Predicting child labour risk at household level

A risk model for cocoa farming households in Ghana

October 2020

*Addended May 2021

Supported by:

Swiss Platform for Sustainable Cocoa

Schweizerische Eidgenossenschaft
Confédération Suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs FDEA
State Secretariat for Economic Affairs SECO

Nestlé®
Cocoa Plan

This study was conducted by ICI in the context of a project financed by the State Secretariat for Economic Affairs (SECO) of the Swiss Confederation, in the framework of the Swiss Platform for Sustainable Cocoa, as well as by the Nestlé Cocoa Plan.

The study was written by Anna Brüderle.

We are grateful to peer reviewers from the University of Sussex for their helpful comments and feedback: Rachel Sabatees-Wheeler, James Sundberg and Keetie Roelen.

CONTENTS

EXECUTIVE SUMMARY

Objective

In cocoa-growing areas of Côte d'Ivoire and Ghana, an estimated one in three children are in child labour.[1] Most of these children work on family farms, alongside their relatives, doing activities that are considered "hazardous" under national legislation. While effective approaches exist to address child labour, their coverage remains limited compared to the overall need – effective approaches to protect children from hazardous work need to be scaled up to fill the gap. To achieve maximum impact with limited resources, it is essential that support is directed to the places where it is most needed, and that the people who need it most are prioritised for assistance. To do so, we need to be able to identify cocoa-farming households at the highest risk of using hazardous child labour.

This paper proposes a **prediction model to identify cocoa farmers in Ghana who are at an elevated risk of using hazardous child labour**, based on basic information about cocoa-producing households and their farms. Such information is commonly available in registers of cooperative members or certified producers in a supply chain; in some communities, local child protection committees maintain community registers, containing similar information.

The model is developed in two stages. First, we use data from child labour prevalence surveys to understand which household and farm characteristics can help predict the likelihood of hazardous child labour and how these characteristics act in combination; we use these data to define and calibrate a model. Second, we feed the model with information on farmers from a potential target group, to assign each farmer a risk score on a scale from zero (very low risk) to one (very high risk). This risk score can be used to ensure that higher-risk households from the target group are prioritised for monitoring and support, so that limited resources can be allocated more efficiently.

Method

To define and calibrate the risk model, we use data from the *Survey Research on Child Labor in West African Cocoa Growing Areas*, collected in 2013/14 by Tulane University.[2] The survey is based on a nationally representative sample of households in cocoa-growing areas in Ghana. While the Tulane data contain a wealth of child, household and farm characteristics, we consider only those characteristics which are typically available in a member register of a cocoa producer organization. First, we analyse these data to **identify the most powerful predictors of hazardous child labour** from the available household characteristics. According to the analysis, these are: education level, age and gender of the household head; the number of children living in household; the number of workers employed; the household's drinking water source and electricity access; the size of land under cocoa cultivation; whether the household cultivates other cash crops; and the use of fertilizer and pesticides. To understand how these combined factors are associated with hazardous child labour, we estimate a logistic regression model. In simplified terms, the logistic regression model is a formula which calculates from the different risk factors a value between zero, indicating no hazardous child labour, and one, indicating hazardous child

[1] Tulane University (2015) *Survey Research on Child Labor in West African Cocoa Growing Areas*

[2] Tulane University (2015) *Survey Research on Child Labor in West African Cocoa Growing Areas*

labour. .[3] Once the formula is defined and calibrated based on the Tulane survey data, **in a second step the model is fed with data on children from the target group of farmers, to produce a hazardous child labour risk score for each child in the household.**

For this second step, we use data from **a separate survey of child labour prevalence,** conducted **among a sample of 705 cocoa producing households targeted by the programme, covering 1,541 children in Asunafo South and Suhum districts of Ghana**. The survey collected characteristics of the household and the farm shown to be associated with hazardous child labour in step one, as well as information about children's actual involvement in hazardous child labour. First, we enter the household and farm data into the model to calculate a risk score for each child, which *predicts* whether or not the child engages in hazardous child labour. Then, to assess how well the model predicts hazardous child labour among the target sample, **we compare the *predicted* outcome to the *observed* outcome in the survey**.

Results

Using the basic first version of the model, which is strictly limited to the set of household and farm characteristics available in our reference farmer register, we find that the model's prediction power is very poor: it correctly predicts hazardous child labour for 60% of the actual hazardous child labour cases observed, but also falsely predicts *hazardous child labour* for 61% of the *non*-hazardous child labour cases.

We **test various modifications of the model to find out how prediction power can be improved**. Two types of modifications turn out to enhance prediction power and yield a model that can be operationalized for a risk-based targeting mechanism: first by **adding relevant information to the model**; and second by **using a measure of child labour severity.**

When we add information to the model, using various indicators associated with hazardous child labour risk according to previous research, **the most powerful parameters to improve the prediction model are the child's age and sex**. The modified model now obtains a level of predictive power that could be used to target interventions more efficiently: the enhanced model correctly predicts 58% of hazardous child labour cases and correctly predicts 63% of the non-hazardous child labour cases. In practical terms, when a project aims to identify cases of hazardous child labour through household visits, this means that instead of visiting every household in a given cooperative or community, only "higher-risk" households could be visited to identify the same number of children in hazardous child labour, thereby reducing the cost of monitoring and increasing the speed at which support to identified cases could be provided.

> A key recommendation emerging from these results is that for the purpose of hazardous child labour risk prediction, data collection for farmer registration should include demographic information about individual children living in the households - specifically, children's age and sex.

[3] Hazardous child labour is defined here following Ghana's Hazardous Child Labour Activity Framework for the Cocoa Sector, which provides a list of tasks in cocoa production which are considered hazardous and therefore illegal for children.

We also explore whether the model can be improved by using different outcome measures which better reflect nuances between "severe" and "less severe" cases of child labour, rather than a binary (yes/no) indicator of hazardous child labour. One alternative measure that proves useful to improve the model is **the number of hours a child has worked during the last week**. A model using this measure of child labour severity **correctly predicts 64% of the 'higher-intensity' child labour cases, and 64% of the "lower-intensity" cases**.

Conclusions

The method presented here for building a hazardous child labour risk model for cocoa farmers is applicable to a **wide range of contexts, where (i) a set of basic household and farm characteristics is available for a large group of potential target households, and (ii) existing child labour prevalence data from a comparable context are available to understand child labour risk patterns**.

The model can make only **probabilistic predictions**, but the paper shows that its use could result in tangible **cost savings**, by reducing the time and resources needed to identify children in hazardous child labour.

We present some simple calculations for a scenario where a Child Labour Monitoring and Remediation System (CLMRS)[4] is rolled out to a group of cocoa-farming households, where the hazardous child labour prevalence rate is 46%. If instead of visiting every single household to identify hazardous child labour, visits are conducted based on risk predicted from their known characteristics, 20% of the monitoring cost could be saved, while still identifying the same number of hazardous child labour cases. This would free up considerable resources for activities to prevent child labour and remediate identified cases.

