RESEARCH REPORT

Risk models for predicting child labour

A review of different approaches to identifying children at risk of child labour in cocoa

September 2021

International COCOA Initiative

Protecting children and their families in cocoa-growing communities

The International Cocoa Initiative is a non-profit membership organisation dedicated to improving the lives of children and adults in cocoa-growing communities. We are experts in child labour and forced labour in cocoa, advising governments and corporations in order to inform their practices and influence decision-making, as well as working with NGOs in the field. We are committed to supporting the development of sustainable cocoa production that protects the rights of children and adults worldwide.

SUPPORTED BY:

giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

This study was conceived and developed by the International Cocoa Initiative (ICI) and was written by Laurent Foubert. It was supported by contributions from ICI's members and the *Deutsche Gesellschaft für Internationale Zusammenarbeit* (GIZ).

We are grateful to the following partners, who shared information about projects to model child labour risk and agreed for them to be included as case studies in this report: Cargill, ECOM, Nestlé and Tony's Chocolonely.

Cover image: awareness-raising session on child labour in Didoko, Côte d'Ivoire, ©Nestle, 2019.

This report was updated in May 2022.

www.cocoainitiative.org | info@cocoainitiative.org

ICI Secretariat in Switzerland
Chemin de Balexert 9,
1219 Châtelaine | Switzerland
+41 22 341 47 25

ICI National Office in Côte d'Ivoire
II Plateaux, 7ème Tranche, Lot 3244, Ilot 264,
Abidjan-Cocody | Côte d'Ivoire
+225 27 22 52 70 97

ICI National Office in Ghana
No. 16, Djanie Ashie Street,
East-Legon | Accra | Ghana
+233 302 998 870

Contents

Executive summary

Considerable progress has been made in recent years towards the creation of risk models capable of predicting child labour at the household or individual level. This report describes different approaches used to develop and test child labour risk models in the context of cocoa-growing areas of West Africa. These models aim to improve the way in which interventions to prevent and address child labour are targeted to where they are most needed, as part of a broader effort aim to scale up efforts to protect children from child labour, as well as to ensure access to their fundamental rights.

This study addresses the following questions:

1. What are the characteristics of the risk models that have been developed to date within the cocoa sector to predict child labour and how do they perform?
2. What has been learned from these experiences and what recommendations have emerged?

What is a risk model?

A risk model is a statistical approach aimed at **predicting an outcome** (e.g. child labour) **for a given unit of observation** (e.g. a child or a household) **from a set of predictors.**

The first step involves "calibrating" the model by applying statistical methods (e.g. logistic regression, multilevel regression, latent class model, etc.) to a data set containing information about the outcome and the predictors for a population similar to the target population. The second step then consists of feeding the model with information about the predictors observed from the target population, in order to obtain a predicted outcome for each unit. If the predicted outcome is a binary indicator (e.g. whether a child engages in child labour), the prediction can be interpreted as a "risk" (or likelihood), that the incident will occur.

Methodology

This report draws on six projects implemented by ICI, its members and other cocoa-sector stakeholders, who use risk models to predict the child labour status of children or households in Côte d'Ivoire and Ghana. Some of these projects were developed within the framework of operational trials, in order to increase the cost efficiency and scalability of child labour monitoring and remediation activities, while others were developed in order to generate theoretical knowledge about the potential of child labour risk modelling. The models use diverse data sources, including national

child labour prevalence surveys,[1] farmer registers kept by cooperatives and data collected by Child Labour Monitoring and Remediation Systems (CLMRS). They also apply different statistical methods, including machine learning and regression-based prediction.

On the basis of all the projects reviewed, we highlight key learnings and propose recommendations for the development of efficient risk models and the rolling out of risk-based approaches. For each of the six approaches reviewed in this report, we also provide information about the specific context of the project, its aims, the method used to predict child labour and the models' performance.

Results

The case studies presented here show that the creation of a good risk model starts with good data, that is, accurate, reliable and up-to-date information. Effective data collection tools and well-trained data collectors are essential, both in order to collect the initial data needed to develop the models and then to test whether the predictions generated reflect the situation on the ground. Without high-quality data, even the "perfect statistical method" will never suffice to obtain a highly efficient model. This means that before a risk model can be developed and rolled out at scale, it may be necessary to first strengthen capacities for data collection and management, especially in cases where data is collected by farmers, community members or cooperatives, given that data collection is not their main occupation.

The case studies presented here show that a good risk model starts with good data: accurate, reliable and up-to-date information.

The development of a child labour risk model should not turn into a quest for the "perfect set of predictive variables". A wide range of different variables can be used to effectively predict child labour risk, depending on the specific context and the availability of data.

It is important to note that effective risk models may use factors that have *no causal link* to child labour (such as gender) and that *may have no statistically significant relationship* to child labour prevalence. That said, many risk models do include some factors that *are* considered among the root causes of child labour. Examples of these factors include access to adult labour, clean water and quality education, all of which are associated with lower prevalence of child labour.

The case studies presented in this paper demonstrate that risk modelling should take a child-centred approach. As the examples show, the availability of basic information about the child and their household always improves the ability of a model to accurately predict child labour.

Beyond these considerations, the usefulness of risk-based approaches for a given actor will depend on the strategy, context and constraints they face. The case studies presented in this paper demonstrate that risk models *can* be used to reduce the number of initial monitoring visits conducted through a child labour monitoring and remediation system, as well as to prioritise at-risk households

[1] See, for example: "Assessing Progress in Reducing Child Labor in Cocoa Production in Cocoa Growing Areas of Côte d'Ivoire and Ghana", Sadhu, S., Kysia, K., Onyango, L., Zinnes, C.F., Lord, S., Monnard, A. and Arellano, NORC at the University of Chicago, 2020. This study is referred to as the "NORC survey" elsewhere in this report.

for preventative support. However, the cost-savings apply *only* to the initial monitoring visit, not to the rest of the system implementation. There are also other possible use cases – for example to increase the number of households or children targeted to receive support.

In other words, the findings of this *paper should not be interpreted as justifying a reduction of the resources allocated to child labour monitoring and remediation*, but rather as an *opportunity to use available resources more efficiently*, and notably to reallocate resources to the provision of support to prevent and remediate child labour, as well as to follow up with children identified at risk.

For every situation in which a child labour risk model is used, care must be taken to determine who the model targets for inclusion or exclusion, which strategies are in place for individuals or households *not* considered to be at risk, how often risk assessment should take place and how the effectiveness of the models being used can be continuously assessed and improved.

Given that there is no *one-size-fits-all* approach to the design and use of child labour risk models, it is especially important that all stakeholders continue to share information about their approaches, results and conclusions in this field, thus enabling others to build upon their efforts, with the broader aim of scaling up activities to address and prevent child labour.

How can risk models be used?

Risk models can be used for a variety of different purposes. The case studies in this report showcase risk models developed for the following reasons:

- To identify households at higher risk of using child labour (e.g. among members of a cooperative or a community), so that they can be prioritised for monitoring visits or support
- To identify cooperatives or communities at higher risk of child labour, so that interventions can be targeted to areas at greatest risk
- To broaden the number of households targeted to receive support, by identifying additional households “at-risk” of child labour, as well as those where child labour has been identified
- To bolster routine child labour monitoring efforts – for example, in the context of a CLMRS – to increase the chance that cases of child labour are identified (e.g. due to the cyclicity of child labour and the possible reallocation of work among children of the same household, in response to shocks).

Practical considerations developing and using child labour risk models

- To develop and calibrate a risk model, **at least one existing data set is required**, which includes the **outcome of interest** (e.g. whether or not households use child labour) and **all the variables to be used as predictors** (e.g. detailed information about households and their members), taken from a sample which is comparable to or overlaps with the target population.
- To use a model to predict the risk of child labour, **data on the target population must be complete and accurate**. Risk-based approaches cannot be used to assess risk among individuals/households for whom data is incomplete or inaccurate.
- **Data management and statistical analysis capacity is required**, either in-house or external, in order to calibrate and fine-tune a risk model, so as to make it responsive to the needs and operational constraints of each implementer.

Learnings and recommendations

Learnings

- **High-performing models can reduce the number of households targeted for initial monitoring visits**, while still **ensuring most children in child labour are reached**: if applied in the context of household-level child labour monitoring – and depending on the performance of the model – a risk-based approach may enable a 50% decrease in the number of initial monitoring visits, while identifying more than 95% of the children in child labour. This would reduce the *initial* cost of monitoring (but not the rest of system implementation) and allow resources to be reallocated to support and follow-up, helping children get assistance faster.
- **There is no one-size-fits-all model**: risk models can and should be tailored to the context of use, to the available data and to operational constraints.

