of the model and the administrative data it feeds on. Once the algorithm has been trained using a sample of historical data, the model is deterministic in the sense that it will provide the *exact* same risk assessment for any two individuals with the exact same background characteristics. Thus, the model does not learn continuously over the course of time. For the model to provide a different risk assessment given certain values for inputs of the model, it is a requirement that we go back to our developmental space on Statistic Denmark's server and recalibrate the model. Hence, by keeping track of which version of the model a given risk assessment is based on, it will always be possible to reconstruct the assessment.

The model reliability criterion is also automatically considered by the choice of the ML method on which the model is based. Our preferred model, the XGBoost model, allows for missing data among the input variables. This means that even in cases of incomplete information about a child in the municipalities' databases, the model will still provide a meaningful risk score based on the non-missing data inputs. This is a very attractive feature of the XGBoost methodology, and a feature not easily accommodated by more standard methods, such as the logistic regression model that was also considered during the developmental phase.

In theory, the model should be applicable for any type of notification. This includes first-time notifications, that is, the first time a child or its family is involved in a notification and notifications regarding newborns. Yet, it could potentially be desirable to set an upper threshold for how much data may be missing before the use of the model is ruled out. In cases with limited information, it is plausible to believe that the model can be imprecise or even misleading, and it might be unethical to use the prediction made by the model to make decisions regarding the child under such circumstances.

In a potential test of the model, the intention is to build in safeguards against this situation, either by flagging predictions based on limited information, or by not calculating the prediction for such cases. This should also enhance the model's trustworthiness in social work practice, and for children and families at risk.

Autonomy

The principle of autonomy is about the protection of autonomy and the ability to make decisions (Morley et al., 2020; High-Level Expert Group, 2019; Leslie et al., 2020). In relation to our case, we discuss human agency and oversight.

Fundamental rights are a central part of the requirements concerning **human agency and oversight**. In relation to human agency (High-Level Expert Group, 2019), AI systems should support the decision-making of the user, and in no way deceive or manipulate users. In addition, users of AI systems should be able to make informed autonomous decisions, and they should possess the knowledge and tools to understand and challenge the system. Leslie et al. (2020) raise similar issues concerning ethics when using ML. Issues of concern include cognitive biases that can influence how users interact with ML models in relation to interpretation of results, as well as potential overuse, underuse, misuse or overreliance.

The predictive risk model we have developed is built to assist social workers in their decision-making process. It provides a risk score that essentially summarizes the information already available to the social worker; an estimated risk that the child is maltreated is represented by an integer score ranging from 1 to 10. In this way, the PRM offers a replicable, but also potentially new perspective, on the notification at hand. It is intentionally built so that it is not capable of suggesting which action should be taken. The interface presents the results of the model by showing the risk score, as well as the information on which it makes its assessment, and how this information has contributed to increasing or lowering the risk score. This also makes it clear that the contents of the present notifiction does not enter into the calculation of the risk score, thereby rendering it impossible to base a decision solely on the PRM, which only captures historical information.

The social workers not only need to understand the scope of the PRM and the correct way of interpreting the risk score, they also need to know the limitations of the PRM. When it comes to the potential use of the PRM, a second pilot study was planned[7] as a way of testing and giving feedback to the building process. It was also planned to test whether the setup around the use of the model provides the needed support for agency and oversight for the social worker and the involved family, so that they – if needed - can challenge both the score and the information used to calculate

[7] This second pilot test phase was postponed due to uncertainties regarding legality. It is obviously a pre-condition that the use of the predictive risk model meets legal requirements. In this article, we will not address issues of the legality, but instead focus on the ethical considerations concerning the construction of the model. The legality of the project has been well examined by an external law firm (see https://childresearch.au.dk/udsatte-boern-unge-og-familier/projekter/underretninger-i-fokus), and discussed with other relevant external stakeholders. However, questions on legality are still debated, and therefore further pilot testing and RCT studies have been postponed.