Addendum, May 2021:

Since the publication of this report, ICI has continued to develop child labour risk models in other operational contexts and has tested different statistical methods for child labour prediction. From these more recent trials, it has emerged that by using multilevel regression models, which better account for nesting of prediction accuracy can be improved considerably data within groups (notably, children within households). In the context of the present study, replacing a basic logit model by a multi-level logit has increased the rate of correctly predicted cases of child labour from around 63% to 72%. This improvement in accuracy further increases the potential cost-savings from using such a risk-based approach.

For further details, please see, *Addendum: Modifying the risk model to improve prediction power by using multi-level regression models* (p18).

[4] A CLMRS is a structure embedded in a supply chain, which aims to identify, prevent and remediate cases of child labour. The system is based on the presence of facilitators within cocoa-growing communities who raise awareness on child labour, identify children in or at risk of child labour, deliver prevention and remediation support to children, and monitor their progress over time.

INTRODUCTION

Child labour is a widespread practice in cocoa-growing households in West Africa. According to research by Tulane University (2015), 34% of children in agricultural households in cocoa-growing areas were involved in hazardous child labour in cocoa production in Ghana and Côte d'Ivoire in 2013/14. Various stakeholders in the West African cocoa sector, including governments, the international chocolate and cocoa industry, and civil society actors, are undertaking effective efforts to promote child protection and prevent hazardous child labour in cocoa production. Given the magnitude of the problem, there is a need to further scale up these efforts.

Until now, many approaches either use needs assessments to select beneficiaries, which can be costly, or select beneficiaries independent of specific needs or vulnerability. Child Labour Monitoring and Remediation Systems (CLMRS), for example, have typically been rolled out to all members of a cooperative, meaning that all farming households in a cooperative are visited by a monitoring agent to assess whether the household is using hazardous child labour. Given the considerable cost implications of visiting each and every household before support can be provided, it has been challenging for such approaches to be rolled out at scale, despite the ambition of many actors to ensure 100% coverage of supply chains.

The same targeting dilemma is relevant for individual and household approaches (like the CLMRS), as well as community-based approaches, where large numbers of communities are assessed – often including costly child labour prevalence surveys – before support is provided. In both cases, we are faced with the same question: with limited resources, how can we direct appropriate support, more quickly, to where it is most needed?

Building on previous work to develop methods to identify *communities* with elevated child labour risk,[5] the focus of this paper is on the challenge of identifying at-risk *households*. We propose a **model to help identify which farming households have an elevated risk of hazardous child labour**, based on basic household and farm information. Such information is commonly available for famers who are part of organised producer groups, such as certified cooperatives. To develop the prediction model, data from child labour prevalence surveys are used to understand which household and farm parameters can help predict hazardous child labour risk among cocoa farming households. The model is then calibrated based on the prevalence survey data. When fed with information on farmers from a potential target group, the model assigns each child a risk score on a scale from zero (very low risk) to one (very high risk). This risk score can help prioritize high-risk households for monitoring or support, thereby channelling limited project resources more efficiently.

This paper explores whether it is possible to use an existing data set on child labour prevalence in cocoa-growing households in a given context to predict hazardous child labour risk among a specific group of cocoa farmers in the same context, for which no child labour specific information exists, but a set of basic household and farm data is available.

[5] See, ICI (2019) *Using community level data to understand child labour risk in cocoa-growing areas of Côte d'Ivoire and Ghana.*

The hazardous child labour risk model was developed as part of an innovation project on supply-chain based child labour monitoring and remediation in Ghana.[6,7] We use this example to develop and test the model. However, the methodology could be applicable to a wide range of contexts. The reference scenario is that a set of basic household and farm characteristics is available for a large group of potential target households, for example from a register of farmers in a producer organization or certified cooperative. Such farmer registers are becoming increasingly common, in response to growing demands for supply chain transparency, and emerging human rights due diligence legislation affecting the cocoa and chocolate industry. By transferring child labour risk patterns observed in a child labour prevalence survey undertaken in a comparable context, these data are used to estimate the risk of hazardous child labour for each farming household.

The risk model predicts the *likelihood* of hazardous child labour among a group of farmers. This paper estimates the potential cost savings of applying the approach across a large group of farmers, to prioritise households with higher predicted risk for support.

METHOD

This paper explores whether an existing data set on child labour prevalence in cocoa-growing households can help predict hazardous child labour risk among *a specific group* of cocoa farmers, for which no child labour specific information, but a set of basic household and farm data is available.

A two-stage procedure was followed to approach the question. First, an existing child labour prevalence data set (Tulane University's child labour prevalence data collected in 2013/14 in cocoa-growing areas of Ghana) was analysed to identify characteristics of cocoa farming households that are indicative of high risk for children in this household to engage in hazardous child labour, and to calibrate a model which predicts hazardous child labour risk based on these characteristics. Second, a child labour prevalence survey was conducted amongst a sample of target farmers, to determine how well the *predicted risk* of hazardous child labour compares to the *observed outcomes* in the survey.

[6] For an introduction to Child Labour Monitoring and Remediation Systems (CLMRS), see a report by ICI here.

[7] The child labour risk model was developed as part of a pilot project titled "Targeted Income Support to Vulnerable Households to Reduce Child Labour", implemented by the International Cocoa Initiative (ICI), in collaboration with a cocoa trading and a chocolate manufacturing company, over a 2 year period from May 2019 to April 2021, with funding from the Swiss State Secretariat for Economic Affairs (SECO). The project is testing new approaches to reducing the prevalence of child labour amongst vulnerable cocoa growing households in Ghana in two key areas: (i) a risk-based approach to monitoring and remediation of child labour in supply chains; and (ii) direct income support to vulnerable farming households. For the first project component, available information on targeted farmers is exploited in order to identify farming households with elevated risk of child labour. A Child Labour Monitoring and Remediation System (CLMRS[7]) is then rolled out, whereby monitoring is focused on higher-risk farmers.

Analysis of hazardous child labour risk factors from a nationally representative child labour survey

Data and method for a survey-based model of hazardous child labour

In the first stage, data from a nationally representative child labour prevalence survey were analysed to identify household and farm characteristics which are indicative of high risk for children in this household to engage in hazardous child labour, and understand how these factors interact with each other. The data used were from the "Survey Research on Child Labor in West African Cocoa Growing Areas », collected in 2013/14 by Tulane University with funding from the United States Department of Labor (USDOL). At the time this research was undertaken, data from the Tulane survey, which is nationally representative of all cocoa-growing areas in the country, were considered the most comprehensive and well-documented data available on child labour in cocoa production in Ghana.

While the Tulane dataset contains a wealth of child and household characteristics, due to the action-orientated nature of the study, we considered *only* those characteristics which typically appear in registers held by cocoa producer organizations, with the aim that such an approach could be replicated by other stakeholders using their own farmer registers.