- **Quality data collected from the target population is essential for generating useful predictions**: measurement errors and outdated data reduce the accuracy of the model, whatever the statistical method, while missing data excludes individuals or households from risk-based coverage. For this reason, strengthening the data collection and data management capacities of local stakeholders may be a prerequisite before a risk-based approach can be used.
- **Information about individual children, such as their age and sex, improves the quality of predictions**: incorporating the age, sex and school status of the child systematically improves the accuracy of models to predict child labour.
- **Child labour status should be predicted at the *household level*, rather than at the child level**: this approach enables cases of child labour to be identified more efficiently, as well as making it easier to plan monitoring visits.
- **Recurring (annual) assessment of child labour risk using a predictive model is likely to be more effective than one-off activity**: every risk model entails a certain level of error, meaning that some children in hazardous child labour will be missed. High-performing risk models that include the age of the child among their predictors seem to have the capacity to eventually identify cases that had been missed earlier, through successive waves of assessment. Thus, assessing the risk for a target population on an annual basis using the most up-to-date data could not only allow children aged 5–17 who were previously absent from the data to be taken into account, but it could also lead to the identification of children who were previously missed, since their risk score changes as they get older.
- **The cost-saving potential of a risk-based approach depends on several factors**: these factors include the prevalence of child labour, the model's performance and the effectiveness of support measures. In locations where child labour prevalence is higher, the potential cost savings of using a risk model are lower, but may be improved over time if support is efficiently targeted, reducing the prevalence rate and increasing potential savings.

Recommendations

Before considering the deployment of a risk-based approach:

- **Clearly define the aim of the risk model** – for example, faster identification of cases or reduced monitoring visits – **and the operational constraints** and develop the model accordingly. Predictive models will always carry a decree of uncertainty; even the best model will "miss" a certain share of cases. This aspect needs to be clearly understood and mitigation measures put in place to manage the operational and ethical consequences of the choices made.
- **Assess the technical capacity and time available** for developing a model. An effective data management system is needed, as well as skilled staff capable of running it and regularly updating the predictive model.
- **Assess the availability of recent, complete and high-quality data**. If one of these three requirements is not met, do not use this data to develop a risk-model – instead focus efforts on getting quality data first. If the approach is to be embedded in local structures (farmer groups, cooperatives, etc.), assess their data collection and management practices. Without the capacity to collect and maintain complete and accurate data sets, there is no point in developing a risk model.

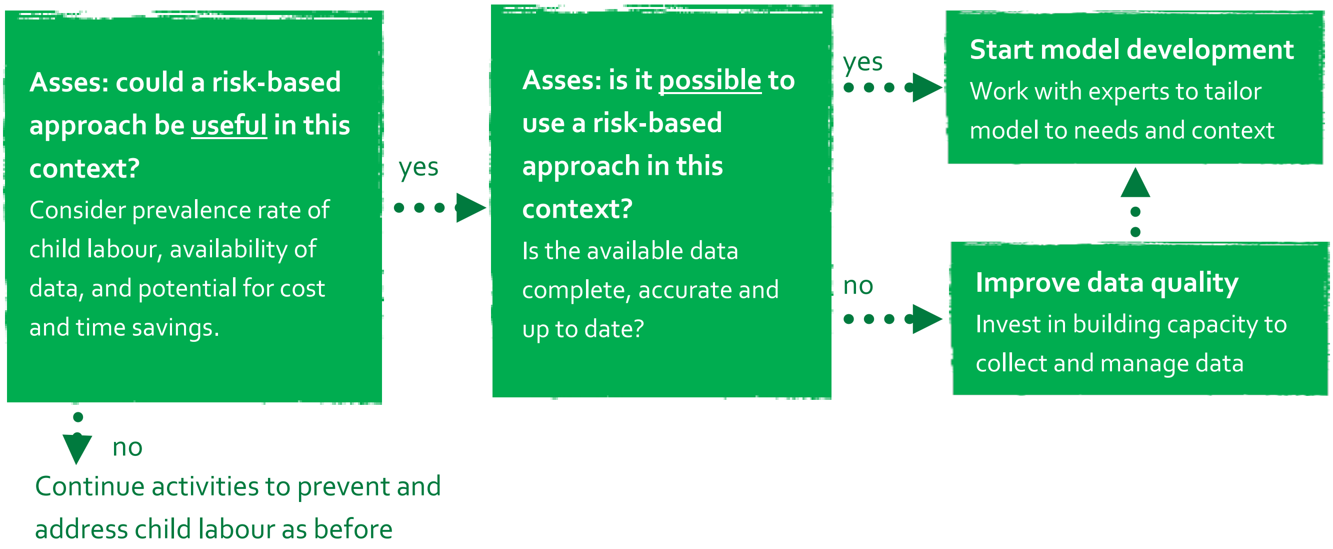
During the development of a risk model:

- **Focus on easy-to-collect, easy-to-assess indicators**, and ensure that every effort has been made to reduce measurement errors and missing data.
- **Limit the number of predictors** used by the model to about 10–12,[2] in order to facilitate data collection and access to high-quality data.
- **Use reliable data to calibrate the model**, ideally from a large sample whose coverage overlaps as much as possible with the geographic area of the target population.
- **Incorporate basic child-level predictors (e.g. age, gender and school status)** into the risk model to improve performance.
- **Set up the model to predict the risk for the *household*, rather than for the child**, as this improves a model's ability to identify cases of child labour, as well as being more practical on the operational level.
- Where applicable**, carefully choose a cut-off value** on the basis of its distribution, local context and prior knowledge (e.g. national prevalence rate), and operational strategy and constraints.

When the model is **in use**:

- Where possible, **run the risk model annually on a recurring basis**, using up-to-date data, so that it can, over time, identify all children at risk.
- **Constantly assess a risk model's performance** against known prevalence rates and the results of monitoring visits and adjust the model if identification rates are lower than expected.

Figure 1: Decision model for risk model development:


Asses: could a risk-based approach be useful in this context?
Consider prevalence rate of child labour, availability of data, and potential for cost and time savings.
yes
Asses: is it possible to use a risk-based approach in this context?
Is the available data complete, accurate and up to date?
yes
Start model development
Work with experts to tailor model to needs and context
no
Improve data quality
Invest in building capacity to collect and manage data
no
Continue activities to prevent and address child labour as before

[2] The high-performing risk models described in this report use between 10 and 12 predictors. There is no evidence that increasing the number of predictors improves the models.

Introduction

In 2018–2019, 1.56 million children were estimated to be in child labour in cocoa production in Côte d'Ivoire and Ghana. In Côte d'Ivoire, 55% of cocoa-growing households had at least one case of child labour, compared to 69% in Ghana.[3] These figures show clearly that child labour is a widespread challenge, but also that some households are more vulnerable than others.

Effective approaches to preventing and addressing child labour exist. but reach only a fraction of households who need them. Over three years, a comprehensive package of interventions – including community development programmes and Child Labour Monitoring and Remediation Systems (CLMRS) – was shown to reduce the community prevalence of child labour in cocoa-growing communities by one-third.[4]

By the end of 2020, between 10-20% of cocoa-growing households in West-Africa were estimated to be covered by any kind of CLMRS.[5] Many stakeholders have recently pledged to scale up coverage, with the aim of ensuring that 100% of households in cocoa-growing areas of Côte d'Ivoire and Ghana are covered by effective systems that prevent and address child labour within the next five years.[6]

In this context, where time and resources are limited, risk-based approaches represent a promising means of scaling up the system. Effective risk models could help to identify vulnerable children more efficiently, enabling assistance to be targeted to where it is needed most, more quickly. This report describes the key steps involved in developing models to predict child labour and summarises the practical experience gathered while putting them into practice. It also includes several examples focusing on the use of risk-based approaches in the context of Child Labour Monitoring and Remediation Systems (CLMRS).

The following questions will be addressed:

- What is a risk model?
- What are the characteristics of the models developed to date within the cocoa sector for predicting child labour[7] and how do they perform?
- What has been learned from these experiences and what are the emerging recommendations?

[3] "Assessment of Effectiveness of Cocoa Industry Interventions in Reducing Child Labor in Cocoa Growing Areas of Côte d'Ivoire and Ghana.", Sadhu, S., Kysia, K., Onyango, L., Zinnes, C.F., Lord, S., Monnard, A. and Arellano, NORC at the University of Chicago, 2020.