it. In this pilot, we planned to follow the use of the PRM through qualitative data. In particular, the perspectives of the social workers and the families are paramount in potential testing or use of the PRM when addressing human agency and oversight. Do the social workers understand the scope and correct way of interpreting and challenging the PRM in decision-making situations? Do they recognize the limitations of the PRM? Are they able to explain these elements to families, and how do families experience and react to this? Finally, after the second pilot, and before any large-scale implementation, the plan was to test the usefulness of the model in a subsequent randomized trial, in which social worker decision-making and subsequent actions are connected to outcomes for children and families (see footnote no 9). In short, pilot testing in a research setup provides important feedback when building a PRM model. Research is paramount, both for providing important knowledge and especially for providing a feedback setup where families and social workers can participate without the fear of repercussions.

Justice

The principle of justice is primarily about fairness, but is also about minimizing discrimination and bias (Morley et al., 2020; High-Level Expert Group, 2019; Leslie et al., 2020). In relation to our case, we discuss bias, discrimination, accuracy and fairness.

Since supervised ML models make predictions based on past social and cultural patterns, regardless of whether these are discriminatory, there is a strong risk of perpetuating **systemic biases**, and to the extent that it affects social workers' decision-making in discriminatory direction, it may even amplify such biases. This causes an ethical problem in relation to fairness. Therefore, we took a closer look at age, gender, ethnicity and socioeconomic status.

The most important issue that we identified is the age of the child. Historically in Denmark, out-of-home placements have mainly occurred for the oldest children (14-17 years old). This will be captured by a supervised ML model trained to predict this outcome. As the true latent child maltreatment will not necessarily follow the same pattern with respect to age, this has been taken into account by constructing decile risk scores based on the age-specific distributions of the model's predictions, hence removing all age differences in the risk score.

Furthermore, we have analysed the extent to which the predictive risk model affects existing differences in relation to gender, ethnicity and socioeconomic status (SES). *In an unpublished work, we also* analysed model predictions in the prediction sample split by ethnicity, gender and SES. We find weak tendencies that social workers may have discriminated slightly on ethnicity. We find no evidence of past discriminatory behaviour by gender. However, we do find evidence of differential decision-making based on SES in the sense that children in high-SES families are much less likely to be removed, given their risk score. Therefore, no variables on SES (nor gender or ethnicity) are included in the information set on which the algorithm can feed. Thus, two notifications regarding children of different SES, but where all other included factors that are identical will be given the exact same score by the model (this has been tested and verified).[8] It is an integral part of the project to continuously monitor the predictions for any systematic biases, and adjust the predictive risk model accordingly. In addition in a potential test of the model, the project intends to introduce and instruct social workers carefully in the potentials and pitfalls of using the predictive risk model in social work, in the hope that we may guide them to using the predictive risk model such that it promotes non-discriminatory behaviour.

In relation to **accuracy,** the purpose of the model is to provide social workers with a risk assessment of children who are at risk of maltreatment. To portray a high level of justice, the model must deliver accurate predictions. In Table 1, we reported a summary of the predictive performance of the four different supervised ML models that were considered in the developmental phase, and concluded that our preferred model, the XGBoost model, had a good predictive performance. As mentioned, the models aim to identify children suffering from maltreatment. Nevertheless,

[8] Of course, it might be that high-SES families have more resources (social and physical capital, such as good relationships with a supportive wider family and social network) on which to draw and make continuing care at home feasibly 'safe enough', even though these factors are not measured by the model. Thus, the *correlation* with the 'discriminatory' factor may be a fact, but may not indicate *causation*. Hence, it may not reflect a bias as much as mediating factors, and a deeper knowledge of the wider context from the perspective of the professionals involved face-to-face in the situation (see also Søbjerg et al., 2018, for a discussion of this).

maltreatment is not a directly observable variable. To help overcome this problem, we follow the same approach as was used to develop a similar model in Allegheny County, Pennsylvania, and use out-of-home placements as a proxy for maltreatment (Goldhaber-Fiebert & Prince, 2019). Despite limitations of this outcome, we find evidence that its external validity suggests that it can be used to distinguish between children at low risk and at high risk of maltreatment.