Generally, such registers contain household and farm level information only, with only limited details on household members other than the farmer. Since the risk model discussed in this paper was developed as a part of a supply-chain based child labour monitoring and remediation project, in partnership with a supplier and retailer, we used a register of certified producers provided by the supplier. The register covers farmer societies in two districts of Ghana, Asunafo and Suhum, and contains basic demographic and socio-economic information about the farmers, their household, cocoa production, and farming practices. The data is collected by Field Trainers from the buying company, through interviews with the farmers and recorded using mobile data collection.

The following information about households was available both in the farmer register and in the Tulane survey data:

- age, gender, marital status and level of schooling of the farmer
- total number of people, and number of children aged 5 to 17, living in the household
- total household income during last 12 months, from cocoa farming and from other sources
- total cocoa production during last 12 months
- other types of agriculture carried out by the household
- size of land owned by the household, and under cocoa cultivation
- use of and spending on fertilizer and pesticides
- household's access to electricity and drinking water
- number of adult and child workers employed in the last 12 months

The Tulane data were used to develop a hazardous child labour risk model based on these parameters. Even though we fed only household and farm level information into the model in this first step, we chose the child as the unit of analysis when building the model. This is because this allowed us to experiment with extending the model by child characteristics to improve its performance in a subsequent step.

A **logistic regression model** was estimated on the Tulane child data set, with:

- an indicator for whether or not the child engaged in hazardous child labour[8] or not as the outcome (the dependent variable); and
- a formula which calculates from the hazardous child labour risk factors (as identified in the first stage of data analysis, described above) a value between zero and one.

In simple terms, the logistic regression model is an equation with a binary outcome indicator (which can take on values 0 or 1) on one side, and a combination of elements on the other side, where each element is a risk factor (e.g. land under cocoa production) multiplied by a coefficient, and then processed through a logistic function which yields a value between zero and one. The logistic regression estimation then finds the best values for these coefficients, i.e., the values that make the equation the "best fit" for the data on which it is estimated. The regression model therefore takes into account the fact that various risk factors are at play simultaneously. Once the coefficients ("weights") are set by estimation on a given sample, the model can be used to predict the outcome by feeding it with values from a different (or the same) sample of farmers. **By nature of the logistic model, it will produce the prediction as a value between zero and one.** A natural cut-off to decide whether the predicted outcome is a "yes" (1) or a "no" (0) would be 0.5; but other cut-offs can be chosen, as we discuss in the following section.

Results from the survey-based risk model

Only children from cocoa-farming households within the Tulane sample were included in the estimation, resulting in a total sample of 1'544 children living in cocoa-growing areas of Ghana. A forward *stepwise logistic regression* was first run to **identify from the list of available household and farm parameters above those most relevant for predicting hazardous child labour**.[9]

Figure 1 shows the parameters that were selected by this procedure and entered the risk model.[10] For each parameter, the table presents average marginal effects resulting from the logistical regression. For each of the explanatory variables, the marginal effects indicate **by how much the risk of hazardous child labour would increase due to a change in the parameter value**. For example, if the head of household had completed primary school rather than having no education (**marginal effect = -0.0253**), the hazardous **child labour risk, measured on a scale from 0 (very low risk) to 1 (very high risk), would decrease by 0.0253 (i.e., by 2.53 percentage points)** for an average child in the sample; if the head of household had secondary education rather than no education, the risk would decrease by a value of

[8] Hazardous child labour, rather than child labour, was used as the outcome variable of interest given that the child labour monitoring and remediation system, which provides the context for this research, targets hazardous child labour. According to the Tulane survey, the large majority of children in child labour – over 90% – are in "hazardous child labour".

[9] A forward stepwise regression is a procedure where in each step, a variable is considered for addition to the set of explanatory variables in a regression model. In each step, the variable is added whose inclusion gives the most statistically significant improvement of the model fit. This process is repeated until none of the variables under consideration improves the model to a statistically significant extent.

[10] Education level and age of the household head are measured as categorical variables. The reference categories, which do not appear in the table, are i) having no education, and ii) being younger than 30 years of age, respectively.

0.0446 (i.e., by 4.46 percentage points); and so forth.[11] The asterisks indicate whether the coefficients are statistically significant, and at what level (***indicates a 1% significance level; **5% significance level, *10% significance level).

Figure 1: Summary of risk factors and their average marginal effects[12]

DEPENDENT VARIABLE: Hazardous child labour	AVERAGE MARGINAL EFFECT
education level of household head = 2, primary	-0.0253
	(0.0464)
education level of household head = 3, lower secondary	-0.0446
	(0.0364)
education level of household head = 4, upper secondary or higher	-0.0325
	(0.0510)
age of household head = 3, 30-39 years	-0.0131
	(0.0662)
age of household head = 4, 40-49 years	0.0688
	(0.0681)
age of household head = 5, 50-59 years	0.107*
	(0.0647)
age of household head = 6, 60-69 years	0.141*
	(0.0762)
age of household head = 7, 70+ years	0.0678
	(0.0743)
household head is male	-0.0241
	(0.0323)
# children living in household	-0.000663
	(0.00785)
# adult and child workers employed last 12 months	-0.00262
	(0.00191)
household has an improved drinking water source	-0.103***
	(0.0358)
household has electricity	0.0432
	(0.0384)
land under cocoa cultivation (acres)	-0.00494
	(0.00328)
land under cocoa cultivation (acres), squared	-4.51e-05
	(5.96e-05)
whether household cultivates other cash crops	0.0752**
	(0.0335)
Observations	1'544
Pseudo R-squared	0.0255

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

[11] For the binary risk factors, e.g. whether the head of household is a single woman, the interpretation of the marginal effects reads as follows: by how much would the child labour risk change if the risk factor changed from 0 to 1. E.g. the child labour risk would increase by 0.076 if the head of household was a single woman rather than a man or a married woman, for an average child in the sample.

[12] Reference category for education level of household head is 1, no education; reference category is age of household head = 2, 18-29 years.

The model was constructed to achieve the best predictive power, which influenced the inclusion and exclusion of different variables. Interestingly, income was *not* amongst the parameters that were retained in the model, as it turned out not to improve prediction of hazardous child labour *within* the Tulane data set. This finding is in some ways convenient, since income data are generally difficult to obtain from household surveys in agricultural contexts and can vary considerably depending on the methods used. For this reason, such data would not have been well-suited to a prediction model.[13]

While only few of the parameters are statistically significant, Figure 2, summarizes the factors which the model suggests are associated with lower and higher hazardous child labour risk.