[4] Ibid.

[5] ICI Strategy 2021–2026, International Cocoa Initiative, 2020.

[6] Ibid.

[7] Note that throughout this paper the general term "child labour" is used, although in several cases, the risk models discussed were developed to identify children in *hazardous child labour*, a subset of all child labour cases.

This paper draws on six projects that use risk models to predict the child labour status of children or households in Côte d'Ivoire and Ghana. Some of these projects were developed as part of operational trials aimed at increasing the cost efficiency and scalability of child labour monitoring and remediation activities, while others were conducted with a view to generating theoretical knowledge about risk models to predict child labour and their potential application. The models have been developed using diverse data sources, including national child labour prevalence surveys, farmer registers kept by cooperatives and data collected by Child Labour Monitoring and Remediation Systems (CLMRS). They also apply different statistical methods, including machine learning and regression-based prediction.

The first section lays out general considerations for the use of models to predict child labour risk. The second section presents the learnings drawn from the six case studies, highlighting key lessons and proposing recommendations for developing effective models to predict child labour, as well as for rolling out risk-based approaches in practice. In the third section, we provide a comprehensive description of each of the six case studies. For each case study, we report the specific context of the project, the method used to predict child labour, the model's performance, and finally the key learnings.

What is a risk model?

Definition

A risk model is a statistical approach aimed at **predicting an outcome of interest** (e.g. child labour) **from a set of predictors for a given unit of observation** (e.g. a child or a household)**.**

For example, a model could assess *the likelihood that a child engages in hazardous child labour*, using the following predictors: *the presence/absence of a primary school, health facilities and a paved road in the community; the level of education and the gender of the head of household; the number of different crops cultivated on the farm; and the age, gender and school status of the child*. The weight of each predictor is then determined by means of a **statistical method**.

In this process, a first data set is used to allow the model to "learn" about how the outcomes vary depending on the predictors (the "calibration" data set). Then, another set of data is used to test the accuracy of the model when making predictions outside the calibration data set. Finally, the model is used to predict the outcome of interest for each unit of observation in the target population.

If the predicted outcome is a binary indicator (e.g. whether a child engages in child labour), the **prediction can be interpreted as a "risk" (or likelihood) that the incident will occur**.

It is very important to bear in mind that **prediction statistics do not draw causal links** between the outcome of interest and the predictors. Rather, they build on correlations between these two elements, meaning that no causal pathway between them is needed, even if it is possible that one exists. Hence, **risk models should not be interpreted in causal terms**.

The main components of a risk model are, therefore, the *outcome of interest*, the *predictors*, *unit of observation* and *prediction* (e.g. the child or the household), and the *statistical method* (logistic regression, multilevel regression, machine learning, etc.).

Risk modelling step by step

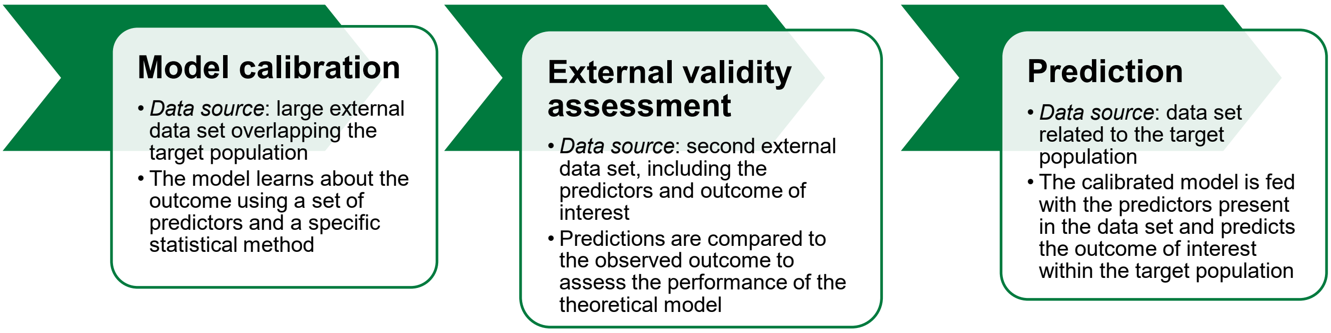
The statistical development of a risk-based approach comprises seven steps:

1. **Selection of a reference data set**: insofar as possible, this should be a high-quality and nationally/regionally representative data set containing observations of all the predictors and outcome of interest from a (large) sample that is considered comparable to the target population,[8] as well as variables which are identical to the variables in the data set in which the risk is to be predicted.
2. **Selection of the predictors**: these are variables common to both the reference data set and the prediction data set which are used to calibrate the model and to make the prediction within the target population. In both datasets, the predictors must be the same type (binary, continuous etc.) and defined in the exact same way.
3. **Calibration**: the reference data set is used to calibrate or “train” the model. At this stage, the statistical method used creates all the parameters enabling the outcome to be predicted from the predictors.
4. **Within-sample validation**: a predicted outcome of interest is calculated by plugging values of the predictor variables into the model. This predicted outcome is then compared with the actual/observed outcome in the sample used for calibration, in order to determine the accuracy of the model (sensitivity, specificity).
5. **Out-of-sample validation**: using a second “validation” data set,[9] which contains observations on the predictors and the outcome of interest, a predicted outcome is calculated by plugging the values of predictor variables into the model. This predicted outcome is then compared to the actual/observed outcome in order to determine the accuracy of the model (sensitivity, specificity) within this validation data set.
6. **Application of the model to the target population**: a data set is available for the target population which contains observations of the predictors, but not of the outcome of interest. The values of predictor variables are plugged into the calibrated model, in order to predict the outcome for each unit (e.g. child or household) in the target population.
7. **Verification/evaluation of the model’s predictive performance**: data on the actual outcome is compared to the predicted outcomes and used to evaluate the model’s performance.

[8] This data set could be a national survey (e.g. the NORC child labour prevalence surveys) or another data set including the same characteristics as the population in which the risk is to be predicted, ideally from the same geographical area.

[9] This second data set may be drawn from a segment of the same population for which the outcome of interest is to be predicted or from another population with similar characteristics. An alternative approach is to split the “calibration” data set into two equal parts, covering different geographic areas or supply chains. One part is then used for calibration and internal validation and the other for a first assessment of the “external” validity of the predictions.

Figure 2: Main steps in the development of a risk model



Model calibration

- *Data source*: large external data set overlapping the target population
- The model learns about the outcome using a set of predictors and a specific statistical method

External validity assessment

- *Data source*: second external data set, including the predictors and outcome of interest
- Predictions are compared to the observed outcome to assess the performance of the theoretical model

Prediction

- *Data source*: data set related to the target population
- The calibrated model is fed with the predictors present in the data set and predicts the outcome of interest within the target population

Output and performance of a risk model

If the outcome of interest is a binary variable (e.g. whether a child engages in child labour), a regression-based risk model generates a predicted value **as a continuous number** ranging between zero and one, where values closer to one indicate a higher "risk" that the observed outcome equals one (e.g. a child in child labour). This value can then be **either transformed back into a binary prediction** (e.g. a prediction that the child is *in child labour* or *not in child labour)* **or into any other categorical variable**[10] (for example, *low*, *medium* or *high* predicted risk of being in child labour), whereby **cut-off values** (see Annex 1 for more details) are defined *according to the purpose of the model.*

By contrast, models based on machine-learning methods generate the predicted outcome as a value of a binary variable (e.g. they would predict that a child is either *in* or *not in* child labour), so the step of transforming a continuous risk value into categories is not applicable for such models.

The **predictions produced by a risk model will always have some degree of uncertainty. The level of uncertainty (or error rate) reflects the performance of the model. This should be assessed and communicated, alongside the model's output**.
In models that produce a binary output (*yes* or *no*), predicted cases will fall into one of four possible categories:

- **true positives** (model predicts "yes" and the true outcome is "yes")
- **false positives** (model predicts "yes" but the true outcome is "no")
- **true negatives** (model predicts "no" and the true outcome is "no")
- **false negatives** (model predicts "no" but the true outcome is "yes")

Figure 3: Defining and using a **risk model always involves a trade-off** between sensitivity and specificity

		Predicted outcome	
		Positive (e.g. *in* child labour)	**Negative** (e.g. *not in* child labour)
Observed outcome	**Positive** (e.g. *in* child labour)	**True positives** *(Sensitivity)*	**False negatives**
	Negative (e.g. *not in* child labour)	**False positives**	**True negatives** *(Specificity)*

[10] All the risk-based approaches reviewed here use binary prediction.