As already mentioned, Rosholm et al. (2024) also explores the relationship between the XGBoost model predictions and other adverse outcomes associated with maltreatment, such as mental illness, criminal charges and illegal school absence. They show that children identified to be at high risk as defined by the model predicting out-of-home placements are also worse off when considering these alternative proxies for child maltreatment. Thus, children identified to be at high risk by the model are arguably the children that are most in need of help.

We therefore, at present, conclude that the predictive risk model has sufficient accuracy and trustworthiness. Note, however, that accuracy was also intended to be further analysed in the randomized trial, and this or any other model should not be recommended for social work practice without a further very detailed analysis of this issue based on results obtained from such a trial.

For a predictive risk model to be applicable in social work in practice, it is of utmost importance that the model is **fair and does not induce discriminatory behaviour**. We have already addressed this issue in part. In this section, we elaborate on these issues.

In order to prevent discriminatory assessments due to basing risk scores on background characteristics such as race and ethnicity, we have decided that the predictive risk model will not incorporate any information about the ethnic origin of the children and their parents.[9] By leaving out information on ethnicity from the set of maltreatment predictors, we can be certain that the differences in risk scores among

[9] The model does not contain information about the number of emigrations and immigrations, since this information might be highly correlated with ethnicity

a group of children cannot be directly attributed to differences in the ethnic origin of the children and their families. Even so, there is still a possibility that ethnicity can indirectly influence the risk score in case there is a strong correlation between some of the input variables and ethnicity. Consequently, it is essential to prospectively monitor that the model does exacerbate inequalities between groups. Our preliminary results reassuringly suggest that leaving out information on ethnicity does not change the predictive performance of the model, nor does it appear that social workers in the past have discriminated against certain ethnic groups in terms of decisions made.

As previously discussed, we remove all age dependencies in the model, and exclude information on SES and gender from the predictive risk model. For this reason, only differences in the risk scores unrelated directly to age, gender, ethnicity or SES drive the variations in placement rates, which is exactly the condition that a well-calibrated risk model must satisfy according to the fairness criterion put forward by Chouldechova et al. (2018). This fairness criterion is labelled the *calibration criterion* and states that, conditional on the algorithmic risk score, there must be no differences in the observed placement rates between every group of individuals (e.g. gender or ethnic groups).

Conclusion

In this article, we present and discuss ethical considerations that arise when building a PRM for potential use in social work with children and families are at risk.

The first important concluding point is the importance of ethical considerations, not only when PRMs are tested or implemented in social work practice, but in particular when the PRM is constructed. This is not always a very transparent process, as it is often performed by tech companies who are not subjected to the demands of transparency.

The second very important concluding point is the need for more both qualitative and quantitative research when trying to build, test or implement PRM in social work with children and families at risk. There are several reasons. First, the strength of connecting research to the processes is that that research is subjected to a review process from other researchers in the field. At the same time, good research is characterized by transparency that is supported both through the review processes, but also in the publications themselves. Mandatory social work with children and families in adversity is a highly sensitive field, so therefore we need open and transparent discussions of decision-making support systems, such as a PRM, if we are ever to understand and create evidence for both potential benefits and risks. Second, qualitative research has the potential to document the perspectives of children, parents and social workers, and thereby document the potential pitfalls and usefulness of a PRM in the lives of children, parents and social workers. Third, quantitative research holds the potential to determine whether a PRM, such as the one presented in this manuscript, actually improves decision-making on a large scale and over time. Fourth, independent research is a potential way to ensure that social workers and families are able to share their experiences and knowledge in a space safe from potential repercussions. However, it is also important to be aware of the fact that researchers entering this field are not necessarily equipped, as traditional research ethics principles are not well suited for research within digitalization; therefore, researchers involved often lack knowledge and training in this specific area to conduct ethically sound research (Ada Lovelace Institute, 2022). In addition, local Research Ethics Committees reviewing this kind of work do not possess the resources, expertise and training to properly review the risks that digitalization poses to a certain field. If research on digitalization fails to address ethical issues, we not only risk reducing public trust in the field, but also risk research potentially contributing to harming children and families at risk instead of helping them.