Figure 2: Factors associated with a lower or higher risk of hazardous child labour[14]

Lower risk of hazardous **child labour**	**Higher risk of** hazardous **child labour**
• if household head has completed at least primary education • if the household uses paid labour on the cocoa farm • if the household has a larger area of land under cocoa cultivation • if the household has an improved drinking water source • if the farmer uses fertilizer	• if household head is over 40 years old • if more children live in the household • if the household is headed by a woman • if the household has access to electricity • if the household cultivates other commercial crops

It is important to note that this model is built on a limited set of available household and farm characteristics, as mentioned above, and **these must be interpreted as risk indicators rather than causes of hazardous child labour**. Some of these factors may in fact be correlated with other household characteristics which are the actual underlying causes of hazardous child labour but are not available in a farmer register. For example, access to water and electricity may be indicators of the household's wealth and the community's access to infrastructure, which in turn may be correlated with other factors more directly relevant for hazardous child labour incidence.

[13] For a discussion on different methods available for measuring household income in cocoa-growing communities, see for example: Bymolt, R., Laven, A., and Tyszler, M. (2018). Analysis of the income gap of cocoa producing households in Ghana. The Royal Tropical Institute (KIT), The Netherlands.

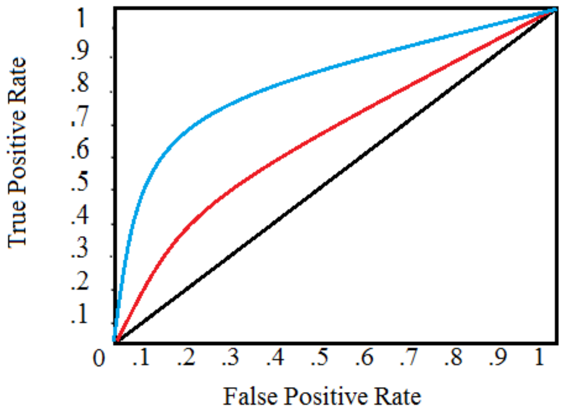
[14] This list contains also factors that are not statistically significant; the relationship between these factors and hazardous child labour is only indicative.

Methods to assess the risk model's ability to predict hazardous child labour

To evaluate how well the model predicts hazardous child labour risk, two concepts are useful to apply: *sensitivity and specificity*. The **sensitivity** of a prediction model is the probability that the model would predict a "yes" outcome for a unit for which the true outcome is "yes", or a "true positive"; in our context, this means the model would **flag high** hazardous **child labour risk for a child who actually engages in** hazardous **child labour**. Conversely, **specificity** is the probability that the model would predict a "no" for a unit with a true "no" outcome, or a "true negative"; in our context, this means the model would **mark low risk for a child who does not engage in** hazardous **child labour**.

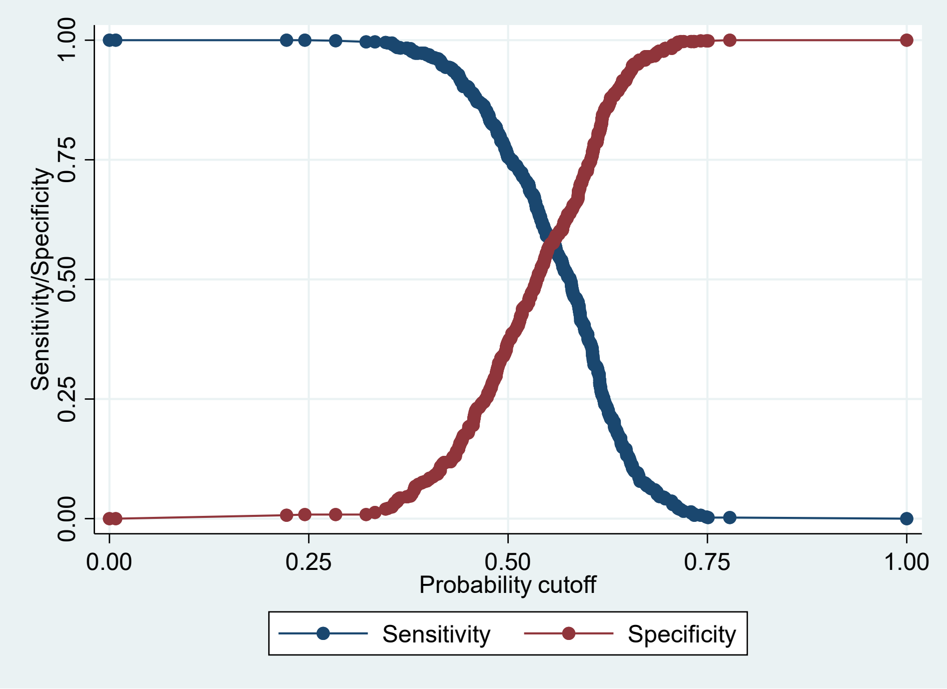
To illustrate this further, if a model predicted hazardous child labour for every child in a sample, it would have a high sensitivity because it would flag all of the actual hazardous child labour cases as high risk; but it would have a low specificity because it would fail to mark the non-hazardous child labour cases as low risk. Hence, we are seeking to build a model that has **both a high sensitivity and a high specificity.** Such a model would have a **high prediction or "discrimination" ability** (indicating that the model is good at discriminating between positive and negative cases). The so-called **receiver operating characteristic (ROC) curve** plots a model's true positive rate (sensitivity) against its false positive rate (1-specificity); and the area under the curve can be used as a measure for the model's ability to discriminate between positive and negative cases (see Figure 3). The **area under the ROC curve** ranges from a value of 1, which corresponds to perfect discrimination ability, to 0.5, which corresponds to a model with no discrimination ability.

Figure 3: Illustrative example of two possible Receiver Operating Characteristic (ROC) curves



The area under the ROC curve for the model as specified above has a value of 0.634. This value provides a benchmark against which we can assess trials to further improve the model, e.g. by adding parameters. Figure 4 plots the model's sensitivity (how well it correctly identifies hazardous child labour cases as "high risk") and its specificity (whether it correctly marks non- hazardous child labour cases as "low risk") when choosing different **cut-off values,** i.e., values of the hazardous child labour risk score which divide the sample into high and low risk. We can see that at a cut-off value of 0.5, the model would correctly predict hazardous child labour for more than 78% of the cases, but would only mark around 33% of the non- hazardous child labour cases as low risk – a large error. The within-sample prediction performance of the model will also be discussed in the next chapter, for the baseline model described here, and for a series of modifications of this baseline model, see upper panels of Figures 6, 7, 9, 10 and 11.