Sensitivity and specificity are two key concepts for assessing the **accuracy of a model**:

1. **Sensitivity** is the **rate of true positive cases predicted by the model** (i.e. excluding false positives), expressed as a percentage.
2. **Specificity** is the **rate of true negative cases predicted by the model** (i.e. excluding false negatives), expressed as a percentage.

The balance between sensitivity and specificity is determined by the performance of the model and the chosen cut-off. When the outcome of interest is binary, this defines the level *above which* predicted scores are to be considered positive and *below which* predicted scores are to be considered negative.

The balance between the sensitivity and specificity of a prediction model is reflected in a "Receiver Operating Characteristic" **(ROC) curve**.[11] For all the prediction models discussed in this paper, we report the area under the ROC curve, referred to as the "ROC area", along with the sensitivity and specificity, as measures of each model's accuracy.

The same types of models can be used to predict outcomes of continuous variables or categorical variables with more than two categories. Examples in the context of child labour include measures of the severity of child labour, such as the number of hours a child has worked per week, the number of hazards a child is exposed to or an index containing several child labour variables (e.g. on a scale from 1 to 10). The regression or machine-learning techniques used to construct prediction models for non-binary outcomes are very similar to those used for binary outcomes. The basic steps involved in constructing the model and generating predictions are the same as those described above for binary outcomes. However, different frameworks are needed for assessing the accuracy of models which predict non-binary outcomes (e.g. the sum of squared errors, R^2).

Contexts in which child labour risk models can be used

Risk models can be used for different purposes:

- To identify households with the highest risk of child labour (e.g. among members of a cooperative), so that they can be prioritised for monitoring visits or support.
- To identify cooperatives or communities with the highest risk of child labour, so that interventions can be targeted to areas at greatest risk.
- To broaden the scope of interventions to include "at-risk" households, as well as those in which child labour has been identified.
- To bolster routine child labour monitoring efforts – for instance, in the context of a CLMRS – in order to increase the chance that cases of child labour are identified (e.g. due to the cyclicity of child labour and the possible reallocation of work among children of the same household, in response to shocks or changing circumstances).[12]

[11] see Annexes for more details.

[12] This hypothesis has not yet been tested and is an example of a potential use case for risk models which should be tested and validated.

Practical considerations governing the use of child labour risk models

When using a risk model to predict child labour, it is important to bear in mind the following practical considerations:

- To develop and calibrate a risk model, **at least one existing data set is required**, which includes the **outcome of interest** (e.g. whether or not households use child labour) **and all the variables to be used as predictors** (e.g. detailed information about households and their members), observed for a sample which is comparable to or overlapping with the target population.
- To use a model to predict child labour risk, **data on the target population must be complete and accurate**. Risk-based approaches cannot be used to assess risk among individuals/households for whom data is incomplete or inaccurate, although there are some strategies to deal with missing data.
- **Data management and statistical analysis capacity is required**, either in-house or external, to calibrate and fine-tune a risk model, in order to respond to the needs and operational constraints of each implementer.

Highlights and recommendations

Data quality matters

Complete and accurate data is crucial for the feasibility of risk-based approaches – missing, inaccurate or outdated data may severely compromise a model's performance and utility.

Key considerations for data on targeted households:

1. **High-quality** and up-to-date data from the target population is a must for any risk-based approach (see **case studies A, D and F**).
2. **Regularly updated data** is key, given that the socio-economic situation of cocoa-farming households and communities may evolve quickly, resulting in readjustments to children's working status within their families (see **case studies A, D and F**).
3. **Where data or indicators are missing, it is not possible to make a prediction,** meaning that children from households for which data is missing or incomplete could be wrongly excluded from the risk assessment. In cases where data for many households is missing or unreliable, this could negate the potential advantages of using a risk-based approach (see "Targeting considerations" below, as well as **case study D**).
4. Where data is missing or incomplete, all households or children with missing prediction data must automatically be considered as "high-risk" cases (see **case study E**).

Key considerations for the calibration of data sets:

5. Both nationally representative data sets and other data sets covering a narrower geographical area are suitable for use as calibration data in risk modelling, provided that they include the area in which the risk model is intended to be used. This is because child labour is sensitive to the context, including characteristics of the household and the community, as well as the existence and differential impact of programmes in a specific area (see **case studies A, B and C**).

Recommendations on data quality

- Before considering the deployment of a risk-based approach with cooperatives or farmer groups, assess current data collection practices, as well as the quality of the data management system in place. In the absence of the capacity to collect complete and accurate data, there is no point starting the development of a model. Therefore, if flaws are identified during the assessment, the strengthening of data management capacity should be prioritised, before starting to develop a risk model.
- When it is not possible to make a prediction due to missing or incomplete data, these households and individuals should automatically be considered as "high-risk".
- Use a large-scale and recent data set of reference (e.g. NORC) to calibrate the model-
- Select predictors for the model that are easy to collect and to assess.[13]
- Design data collection tools and train data collectors to minimise measurement errors and missing data as much as possible.

[13] For example, "number of primary schools in the community" rather than "degree of involvement of the school management in violence reduction".

Selecting "good" predictors

Risk-based approaches are not tied to causal considerations and are not intended to describe causal pathways. For this reason, the performance of the model depends neither on the inclusion of predictors that have been demonstrated to be root causes of child labour nor on a set of variables tracking every single aspect of the child's environment. Rather, predictors are useful if they correlate strongly with child labour outcomes and if that correlation is present universally across all subsets of the target population.

Key considerations for selecting predictors:

1. **Including basic child-level variables** (e.g. age, gender and school status) as predictors has proven to increase model performance across the range of projects examined in this study and seems to be an essential element in the most accurate risk models (see **case studies A, B, D and F**).
2. Including the age of the child **improves the sensitivity of the model over time, in the context of a regular risk assessment** (e.g. annual), since child labour risk increases with the child's age. Hence, if children's ages are updated over time in the prediction data set and the risk model is rerun using this updated information, the model will automatically predict more child labour cases among the children assessed during the previous year(s): children previously missed would be flagged as being at risk in this new wave of assessment, due to the effect of age. Provided that every at-risk child is monitored during the year of assessment or predicted cases of child labour are accumulated over time, models based on up-to-date age information witness the number of false negatives (i.e. missed cases of child labour) progressively decrease (see **case study F**).
3. The inclusion of variables whose association with the outcome is *not* statistically significant can still improve the performance of a risk model, meaning that it is not mandatory to remove these (see **case study E**).
4. It is possible to develop a risk model to predict child labour in cocoa, even when no information related to cocoa-production is used in the model (see**case study E**).
5. Where available, the inclusion of a geography-specific child labour prevalence variable among the predictors of a risk model may improve a model's performance (see **case study B**).
6. It is not always necessary to collect new data to develop a risk model – an efficient risk model can be developed using existing data collected for another purpose, provided that the data set is up to date, complete and of good quality (see **case study B**).
7. Risk models with 10–12 variables offer a good compromise between operational constraints related to data collection and quality, on the one hand, and model performance, on the other (common to all projects).

Recommendations on selecting predictors

- Include variables/predictors with the following characteristics in the risk models, since they appear to produce the most accurate ones:
    - ○ Use variables with as little missing data as possible.
    - ○ Use basic child-level variables (e.g. age, gender and school status).[14]
    - ○ Use other variables describing the household and community level.[15]
    - ○ Collect and use a limited but accurately measured set of variables, rather than an exhaustive data collection on a given topic or level of description (e.g. community, farm or household), where measurements are often inaccurate.
    - ○ Do not disqualify *a priori* variables with an undocumented causal relationship with child labour or that display a statistically non-significant correlation with the outcome of interest during the predictor selection process.
- Run the risk model annually on a recurring basis, using up-to-date data, when the age of the child is among the predictors of the model, in order to capture cases that were initially missed and not only generate risk scores for the individuals or households that are newly recorded in the database.