Our final concluding point is the importance of a close collaboration between model builders, social work practice and children and families at risk when building and testing a PRM, so that any potential harm a model might create is solved in the way it is built or in the way it is potentially implemented, rather than it being a matter of social workers and children and families adapting to the model.

References

Ada Lovelace Institute. (2022). *Looking before we leap: Ethical review processes for AI and data science research*. https://www.adalovelaceinstitute.org/report/lookingbefore-we-leap/Ethics and accountability in practice

Cheng, H. F., Stapleton, L., Kawakami, A., Sivaraman, V., Cheng, Y., Qing, D., ... & Zhu, H. (2022, April). *How child welfare workers reduce racial disparities in algorithmic decisions* [Conference presentation]. 2022 CHI Conference on Human Factors in Computing Systems, New Orleans. https://doi.org/10.1145/3491102.3501831

Chouldechova, A., Benavides-Prado, D., Fialko, O. & Vaithianathan, R. (2018). *A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions* [Conference presentation]. 1st Conference on Fairness, Accountability and Transparency https://proceedings.mlr.press/v81/chouldechova18a.html.

Coulthard, B., Mallett, J., Taylor, B.J., 2020. Better decisions for children with 'big data': can algorithms promote fairness, transparency and parental engagement? *Societies, 10*(4), 97. https://doi:10.3390/soc10040097.

Cuccaro-Alamin, S., Foust, R., Vaithianathan, R. & Putnam-Hornstein, E. (2017). Risk assessment and decision making in child protective services: Predictive risk modeling in context. *Children and Youth Services Review*, *79*, 291-298. https://doi.org/10.1016/j.childyouth.2017.06.027

Devlieghere J., PGillingham, P. & Roose, R. (2022): Dataism versus relationshipism: a social work perspective, *Nordic Social Work Research*, *12*(3), 328-338. https://doi.org/10.1080/2156857X.2022.2052942

Gilbert, N. (ed). 2007. *Combatting Child Abuse: International Perspectives and Trends*. Oxford university press.

Gilbert, P., Parton, N. & Skivenes, M. (eds.). (2011). *Child Protection Systems: International Trends and Orientations*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199793358.001.0001

Gillingham, P (2016): Predictive Risk Modelling to Prevent Child Maltreatment and Other Adverse Outcomes for Service Users: Inside the 'Black Box' of Machine

Learning. *British Journal of Social Work, 46*(4), 1044-1058. https://doi.org/10.1093/bjsw/bcv031

Goldhaber-Fiebert, J. D. & Prince, L. (2019). *Impact evaluation of a predictive risk modeling tool for Allegheny county's child welfare office*. Allegheny County.

High-Level Expert Group. (2019). *Ethics Guidelines for trustworthy AI*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

High-Level Expert Group on Artificial Intelligence. (2018). *A definition of AI: main capabilities and scientific disciplines.* European Commission. https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

Kawakami, A., Sivaraman, V., Cheng, H. F., Stapleton, L., Cheng, Y., Qing, D., ... & Holstein, K. (2022, April). *Improving human-AI partnerships in child welfare: understanding worker practices, challenges, and desires for algorithmic decision support*. [Conference presentation]. 2022 CHI Conference on Human Factors in Computing Systems, New Orleans. https://doi.org/10.1145/3491102.3517439

Keddell, E. (2023). The Devil in the Detail: Algorithmic Risk Prediction Tools and Their Implications for Ethics, Justice and Decision-making. In B. Taylor, J. D. Fluke, J. C. Graham, E. Keddell, C. Killick, A. Shlonsky & A. Whittaker (eds.) *The Sage Handbook of Decision Making, Assesment and Risk in Social Work* (pp. 405-420). Sage Publications. https://doi.org/10.4135/9781529614657.n51