Figure 4: Sensitivity and specificity of a model



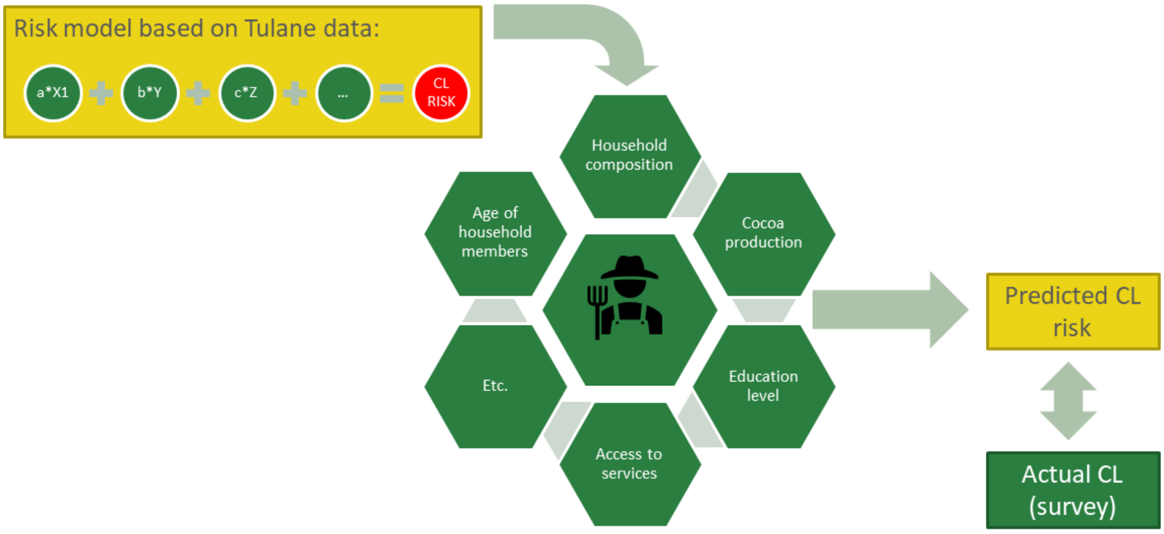
Note on methodology:
A child labour risk model must perform on two objectives: it should correctly flag child labour cases as high risk (sensitivity, finding "true positives"), and correctly mark non-child labour cases as low risk (specificity, finding "true negatives").
The model's sensitivity and specificity are defined by i) setting the model parameters ; and ii) choosing a cut-off point on a risk score between 0 and 1. When building a risk model to be applied for a risk-based targeting mechanism, the optimal choice of the cut-off point will depend on the context and the overarching objective:

- Are we seeking to **capture as many child labour cases as possible**? Then we should aim at **high sensitivity**.
- Are we seeking to **maximize cost savings, by excluding as many non-child labour cases as possible** from the households selected for monitoring or prevention? Then we should aim at **high specificity.**

Hazardous child labour prevalence survey amongst target farmers

In the second stage of the child labour risk model development, a child labour prevalence survey was conducted among a sample of target farmers, who were members of two farmer societies in Asunafo and Suhum districts in Ghana. Both groups are registered as certified suppliers. The survey results allowed to check how precisely the hazardous child labour risk predicted by the model based on Tulane data corresponds to actual hazardous child labour use among this sample of target farmers (as illustrated in Figure 4).

Figure 5: Applying the Tulane risk model to predict hazardous child labour risk among a target group of cocoa farmers



Survey implementation and descriptive results

The survey was implemented by a local consultant with support from ICI for the conception of questionnaires and sampling strategy. Data was collected in June and July 2019.

The sampling phase revealed that even if a cocoa trading company has the clear intent to ensure traceability of their product through their supply chain, the maintenance of an up-to-date register of smallholder producers can be challenging. While the farmer registers used for this project were supposed to be updated on an annual basis, detailed information was missing for many farming households, since it had not yet been collected. When the data collection team conducted the prevalence survey in the field, it also turned out that a large share of farmers for whom data was available were no longer selling cocoa to the same supplier.

> **Lesson learned:** When applying a child labour prediction model to prioritize target farmers in a project, it is important to first check the completeness and accuracy of available data in farmer registers, since the model can *only* predict risk when the complete set of risk indicators is available. If the farmer register turns out to have many missing entries on certain household and farm characteristics, these cannot be considered for inclusion in the model.

The survey was administered to a total of 1,541 children across 705 cocoa producing households from 61 communities in Asunafo and Suhum districts of Ghana. The survey found that **46% of children in the sample had engaged in at least one hazardous activity during the week prior to the survey; and 65% of the children had engaged in a hazardous activity during the 6 months prior to the survey**. Hazardous child labour was more prevalent amongst boys (51.3) than amongst girls (40.8%); and more prevalent amongst older children (36.7% amongst children aged 5-11; 55.9% amongst children aged 12-14; and 59% amongst children aged 15-17).

Breaking of cocoa pods with sharp tools, carrying heavy loads beyond permissible weight, working without basic foot or protective clothing, and using sharp tools for weeding or pruning were the most mentioned activities children in hazardous labour engaged in. These patterns of child labour incidence are in line with findings from other child labour prevalence surveys, such as the Tulane survey, and also with data emerging from Child Labour Monitoring and Remediation System (CLMRS) projects. Figure 5 provides some additional summary statistics on key household and child variables.

Figure 6: Summary of key findings from child labour prevalence survey

	Results from child labour prevalence survey in Asunafo and Suhum districts of Ghana	**Results from Tulane 2013/14 child labour prevalence survey, cocoa-growing households in Ghana**
Average land size under cocoa cultivation	9.3 acres (3.8 ha)	6.4 acres (2.6 ha)
Average amount of cocoa produced per farmer	20 bags (1,300 kg)	12 bags (780 kg)
Average annual income obtained from cocoa sale	GHC 9807.00 (approx. USD 1961.00)	-
Share of households with a LEAP[15] beneficiary	9.6%	-
School enrolment amongst boys aged 5-17	96.6%	94.7%
School enrolment amongst girls aged 5-17	97.8%	94.0%
Share of households with at least one child doing hazardous tasks, 7 day reference period; n=704	45.2%	75.6%
Share of children doing hazardous tasks, 7 day reference period; n=1,541	46.1%	55.1%

A comprehensive documentation of the sample and descriptive results of the survey are provided in the separate report *Prevalence of Child Labour in the cocoa sectors of Asunafo and Suhum Districts of Ghana, prepared* by the consultant and available upon request from ICI.

Applying the risk model to target farmers

By applying the risk model to the target farmers and comparing the predicted risk with the observed hazardous child labour outcome for each child, we can learn whether the child labour use patterns observed in a nationally representative sample of cocoa farmers are transferable to a specific group of target farmers and can be used to predict risk.