Statistical methods

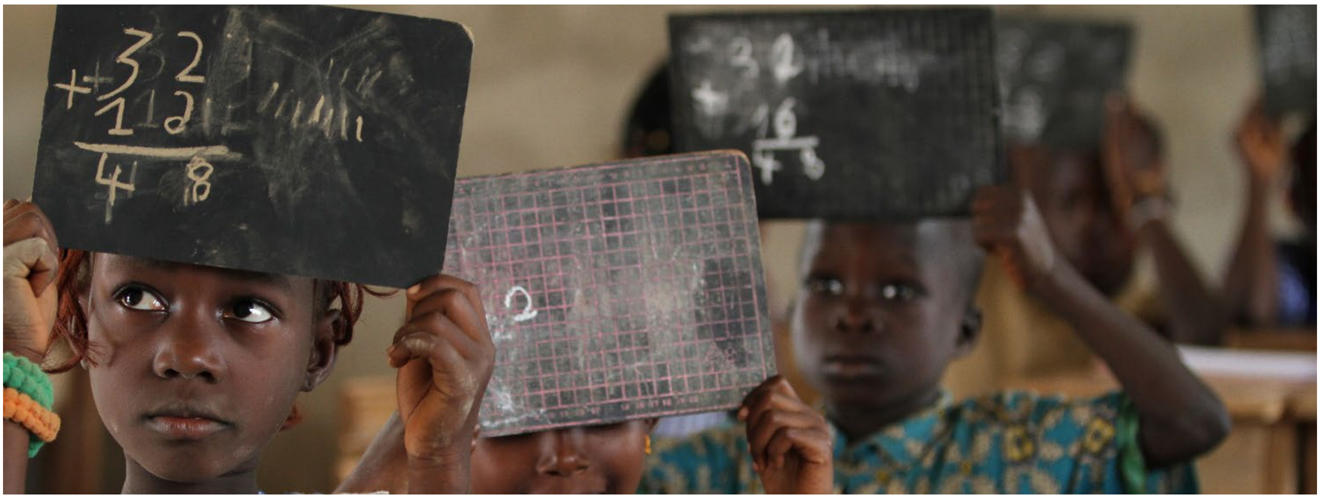
Key considerations governing the choice of a statistical method:

1. The lower the level of modelling, the more accurate the prediction (see Annex 1 for more details about levels). The performance of the models seem to be stratified in virtue of the *lowest level* (e.g. the household or even the child[16]) that they account for (see **case study E**):
    a. Logistic regression – unit of observation: child or household, not accounting for nesting of data –> ROC area ~ 65–70%
    b. Multilevel regression – unit of observation: child or household, accounting for nesting of data at the *community level* –> ROC area ~75–80%
    c. Multilevel regression – unit of observation: child, accounting for nesting of data at the *household level* –> ROC area ~85–90%
2. Choosing to predict a **household-level score** (e.g. the probability that the household has at least one case of child labour) may improve **operational efficiency**, since the cost base for the agent to visit the household is high when compared to the marginal cost for each additional child interviewed.

[14] Since this level of indicator is the most likely to display variations, which will allow the model to generate a fine-grained learning about the outcome of interest.

[15] Since these indicators allow us to capture variations in the child's environment, but also potentially project-related changes from a baseline/end-line perspective.

[16] Variations may occur even at the child level (e.g. the same child may have a changing child labour status) with longitudinal data.



3. Moreover, a **household-level score** is both suitable to operational constraints, because it is easier to plan visits, and statistically efficient, because the model's performance improves (see **case study E**).
4. Finally, a **risk-model approach at the household level** is potentially a suitable way of overcoming the issues of the cyclicity of child labour and of the in-family reallocation of work among the children, both of which cause the child's status regarding child labour to be unstable and identification visits to fail (create false negatives). This is because:
    a. in such an approach, all the children in the same household are tagged as being at risk (and potentially prioritised for a further visit or intervention), which makes their at-risk status robust in the face of the reallocation of work within the household.
    b. if the model is accurate enough, the practitioner can rely on the prediction and choose to operate prevention or remediation actions, *whatever the result* of the identification visit at a specific time-point, thus avoiding "denying" a remediation to a child who is *currently* out of child labour, but *will be back* to it (see **case study F**).

Even the "best" statistical method cannot produce efficient risk models when the data set is inaccurate or contains many missing variables. Promising statistical approaches may yield deceptive results when applied to low-quality data. It is therefore always useful to compare the results of one model by applying it to several data sets (see **case study D**).

Recommendations on statistical methods

- Model the outcome of interest at the lowest level of variation possible (the household or even the child, in the case of longitudinal data sets) using the appropriate statistical method.
- Use a household-level prediction in order to better capture the status of the child, to obtain an actionable output from the risk models and to address issues related to the dynamic of child labour within the households.
- Continuously assess the risk model against the reality unveiled by the monitoring visits and amend it when identification rates are lower than expected.

Targeting considerations

The risk-model approaches presented in this report presuppose that the practitioner retains control over the number of households that the model flags as priority households and can fine-tune this number to balance out ethical considerations (e.g. exclusion of children due to error) and operational constraints. The higher the quality and completeness of the available data, the more flexibility the practitioner has in developing the model (e.g. setting the cut-off value) in order to pursue operational/strategic goals.

1. Depending on the practitioner's choice or capacity, missing data can be dealt with by either: a) ignoring the individuals/households with missing predictions, which may raise ethical concerns, or b) considering all missing predictions as positive cases, which will inflate the number of risk-driven monitoring visits. In both cases, part of the control over the share of children excluded (option a) or included (option b) is lost, which may severely impair the risk-based strategy, if there is extensive missing data: either the risk-based strategy will turn out to be difficult to endorse ethically or the cost savings in relation to monitoring visits / targeted support will shrink, turning the risk-based strategy into a quasi-blanket CLMRS (see **case study E**).
2. The choice of cut-off value directly determines the number of individuals/households flagged as being at risk and thus potentially prioritised. It therefore has substantial operational and ethical implications. Allowing the choice of cut-off to be guided by the inspection of the distribution of the predicted scores may help ensure that these ethical concerns are properly addressed, while also enabling the model to be fine-tuned to suit operational/strategical purposes (see **case study E**).

Recommendations on targeting

When choosing a cut-off value:

- Entrust this task to a person who fully understands the meaning/functioning of the distribution of the predicted scores, in order to produce meaningful and reasonable cut-offs that are fit for purpose.
- Use the median of the predicted scores when the goal is to narrow the scope of monitoring, in order to prioritise at-risk individuals/households. This is a conservative strategy that is recommended if one has sufficient confidence in the model’s accuracy: in most contexts, it misses fewer cases, while selecting 50% of the individuals/households.
- Use the mean of the predicted scores, if it is higher than the median. Otherwise, there is a risk of excluding individuals who are indeed part of the most at-risk share of the population in a high-prevalence area.
- Use the percentiles of the distribution of the predicted scores (e.g. 25th percentile) when the purpose is to select an exact proportion of the target population to suit operational purposes or constraints.
- Choose a cut-off allowing to select a share of the population close to the national or regional prevalence in the absence of specific operational constraints or prior knowledge about the local prevalence rate.
- In the future, more models could be developed to predict the severity of child labour, in order to help practitioners decide the kind and the intensity of prevention/remediation actions to be provided to the children/households, ranked in terms of severity levels.

The cost efficiency of risk-based approaches

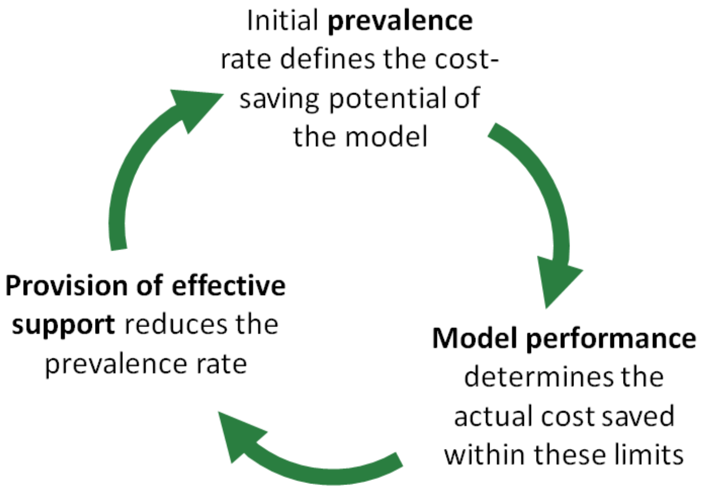
In this report, we have only examined the **potential cost savings** in the context of a CLMRS, in which a risk model could be used to target at-risk children/households for monitoring visits and/or support, as an alternative to conducting a full census and providing support to all the households of an area.

These **potential cost savings** are **limited to the initial phase of the intervention only**: by reducing the number of initial monitoring visits or speeding up the targeting process. These efficiencies allow resources to be reallocated from identification to action, helping children receive support more quickly, rather than reducing the overall intervention cost.

The use of predictive models will not reduce the cost of other key elements of approaches to preventing and addressing child labour: awareness-raising, providing support or conducting follow-up visits with cases identified.