Kriz, K. & Skivenes, M. (2013). Systemic Differences in Views on Risk: A Comparative Case Vignette Study of Risk Assessment in England, Norway and the United States (California). *Child and Youth Services Review, 35*(11), 1862–1870. https://doi.org/10.1016/j.childyouth.2013.09.001

Leslie, D., Holmes, L., Hitrova, C. & Ott, E. (2020). *Ethics review of machine learning in Children's social care.* The Alan Touring Institute. Oxford University. https://whatworks-csc.org.uk/research-report/ethics-review-of-machine-learning-in-childrens-social-care/

Lehtiniemi, T. (2024). Contextual social valences for artificial intelligence: anticipation that matters in social work. *Information, Communication & Society*, 27(6), 1110-1125. https://doi.org/10.1080/1369118X.2023.2234987

Lundberg, S.M. & Lee. S. I. (2017) A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (eds.), *Advances in Neural Information Processing Systems 30 (NIPS 2017).* https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S. & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 1-21. https://doi.org/10.1177/2053951716679679

Morley, J., Floridi, L., Kinsey, L. & Elhalal, A. (2020). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and engineering ethics*, *26*(4), 2141-2168. https://doi.org/10.1007/s11948-019-00165-5

Munn, L. (2023). The uselessness of AI ethics. *AI and Ethics*, *3*(3), 869-877. https://doi.org/10.1007/s43681-022-00209-w

Pösö, S., Skivenes, M. & Hestbæk, A.-D. (2013). Child Protection Systems within the Danish, Finnish and Norwegian Welfare States – Time for a Child Centric Approach?. *European Journal of Social Work 17*(4), 475–490. https://doi.org/10.1080/13691457.2013.829802

Rosholm, M., Bodilsen, S. T., Michel, B. & Nielsen, S. A. (2024). Predictive risk modeling for child maltreatment detection and enhanced decision-making: Evidence from Danish administrative data. *PLOS ONE, 19*(7), e0305974. https://doi.org/10.1371/journal.pone.0305974

Søbjerg, L. M., L. Nirmalajaran & A. M. Villumsen. (2020). Perceptions of Risk and Decisions of Referring Children at Risk. *Child Care in Practice 26*(2), 130-145. https://doi.org/10.1080/13575279.2019.1685460

Søbjerg, L.M., Taylor, B.J., Przeperski, J., Horvat, S., Nouman, H. & Harvey, D. (2020). Using risk-factor statistics in decision making: prospects and challenges. *European Journal of Social Work, 24*(5), 788-801. https://doi.10.1080/13691457.2020.1772728

Taylor, B. J. (2020). Teaching and learning decision making in child welfare and protection social work. In J. Fluke, M. López López, R. Benbenishty, E. J. Knorth, & D. J. Baumann (Eds.), *Decision making and judgement in child* *welfare and protection: Theory, research and practice* (pp. 281–298). Oxford University Press. https://doi.org/10.1093/oso/9780190059538.003.0013

Vaithianathan, R., Maloney, T., Putnam-Hornstein, E. & Jiang, N. (2013). Children in the public benefit system at risk of maltreatment: Identification via predictive modeling. *American Journal of Preventive Medicine*, *45*(3), 354-359. https://doi.org/10.1016/j.amepre.2013.04.022

Villumsen, A. M. (2017). Hvorfor det ikke er så lige til med udsathed hos børn og unge. In. D. Graversen (ed.), *Pædagogik: introduktion til pædagogens grundfaglighed* (1st ed.). Hans Reitzels Forlag.

Villumsen, A. M., & Søbjerg, L.M. (2020).: Informal Pathways Informal pathways as a response to limitations in formal categorization of referrals in child and family welfare*. Nordic Social Work Research, 13*(2), 176-187. https://doi.org/10.1080/2156857X.2020.1795705