To predict hazardous child labour for each child in the target farmer sample, the relevant **household and farm parameters** collected as part of the child labour prevalence survey were **plugged into the risk model equation, to compute for each child a risk score value between 0 (very low risk) and 1 (very high risk)**. Using a cut-off at 0.55 on the risk score,[16] each child was then assigned a hazardous

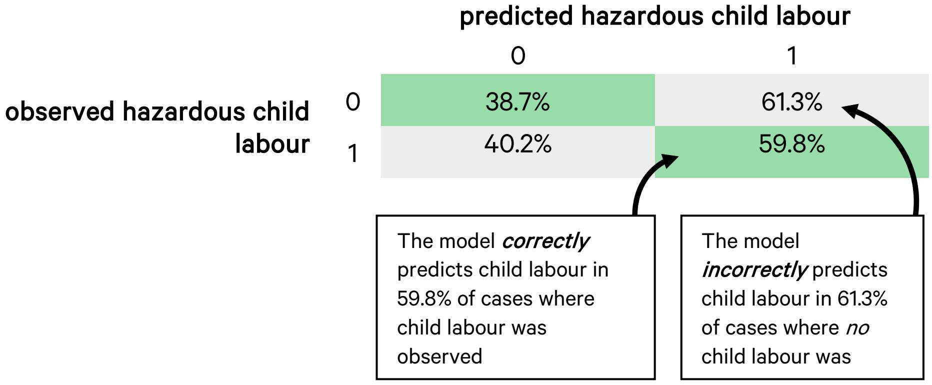
[15] Ghana's Livelihood Empowerment Against Poverty programme, a social protection scheme targeting poor households.
[16] The cut-off value at 0.55 yields, within the Tulane sample, a predicted hazardous child labour rate closest to the observed rate.

child labour prediction indicator, where value 0 predicts that the child *does not* engage in hazardous child labour, and value 1 predicts that the child *does* engage in hazardous child labour.[17] Amongst the target farmers, the predicted hazardous child labour rate was 61%, using the Tulane-based model and a risk score cut-off at 0.55, hence 15 percentage points *higher* than actually observed rate of 46%.

To evaluate the transferability of the model, Figure 7 shows the shares of predicted hazardous child labour against the share of observed hazardous child labour in the target sample.

It turns out that this basic first version of the model, which uses exclusively information available in our reference farmer register, cannot be operationalized for hazardous **child labour prediction**. While the model correctly predicts hazardous child labour for 60% of the actual hazardous child labour cases observed, it also predicts hazardous child labour for 61% of the non- hazardous child labour cases.

Figure 7: Child labour risk model based on parameters available in the supplier farmer register; outcome: hazardous child labour indicator



		predicted hazardous child labour	
		0	1
observed hazardous child labour	0	38.7%	61.3%
	1	40.2%	59.8%

The model *correctly* predicts child labour in 59.8% of cases where child labour was observed

The model *incorrectly* predicts child labour in 61.3% of cases where *no* child labour was

Modifying the risk model to improve prediction power: adding parameters

In order to further improve the model, we test adding different characteristics which are *not* currently collected as part of the target farmer register, but which could easily be added to a standard questionnaire administered in regular intervals to members of a producer group. We tested various parameters, including basic demographic characteristics of the family members, whether children are biological children of the household head, whether the family has migrated, and whether children attend school. Among them, **we identified two parameters that significantly improve the model: the sex and the age of children in a household**. The data show that boys are at higher risk of engaging in hazardous child labour than girls and the risk of hazardous child labour increases with the child's age.

When adding the child's sex to the parameters listed in Figure 6, the area under the ROC (a measure of combined performance of sensitivity and specificity, see section 2.1.3) increases from 0.601 to 0.631. When adding the child's age group, the area

[17] For this exercise, the model had to be slightly adjusted because data on farmers' use of fertilizer and pesticides was not collected as part of the survey, and data from the farmer register, which did contain this information, was available for only a small sub-set of farmers in the final sample. Dropping the fertilizer and pesticides variables from the model resulted in a slightly decreased prediction ability of the model within the Tulane data, with the area under the ROC curve decreasing from 0.634 to 0.601.

under the ROC increases to 0.802; and when adding both the child's sex and age group, it increases to 0.816. **The addition of other easy-to-collect parameters does not help to improve the prediction ability of the model**.

Figure 8 shows results for **a model including the child's sex and age in addition to the other characteristics** in the target farmer data, applying a 0.55 cut-off as before. We can see that this model **correctly predicts 63% of non-** hazardous **child labour cases, and 58% of** hazardous **child labour cases**.

To conclude, **a risk model calibrated on a nationally representative sample of cocoa households and using *only* basic household and farm characteristics does not predict child labour** outcomes with the desired precision. **However, by adding basic demographic information** about individual *children* in the household to the model, **namely their age and sex, the model has a strong enough predictive power to use for risk-based targeting of interventions**.

> **Key recommendation:** These results suggest that **for the purpose of child labour risk estimation, information about the age and sex of children should be included in data collected on producer households**. This information could be collected during the registration of cooperative/farmer group members, in the context of supply chain traceability or certification.

Figure 8: Child labour risk model based on parameters available in the target farmer register, plus child's age and sex; outcome: hazardous child labour indicator

		predicted hazardous child labour	
		0	1
observed hazardous child labour	0	62.6%	37.4%
	1	41.8%	58.3%

Modifying the risk model to improve prediction power: Choosing a more appropriate reference sample

While, at the time of writing, the Tulane data are considered the best available reference data to understand child labour patterns in cocoa production in Ghana, there are various reasons why they may not be representative for the target farmers in this project. First, the Tulane sample represents all six cocoa growing regions of the country, while farmer societies in the target sample operate in only two of these regions: Eastern and Brong Ahafo Region. Second, Ghana has seen rapid economic and social development since 2013/14, when the Tulane data were collected, implying that some of the social and behavioural patterns observed in the Tulane study may have changed over time. Third, as the project targets certified cooperative members only, these farmers have a distinct socio-economic profile (as can be seen from some of the differences noted in Figure 5) which may imply different modalities and trends in their use of hazardous child labour.

All these factors may constrain the ability of a Tulane-calibrated model to predict child labour among the target farmers. While we cannot test what role the time lapse between the two data sets plays, we can investigate whether taking into account the geographic context and the socio-economic profile of farmers could improve the predictive power of the model.

In order to test whether the difference in geographical scope is a key impediment to prediction, we re-calibrated the model on the Eastern region sub-set of the Tulane data (and the Brong Ahafo region, subsequently), and then applied that model to the target farmers in the Eastern region (and the Brong-Ahafo region, respectively). It turns out that ***this geographically adjusted model does not have a better prediction ability than the reference model based on the full sample*** (Figure 7); on the contrary, it performs worse, probably due to the significantly reduced sample size.

In order to test how the **difference in average farmer profiles** across the two data sets may affect prediction performance, we used the difference in farmers' average cocoa production, which was one of the parameters for which we observed a striking difference between Tulane and target farmers (see Figure 5).[18] To do this, we selected from the Tulane data only farmers whose annual cocoa production was above 10 bags, and hence closer to the average production volume of our target farmers, and re-calibrated the model. This model performed equally well, but not better than the reference model.

> These tests suggest that neither differences in geographical scope, nor in production volume, appear to be key obstacles to predicting hazardous child labour risk from a given prevalence survey to a specific target farmer group.

Addendum: Modifying the risk model to improve prediction power by using multi-level regression models

From more recent trials of child labour prediction based on regression models using readily available child labour data, ICI has found that **prediction accuracy can be improved considerably by using multilevel regression models**. These models account for *nesting* of data within groups, for example children nested within households, or households nested within communities.