Moreover, cost savings are limited **both by the actual child labour prevalence rate** (since reducing the share of farmers visited below the prevalence rate would amount to voluntarily ignoring a certain share of children in child labour) **and the performance of the model**. Therefore, the costs saved by using a risk-based approach vary greatly, depending on the model and the context (see Figure 5). However, in a longer-term perspective, the **cost-saving potential of a risk-based approach is not fixed and might result in a reduction of the overall costs of the intervention**. In a **virtuous circle**, improving the targeting of remediation via a high-performing risk model may improve the effectiveness of the prevention/remediation actions undertaken and subsequently increase the cost-saving potential of the risk-based approach, and with it the cost of the overall intervention.

Figure 4: Virtuous cycle of cost-effectiveness in a risk-based approach



Initial **prevalence** rate defines the cost-saving potential of the model

Model performance determines the actual cost saved within these limits

Provision of effective support reduces the prevalence rate

The initial prevalence rate temporarily sets the limit of the potential costs saved in monitoring visits, while the performance of the risk model determines how much of this potential is achieved.

The effectiveness of support may, in turn, reduce the prevalence rate and increase the cost-saving potential of a risk-based approach.

Figure 5: Relative cost-effectiveness of different methods of risk-model creation

Method	Households flagged at risk in the sample (%)	Targeting quality score (%)	Cost-saving score (%)	Cost-effectiveness score
Multi-level logit, household-level nesting	41%	98%	54%	0.90
	36%	99%	56%	0.92
	40%	98%	56%	0.91
Multi-level logit, household-level nesting	46%	83%	36%	0.67
	41%	84%	38%	0.69
	47%	84%	39%	0.70
Logistic regression	47%	76%	30%	0.60
	35%	77%	31%	0.61
	48%	77%	30%	0.60

In *Figure 5,* above, we compare:

- the **targeting quality** of several different risk models, defined as the *theoretical* capacity of the predictions made by the model to target all the children in child labour – that is the number of children correctly predicted in child labour divided by the actual number of children in child labour[17] in the sample, expressed as a percentage. This measure is *theoretical*, because it depends on an *estimate* of the number of children in child labour.
- the **cost-saving scores** of these models, defined as the number of monitoring visits needed to identify a certain number of children in child labour based on a risk model, divided by the number of visits needed to identify the same number of children using a full-census

[17] Two possibilities for calculating this figure exist: either by using the prevalence rate *observed* in the external validation data set (if it is reliable, as well as geographically and demographically close enough to the target population) or by using the national or regional prevalence rate reported by a high-quality national survey like NORC. For instance, using the figures from the NORC report, the estimated number of children in child labour in a sample of 1,000 children in Ghana is 1,000 x 0.55 = 550.

approach,[18] expressed as a percentage. This measure is *theoretical*, because it depends on an *estimate* of the prevalence of children in child labour.

- the resulting adjusted **cost-effectiveness score**, defined as the product of the targeting quality and the cost-saving score, expressed as a ratio ranging from 0 to 1 (the closer to 1, the better). In practice, the cost-effectiveness score will never reach 1, since the actual prevalence rate of child labour sets the limit to the model's cost-saving potential. The cost-effectiveness score is computed according to the following formula:

(Targeting quality score x cost-saving score) + actual prevalence rate

This cost-effectiveness score therefore captures the extent to which a specific risk-based approach theoretically manages to **strike the balance between operational concerns and due diligence obligations/ethical concerns** – that is, to what extent it manages to capture all the cases of child labour in a given target population, while reducing the cost of the system and making it more easily scalable.

For example, a risk-based approach that selects 50% of the households but misses 50% of the children in child labour could be considered efficient if assessed only on the basis of its capacity to reduce the number of monitoring visits. It would, however, have a poor cost-effectiveness score, reflecting the inaccuracy of the underlying model and the ethical concerns raised by using it to guide operations.

[18] This figure is also calculated based on an estimation of the child labour prevalence rate. For example, if a risk model enabled 100 true cases of child labour to be identified in Côte d'Ivoire, then the number of visits needed to find this amount of cases with random/classical census visits would be: number of true cases found / estimated prevalence rate – that is, e.g., 100/0.38 = 263 (based on the prevalence rate in the NORC report).

Case studies of risk-based approaches

In this section, we provide a comprehensive description of the six examples of risk models recently developed to predict the risk of child labour in the cocoa-growing areas of West Africa, from which the lessons learned and recommendations presented in this report were drawn.

In each instance, we report the specific context of the project, the method used to predict child labour, the performance of the model(s) and other results, and finally the key learnings.

Contents

A. Predicting hazardous child labour from a register of certified farmers in Ghana

Context

A risk model was developed in the context of a CLMRS innovation pilot implemented by ICI in Ghana, which was launched in 2019.[19] The aim of this project was to test the feasibility of using a readily available farmer register to predict the risk of child labour, which would then make it possible to prioritise high-risk households for monitoring visits under a CLMRS. A publicly available data set on child labour prevalence (the 2015 Tulane survey) was used to calibrate the child labour risk model.[20] Only those variables were considered as predictors that also featured in the existing register of certified cocoa producers in the ECOM-Nestle supply chain in Ghana.

Stakeholders involved

This risk model was developed by ICI, in the context of a project jointly funded by the Swiss State Secretariat for Economic Affairs (SECO) and Nestlé. ECOM provided data from the farmer registers of two cooperatives in the districts of Asunafo and Suhum, Ghana, which were used to test the model.

Method

1. Using the 2015 Tulane data set for Ghana as the reference data set, 1,541 records for the children of cocoa-growing farmers in Ghana were used to **calibrate** the model. In a first iteration of the model, only variables that were also available in a register of certified cocoa producers held by ECOM were considered as predictors.
2. A child labour prevalence survey was used as the **validation** data set. The survey was conducted amongst a group of members of two certified cooperatives, in order to assess the feasibility of the risk modelling approach and to determine the accuracy of the model.

The model used 12 variables to predict: a) each child's hazardous child labour status and b) three measures of child labour severity.[21] Several iterations of the model were developed using different statistical methods. First, *logistic regression* was used to predict a child's hazardous child labour status, while *linear regression* was used to predict the severity of child labour. Second, a *multilevel regression* was used, to predict the likelihood of at least one case of hazardous child labour in a household.

[19] ICI (2020). *Predicting child labour risk at household level: A risk model for cocoa farming households in Ghana*.

[20] Tulane University (2015). *Survey Research on Child Labor in West African Cocoa Growing Areas*. Even though the Tulane survey has recently been shown to contain methodological flaws, it provided the best nationwide data available at the time the risk model was under development. Moreover, no specific data-quality issues were detected during the development of the models. Nonetheless, the fact that the data was collected in 2013 and used in a model developed in 2020 may have affected the performance of the models used in this project.

[21] Even though the initial idea was to test a model which would only use information available in the ECOM farmer register, this did not yield a model with any useful predictive power. As a first improvement, two child-level predictors were added to the model, namely the child's sex and age (these were not readily available in the ECOM farmer register, but were found in the reference and validation data). This modification brought about an increase in the predictive power of the model to a level where it started to be interesting from an operational point of view.

Results

Child-level prediction:

- A logistic regression model with child labour as a binary outcome, using a cut-off value of .55, yielded 58.3% sensitivity and 62.6% specificity (ROC area: 0.63 – see Annex 1 for more details about ROC area); 68% of the households in the sample would have been considered “high risk” and selected for monitoring visits.
- This model could be further improved by using multilevel logit regression with household-level nesting, which increased sensitivity to 88% and specificity to 85% (ROC area: 0.86).
- A linear regression model, with the number of hours a child had worked in the reference week as the outcome variable, yielded 64.4% sensitivity / 63.6% specificity (ROC area: 0.64).

Household-level prediction:

- At the household level, a logistic regression model with at least one child in child labour as the outcome performed significantly better, with a sensitivity of 80.25% and a specificity of 50.70% when using a 0.58 cut-off (ROC area = 0.7).

Learnings

- The **child’s sex and age** turned out to be by far the most powerful predictors of hazardous child labour from amongst a set of child and household characteristics. Without this information, it was not possible to produce a model with sufficient predictive power to be operationalised.
- When using a continuous indicator of child labour severity (specifically, the number of hours a child works), the model’s performance was slightly better than when using a binary indicator of child labour.
- Aggregating predicted scores to obtain scores at higher level of hierarchy (e.g. child-level scores aggregated into household-level scores) improved the accuracy of the predictions.
- Models based on multilevel regressions seem to perform better than linear regressions.