To test this approach here, we replace the logit regression model for predicting a binary child labour outcome with a multilevel logit model which accounts for the nesting of children within households, while using the same predictors and cut-off as in the reference model. **The multilevel model** indeed **performs better** than the basic logit model, **correctly predicting 72% of the non- hazardous child labour cases, and 71% of the hazardous child labour cases** (area under the ROC increases to 0.83).

Modifying the risk model to improve prediction power: Using measures of child labour severity

In the prediction model constructed above, we considered only whether or not a child engages in any type of hazardous work, ignoring any nuance between 'severe'

[18] We cannot test cooperative membership directly, since the Tulane sample contains only very few farmers who declare they sell cocoa to a cooperative.

and 'less severe' cases of child labour.[19] One may argue that a more powerful model should be able to distinguish between a child who works intensely, is exposed to multiple hazards and prevented from going to school, and a child who does some type of hazardous work once in a while.

We therefore test whether the child labour risk model can be refined by using as the outcome different measures of severity of child labour, rather than a binary hazardous child labour indicator. We apply three different measures of *child labour severity* that can be constructed from both the Tulane data and the child labour prevalence data from the target farmers:

- a binary indicator on whether or not a child engages in hazardous tasks and *also* has to compromise on schooling (either not going to school, reporting that schooling is affected by work, or reporting to have worked on school days during the reference week)
- the number of different hazards a child is exposed to[20]
- the total reported number of hours worked during the reference week.

Summary statistics for these measures from the two survey samples are provided below.

Figure 9: Summary of indicators of child labour severity

	Tulane Ghana data	Target farmer sample
Hazardous child labour plus schooling affected	4.67%	4.98%
Average number of hazards a child is exposed to (incl. zero hazards for non- hazardous child labourers)	1.55	1.35
Total number hours worked during last week (incl. zero hours for non- hazardous child labourers)	2.45 hours	2.97 hours

The share of children who engage in **hazardous child labour *and* compromise their education** is around 5 percent in each of the two samples. We calibrate the model using the same risk parameters as before, including the child's age and sex, but replace hazardous child labour by *hazardous child labour plus schooling compromised* as the outcome, and apply a risk score cut-off at 0.05 (reflecting the actual incidence).

We find that amongst children of the target farmers this model correctly predicts 55% of the incidences, and 66% of the non-incidences (see Figure 9). **Overall, the**

[19] The ILO Convention 182 states that hazardous work is considered as "*work which, by its nature or the circumstances in which it is carried out, is likely to harm the health, safety or morals of children" [...] The types of work referred to under Article 3(d) shall be determined by national laws or regulations* » . This means that each state party government determines its own list of hazardous activities and while under the law, they all fall under the category hazardous work" or "worst form of child labour", it is important to recognise that some could potentially be described as more harmful to the child (e.g. spraying pesticides in Ghana law) than others (e.g. not wearing gum boots on a farm) in Ghana law).

[20] The list of different hazards that can be derived from the Tulane data set does not correspond precisely to the list of hazards available in the child labour prevalence data. In order to construct a measure that is comparable across the two surveys, we use the standardized version of the respective variable in each data set (subtracting the mean and dividing by the standard deviation).

performance of this model is slightly poorer than the reference model, which used only engagement in hazardous tasks as an outcome.

Figure 10: Child labour risk model based on parameters available in target sample farmer register, plus child's age and sex; outcome: indicator for hazardous child labour *and* schooling compromised

		predicted hazardous child labour *and* schooling compromised	
		0	1
observed hazardous child labour	0	65.9%	34.1%
	1	44.9%	55.1%

Second, we test a model which uses as an outcome the **number of different hazards a child is exposed to** (standardized for comparability across the data sets). Since we no longer work with a binary hazardous child labour indicator, but with a continuous measure of child labour severity, we adjust the modelling procedure by estimating a *linear* rather than a *logistical* regression model in the first stage, but using the same set of risk parameters as before (including child's age and sex).

In order to evaluate this model against the benchmarks used above, we create an indicator variable which takes the value one if a child is exposed to more than the average number of hazards within the sample (which is 1.55), and zero if a child is exposed to less than the average number of hazards. This also corresponds to the risk-based targeting logic, which requires a mechanism for flagging higher-risk cases.

This model now performs very similarly to the reference model, as shown in Figure 10. When used to predict cases of high exposure to hazards among children in the target farmer sample, it correctly flags 59% of the high-exposure cases, and correctly marks 62% of the low-exposure cases as such. Again, the performance of this model is slightly poorer than the reference model.

Figure 11: Child labour risk model based on parameters available in target sample farmer register, plus child's age and sex; outcome: number of hazards to which a child is exposed

		predicted exposure to a *higher than average number of hazards*	
		0	1
observed hazardous child labour	0	61.6%	38.4%
	1	40.1%	59.0%

Finally, we test a model which uses as an outcome the **total number of hours a child has worked during the reference week**. Again, this is a continuous measure of child labour intensity, and we therefore apply a similar procedure as above to construct the prediction model: we use the same risk parameters (including child's age and sex), a linear regression model, and apply as the cut-off a value of the 2.5 hours, the mean number of hours worked in the Tulane sample.

This model slightly outperforms the reference model in terms of its prediction ability on the target sample farmers (see Figure 11). Amongst children in the target farmer households, **this model correctly flags 64% of the ‘high-intensity’ child labour cases, and also correctly marks 64% of the ‘low-intensity’ child labour cases** as such.

Figure 12: Child labour risk model based on parameters available in the target farmer sample, plus child’s age and sex; outcome: number of hours worked last week

		predicted exposure to hazardous child labour for *more than 2.5 hours per week*	
		0	1
observed hazardous child labour	0	63.6%	36.9%
	1	35.6%	64.4%

To conclude this section, **an effective child labour prediction model can be built using either a binary indicator of hazardous child labour or using measures which better reflect nuances between severe and less severe cases** of child labour.

> Of the child labour measures tested as outcomes of the prediction model, the number of hours a child has worked during the last week proves to be a useful alternative measure. **A model using the number of hours worked per week as the outcome can predict a higher share of high-intensity cases,** compared to a model based on a binary indicator for hazardous child labour.

COST SAVING POTENTIAL OF A RISK-BASED TARGETING MECHANISM

The child labour risk model we have developed allows to improve the targeting of interventions, by channelling more resources to households most at risk of using hazardous child labour. The benefits of improved targeting can be measured in terms of number of hazardous child labour cases reached with a given amount of resources, or in terms of cost saved for reaching a given number of hazardous child labour cases.

In order to quantify the benefits of applying a risk-based targeting mechanism, we need a good understanding of the predictive power of the risk model, as discussed in the previous sections, and a reference intervention with a well-defined cost structure. In order to illustrate the benefits of a risk-based targeting approach, we present here some calculations of cost saving potential for the risk model described above in Figure 8.

This model includes basic household and farm parameters, plus child’s age and sex, and uses a binary indicator for hazardous child labour as outcome. The reference intervention we consider here is a Child Labour Monitoring and Remediation System (CLMRS), where one costly element of the intervention is a household or farm visit to every farmer in a producer group, in order to identify cases of hazardous child labour before they can be addressed. If these monitoring visits could be focussed on only those households at higher risk, a given number of visits could identify a higher number of cases of hazardous child labour. As a result, **either a higher share of the project funds could be available to support identified cases of** hazardous **child labour, or a larger number of farmers could be covered by the system.**

To calculate potential cost savings, we assume that among cocoa farmers covered by a CLMRS, the actual hazardous child labour rate is 46% (as in the target farmer sample), and that a monitoring visit would detect a case of hazardous child labour if it were present. Then, if every household is visited independent of the household's predicted hazardous child labour risk, 1000 children would need to be visited to identify 460 child labour cases. **If instead the monitoring were to follow a risk-based approach**, where *only* children flagged as high risk are visited (children for whom the model had predicted incidence of hazardous child labour, using a 0.55 cut-off on the risk scale), **800 children would need to be visited, rather than 1000, to identify the same number of child labourers.**[21]

If one hazardous child labour case corresponded to one household – another simplifying assumption for illustrative purposes –[22] costs for 200 household visits (or 20% of the monitoring cost) could be saved, as shown in the grey column on Figure 13, below.

The table shows the outcomes of several different scenarios, each with a different cut-off point. As above, we assume a target group of 1000 children to be covered by a CLMRS (assuming for simplicity that each child is living in one household); that within the target group, the true hazardous child labour rate is 46%; and that a household visit would detect hazardous child labour, if it were present.

Figure 13: Cost saving potential for a risk-based child labour monitoring system, using different cut off points to mark a child as "high-risk"

		Cut-off point used to mark a "*high-risk*" child:						
		0.15	**0.25**	**0.35**	**0.45**	**0.55**	**0.65**	**0.75**
A	Within 1000 targeted children, # of *high-risk* cases to be visited	960	803	666	530	470	430	370
B	Amongst children flagged *high risk*, rate of actual hazardous child labour cases	47%	49%	53%	56%	58%	60%	61%
C	# of hazardous child labour cases identified	451	396	350	296	270	257	226
D	Within 1000 targeted children, number of *low-risk* cases *not to be visited*	40	197	334	470	530	570	630
E	Share of hazardous child labour cases among the *low-risk* cases *not to be visited*	26%	35%	34%	36%	37%	36%	38%
F	# of hazardous child labour cases *not* identified (missed)	10	69	115	168	195	207	239
G	# of visits needed to find an equal # of hazardous child labour cases *without* risk-based targeting	981	861	762	644	588	559	491
H	Relative cost saving achieved through risk-based targeting	2%	7%	13%	18%	20%	23%	25%

[21] The share of child labour cases in the high-risk group would be 57.5%; 460/0.575=800.

[22] Given that several of the risk factors entering the model are at household level, the predicted child labour risk will not be distributed equally across households. Rather, children at higher risk of child labour can be expected to be concentrated within a smaller number of households. The cost saving estimates presented here are therefore likely underestimate the true cost saving potential of using such a model.

Using the lowest cut-off point (0.15), 96% of children are marked as *high-risk.* This scenario identifies the most hazardous child labour cases and misses the least hazardous child labour cases but results in the smallest cost saving. In contrast, using the highest cut-off point (0.75), 37% of children are marked as *high risk.* This scenario identifies the least hazardous child labour cases and misses the most – in fact it misses more hazardous child labour cases than it identifies – but results in the largest cost saving.

Depending on the cut-off point selected, row A shows the number of children that would be flagged as *high-risk* and receive a visit; rows B and C show the rate and number of true hazardous child labour cases that would be detected through these visits. Row D shows the number of children that would be flagged as *low-risk* and therefore not receive a visit; rows E and F show the rate and number of true hazardous child labour cases that would be "missed" by the risk-based approach and would therefore go undetected.

The balance between detected cases of hazardous child labour (B and C) and non-detected cases (E & F) will need to be weighed against the cost saving realized through applying the risk model (row H). The cost saving is determined based on comparisons with the number of household visits needed to determine the same number of hazardous child labour cases in a conventional system (row G).

> **The use of a predictive model to target only high-risk households, will always involve some trade-offs** between hazardous **child labour cases identified**, hazardous **child labour cases "missed"**, and **cost savings made**. All these will depend on where we set the "high-risk" cut-off point.

In practice, many other factors need to be considered alongside any potential cost savings. For example, what level of exclusion error are stakeholders willing to tolerate? How does the physical proximity of households affect the cost savings related to their inclusion or exclusion? And could the exclusion of some households lead to tensions within cooperatives or communities?

CONCLUSION

This study describes the development of a child labour risk model, using nationally representative data on child labour in Ghana. It demonstrates that information commonly available in farmer registers, such as those held by cooperatives, can be used to predict the likelihood that a child living in a cocoa-farming household engages in hazardous child labour. However, the model can only correctly predict the presence or absence of hazardous child labour in most cases when information on the sex and age of children in farming households is available. Where farmer registers do not include this information, or if farmer records are incomplete, then the model cannot be used.

If such a risk-prediction model were to be operationalised in the context of a Child Labour Monitoring and Remediation System, for example to identify households to prioritise for monitoring and support, it could lead to a potential cost saving of around 20%. Its use could also mean that vulnerable households and children are identified more quickly, allowing them to receive assistance faster.

However, the use of a risk-prediction model has limitations, which need to be clearly understood when deciding to use such a model. Stakeholders responsible would need to be aware that any model will *incorrectly* predict the presence or absence of child labour in some cases, meaning that some vulnerable households will not be identified and may be excluded from support programmes. The selection of the cut-off point used to define a "high-risk" household must be done with these limitations in mind.

This paper summarises one possible approach to constructing and testing a child labour risk model for Ghana. As more recent data on child labour becomes available, these data could be used to re-calibrate the risk model, potentially leading to an improvement in its predictive power.

The same approach could equally be used to develop child labour risk models specific to other smallholder agriculture contexts, so long as basic demographic information on farming households is readily available, and reliable datasets on child labour within the same context exist.

Addendum, May 2021

Since the publication of this report, ICI has continued to develop child labour risk models in other operational contexts and has tested different statistical methods for child labour prediction. From these more recent trials, we have learned that prediction accuracy can be improved considerably by using multilevel regression models, which better account for nesting of data within groups. In the context of the present study, replacing a basic logit model by a multi-level logit which accounts for nesting of children within households has increased the rate of correctly predicted cases of child labour from around 63% to 72%. This would imply a considerable additional cost saving potential if used to inform a risk-based CLMRS.