Summary of model characteristics and performance

Data source	Outcome(s) of interest	Predictors (and respective levels)	Statistical method	Sensitivity/ specificity (ROC)
2015 Tulane child labour prevalence survey (calibration) + locally conducted prevalence survey (validation)	*Child level* Hazardous child labour status # of hours worked per week	*Child level* Age, gender *Household level* Education, age, gender of the head of household # of children in household # of workers employed Household's access to drinking water and electricity Land under cocoa cultivation (ha) Other cash crops cultivated? Use of fertiliser and pesticides?	Logistic regression, predicting hazardous child labour at child level	58% / 63% (ROC area = 0.63)
	Household level Hazardous child labour status is positive for at least one child in the household		Linear regression, using hours worked as outcome to predict high child labour risk	64% / 64% (ROC area = 0.64)
			Multilevel regression	ROC area = 0.86

B. Predicting hazardous child labour at household level using machine learning

Context

Following a public commitment to achieve 100% coverage of Child Labour Monitoring and Remediation Systems (CLMRS) by 2026, Cargill began developing risk models to predict child labour. The aim was to identify vulnerable households to prioritise for child labour monitoring visits.

The model development process included several tests to assess whether low-cost, *existing* data sets – in this case geospatial and demographic information – could improve the accuracy of child labour predictions and therefore improve the cost-effectiveness of the system. Since the suitability of different data sources for predicting child labour can vary significantly, it is interesting to test to what extent existing data, which can be acquired at a relatively low cost, can be used to make effective risk-models. This case study compares the performance of a series of models, one using low-cost data alone, and the others using such data in combination with more granular data about communities and households.

Stakeholders involved

This risk model was developed by Cargill, using data collected in-house to test the model.

Method

- Geospatial and demographic information (e.g. number of farms per cooperative, farm size, number of schools within 5 km), coaching surveys of farming practices and basic household information, and CLMRS data were used to **calibrate** the models running a machine-learning analysis based on the Extreme Gradient Boosting method
- CLMRS data[22] was used to **validate** the predictions of the actual child labour cases.

The household's child labour status (i.e. whether or not the household reported at least one case of hazardous child labour) was the outcome of interest. Four risk models were selected to identify the best predictors among a set of variables describing the regional, cooperative, farm, household and farmer level(s).

Results

Several different models were created, which performed as follows:

- a 3-predictor model (M1) using only publicly/easily accessible data yielded a 0.75 ROC area (see Annex 1 for more details about the ROC area)
- a 5-predictor model (M2) adding neighbourhood information (i.e. child labour prevalence density and a risk score) to the former model yielded a 0.83 ROC area
- an 8-predictor model (M3) adding farmer-, farm- and household-level information to publicly accessible data yielded a 0.85 ROC area

[22] This is the particularity of machine-learning methods, namely that they use part of a data set to learn about the data (calibrate the model) and another part to predict the outcome of interest.

- a 10-predictor model (M4) including the variables of the three other models yielded a 0.89 ROC area. When using this model, targeting a 90% sensitivity resulted in a false positive rate of about 35%.

Learnings

- Combining information from multiple sources makes it possible to develop a risk model that performs better than randomly targeted household visits.
- Including data related to child labour in the targeted areas (M2, M4) improves model accuracy, as does including more granular information about the context (i.e. farm-, farmer- and household-level predictors).

Summary of model characteristics and performance

Data source	Outcome(s) of interest	Predictors (and respective levels)	Statistical method	ROC
Geospatial and demographic information, coaching survey, CLMRS data (calibration) / CLMRS data (prediction)	*Household level* Hazardous child labour status is positive for at least one child in the household	*Region level* # of schools Schools within 5 km Child labour propensity Demographics *Cooperative level* Maturity # of farms/cooperatives *Farm level* Productivity Size Household level # of children Age, gender of children *Farmer level* Literacy, income, gender, age	Extreme Gradient Boosting	M1: ROC area = 0.75 M2: ROC area = 0.83 M3: ROC area = 0.85 M4: ROC area = 0.89

C. Using risk models to *increase* the number of children targeted for closer monitoring

Context

In contrast to other child labour risk models presented in this report, which use risk-based approaches to *decrease* the number of children targeted for monitoring visits, this model was created with the aim of *increasing* the number of children who would receive closer monitoring. It was developed in the context of an existing CLMRS, in which a first round of farm visits identified relatively few cases of hazardous child labour. Normally, only households in which a hazardous child labour case is identified would be prioritised for closer monitoring (i.e. would receive follow-up visits every 6 months, rather than every 12 months). In this case, the model was designed to identify additional children at risk of hazardous child labour, to ensure that no case of hazardous child labour had been missed and to prevent more vulnerable children from falling into child labour. Any household identified as high risk by the predictive model, but in which no case of hazardous child labour was identified, was added to the list of households to receive closer monitoring.

Stakeholders involved

The risk models were developed by ICI for use in Tony's Chocolonely supply chain in Côte d'Ivoire.

Method

1. Two data sets were used to **calibrate** the risk models described in this case study:
   (1) a subset of the NORC[23] data set, limited to the regions of Côte d'Ivoire in which Tony's Chocolonely's suppliers are concentrated (479 records from children, models M1);
   (2) data from Tony's Chocolonely's CLMRS (13,017 records from children, models M2).
2. The outcome to be predicted was whether a child engaged in hazardous child labour. Two regression methods were used to calibrate the models: logistic regressions (M1a, M2a) and multilevel logistic regressions accounting for the variations in the outcome at the household level (M1b, M2b), resulting in four models for comparison.
3. The CLMRS dataset was used to **validate** the predictions of hazardous child labour cases from the M1 models.
4. The best model from each series (M1, M2) was selected to produce two competing predictions (see the table below for the respective performances of each of the four models).

The models used up to 12 variables to predict each **child's** hazardous child labour status (i.e. in/out of hazardous child labour). In line with the aim of the exercise, which was to increase the number of children receiving more intense monitoring visits, and taking into account the fact that the hazardous child labour identification rate amongst the target population was substantially lower than 25%, the 75th percentile of the predicted estimates was chosen as the cut-off value, in order to select the 25% of children considered "most at risk" within the sample.

[23] "Assessing Progress in Reducing Child Labor in Cocoa Production in Cocoa Growing Areas of Côte d'Ivoire and Ghana.", Sadhu, S., Kysia, K., Onyango, L., Zinnes, C.F., Lord, S., Monnard, A. and Arellano, NORC at the University of Chicago, 2020.

Results

- The model calibrated with the NORC dataset yielded a 77% specificity and a 99% sensitivity (ROC area: 0.88; see Annex 1 for more details about ROC area).
- The model calibrated with the CLMRS dataset yielded 78% specificity and 100% sensitivity (ROC area: 0.89).
- The first model performed even better with a cut-off value set to the mean of the predicted scores (95% specificity, 100% sensitivity, ROC area: 0.98, marked * in the table below)
- In all cases, multilevel logistic regressions outperformed the logistic regressions which ignored nesting of data at the household level (see table below).
- Overlapping predictions of hazardous child labour cases between both models predicted hazardous child labour with 87% accuracy (and 87% sensitivity).

Learnings

- Child labour prevalence data from representative samples of cocoa farmers (e.g. region-specific extracts from the NORC data set) provide suitable calibration data for predicting hazardous child labour in specific supply chains within the same regions, even when the calibration data sample size is relatively small.
- Using multilevel regression can improve the accuracy of the risk models, provided that it accounts for variations in the outcome at the lowest levels (i.e. household, rather than community or region).
- It is possible to control the number of individuals selected by the model as “at risk” by adjusting the cut-off value according to the target number of children/households to be supported.
- Using the mean of the predicted estimates as a cut-off value allowed a better balance between sensitivity and specificity.

Summary of model characteristics and performance

Data source	Outcome(s) of interest	Predictors (and respective levels)	Statistical method	ROC
NORC restricted to two regions in Côte d’Ivoire + CLMRS (calibration), CLMRS (validation)	*Child level* Hazardous child labour status	*Household level* Education, age, gender of the head of household # of children under 5 in household # of children out of school, in primary and in lower secondary # of male and female children in household # of adults in household *Farm level* # of sharecroppers on the cocoa farm # of cocoa crops under cultivation in the farm *Community level* Access to electricity Access to improved drinking water source Presence of a primary health centre	Logistic regressions	M1a: 59%/76% (ROC area = 0.68) M2a: 13%/81% (ROC area = 0.47)
			Multilevel logistic regressions	M1b: 99%/77% (ROC area = 0.88) M1b*: 100%/95% (ROC area = 0.98) M2b: 100%/78% (ROC area = 0.89)

D. Predicting a child's child labour status from cooperative registers in Côte d'Ivoire

Context

This project aims to support cooperatives to identify members most likely to use child labour so that they can better target interventions to prevent and address child labour. The project aims to create a hazardous child labour risk model based on existing information from farmer registers.

Similar to case study A in this paper, this project tested the feasibility and performance of a risk model based on the limited set of variables *already* available in registers held by five cooperatives, as a first step towards scaling up a risk-based approach to targeting a further 40 cooperatives in Côte d'Ivoire. While data was shared by five cooperatives during this first stage, much of it was incomplete or only available in paper-based format. As a result, data from only one cooperative, was used to develop the model described in this case study.

Stakeholders involved

This model was developed by ICI, as part of a project funded by GIZ/*Centre d'Innovations Vertes* (CIV). Five cooperatives in Côte d'Ivoire shared data and were involved in the initial development of a risk model. Data from the cooperative NECAAYO was used to develop the risk-model. In a later phase of the project, the risk model will be piloted in a total of 40 cooperatives.

Method

Four **independent** steps were taken (see points 1, 2, 3 and 4 below) to test the feasibility of building a model **using existing information from cooperative registers**. Since no child-related information was available in the data sets of the five cooperatives, an additional field survey on hazardous child labour prevalence was conducted, which served: a) to assess the hazardous child labour status of the children living in the households of members of the five cooperatives; b) to collect up-to-date and complete data (including child-level data) from members of the cooperatives; c) to evaluate the quality of the data collected and managed by the cooperatives.

1. The NORC Côte d'Ivoire data set (2,127 observations) was used to **calibrate** an initial risk model, while the data set of the only cooperative with a complete register, NECAAYO (774 observations), was used for the **validation of the predictions**. Information gathered on children's hazardous child labour status from the field survey was inputted into NECAAYO's data set, which covered the same farmers.
2. The NORC subset was split in two equal parts balanced between the regions, with the first serving to **calibrate** the model and the second to **validate the predictions** of hazardous child labour.
3. The whole NORC subset was used to **calibrate** the model, and the survey data for all the cooperatives (including NECAAYO) were used to **validate the predictions** of hazardous child labour.
4. We used the same data sources as in step 3, but included three child-level variables in the model (i.e. age, gender and school status), borrowing information made available in the surveys, in order to compare their respective levels of efficiency.

A child's hazardous child labour status was the outcome to be predicted by means of multilevel logistic regressions – accounting for the nested nature of the data at the community and household level – and 9 or 12 predictors (see table below).

Results

- At the first step, the model had a good internal validity (ROC area = 0.92 – see Annex 1 for more details about ROC area), but performed very poorly when comparing the predictions to the actual child's status in NECAAYO (**ROC area = 0.5**, while selecting 55% of the households in the sample). The hypotheses advanced to explain this underperformance were that data available at the cooperative level was either outdated or deficient, as the result of unreliable data collection, inadequate data storage (i.e. on paper) or faulty post-digitalisation.
- At the second step, the model had a sensitivity and specificity of 89% (**ROC area = 0.89**), while selecting 44% of the households of the sample.
- At the third step, the model had a sensitivity and specificity of 85% (**ROC area = 0.85**), while selecting 39.75% of the households in the sample, when applied to the data collected by the survey among NECAAYO's children, and a sensitivity of 87% and a specificity of 81% (**ROC area = 0.84**), while selecting 36% of the households of the sample, when applied to the data collected by the survey among all the children in the five cooperatives.
- At the fourth step, the model had a sensitivity of 92.5% and a specificity of 87% (**ROC area = 0.90**), while selecting 44.6% of the households of the sample, when applied to the data collected by the survey among all the children of the five cooperatives. The difference in the ROC curves between the one produced by this model and the one in step 3 was strong and significant (χ^2=36.44, p<0.0001).

Learnings

- It is possible to develop an effective risk model using only the limited set of variables available in cooperative registers. However, if the predictor data fed into the model is outdated, inaccurate or incomplete, the model will produce poor results.
- The quality and completeness of available data, the regularity of data collection, and the capacity to digitalise and manage data at cooperative level are therefore key for the feasibility of developing or using such a risk model.
- While it is possible to make an accurate prediction of hazardous child labour even when no child-level data was available (see results of steps 1–3 below), including child-level information, such as age, gender and school status, improves the model's accuracy.

Summary of model characteristics and performance

Data source	Outcome(s) of interest	Predictors (and respective levels)	Statistical method	ROC
NORC (calibration) + Cooperative data, NORC, prevalence survey (validation) NB: child-level information (outcome of interest and child-level predictors) come from the prevalence survey	*Child level* Hazardous child labour status	*Household level* (for the 3 steps where the predictors are limited to the variables available in the cooperative registers) Education, age, gender of the head of household # of permanent adult workers on farm Volume of cocoa produced in the last year (tonnes) Class of cocoa farm according to tonnage (5 categories[24]) Land under cocoa cultivation (ha) Yields of cocoa cultivation (tonnes/ha) Non-cocoa land under cultivation (ha) *Child level* (for step 4 only) Age, gender, school status	Multilevel logistic regression	Step 1 57% / 44% (ROC = 0.5) Step 2 90% / 89% (ROC = 0.89) Step 3 86[25]-87[26]% / 85-81% (ROC = 0.84-0.85) Step 4 92.5% / 87% (ROC = 0.90)

[24] Category is one if tonnage < 1, two if 1 ≥ tonnage < 3, three if 3 ≥ tonnage < 5, four if 5 ≥ tonnage < 7, and five if tonnage > 7.

[25] When prediction is made based on the NECAAYO survey sample alone.

[26] When prediction is made by including the survey sample of all cooperatives.

E. Assessing the benefits of predicting child labour at household vs. individual level

Context

Rather than predicting the hazardous child labour status of each individual child, a risk model can also be designed to predict the likelihood that at least one case of hazardous child labour is present in a given household. While the ultimate aim of the risk model remains the identification of children in hazardous child labour, making a prediction at the household level presents obvious operational advantages while potentially improving the model's predictive power. This is particularly relevant in operational contexts where interventions or monitoring activities are targeted at the household level, making the ability to identify high-risk households valuable.

This project assessed the performance of two strategies for predicting the hazardous child labour status of the household (i.e. whether at least one case of hazardous child labour is present in a household). The project also explored ways of dealing with missing data from target households, by considering missing predictions as "positive" cases, that is, automatically assuming they represent cases of hazardous child labour.

Stakeholders involved

This model was developed by ICI as part of its ongoing research activities.

Method

1. The Côte d'Ivoire NORC subset (2,734 child observations) was randomly split into two equal subsamples balanced between the regions. The first one (1,371 observations) served to **calibrate** three series of risk models (18 models in total), each using a different set of variables, three different regression methods (logistic regression, multilevel regression at community-level, multilevel regression at household-level) and two strategies for producing a hazardous child labour prediction at the household level:
    - directly predicting the *household's* status (dependent variable in the regression),
    - predicting the *child's* status (dependent variable) and then using aggregated child results (presence of at least one predicted case of hazardous child labour) to predict the household status (two-step prediction).
2. The second subsample (1,363 observations) was used to **validate** the capacity of these predictions made **at household level** to accurately **identify**: 1) **households** with at least one case of hazardous child labour and 2) **children** in hazardous child labour.[27]
3. The models used up to 13 variables to predict hazardous child labour status (see "More details on the models and results" below). As previously mentioned, logistic regressions and multilevel logistic regressions were used to account for the variation of the dependent

[27] This made sense since a true positive household may have several children in child labour. Therefore the "return" of one true positive prediction at the household level may greatly vary from 1/1 to 1/many positive children. Therefore, a model that predicts the status of the households with 80% accuracy may predict the status of the children with 90% accuracy. In other words, a household-level prediction is always at least as good at predicting the child's status than it is at predicting the household's status.

variables at the community or household level. Along with the validation step, the operational and statistical impact of considering missing predictions as positive cases was reported.

Results

- Multilevel regressions accounting for the variation of the outcome at household-level (ROC ≈ 0.9 – see Annex 1 for more details about ROC area) systematically performed better than those accounting for variations at the community level (ROC ≈ 0.7). In both cases, the accuracy of multilevel models was better than that of logistic models which ignored nesting of the data (ROC ≈ 0.65) (see below).
- The two-stage strategy systematically performed at least as well as the direct one, while making it possible to use multilevel regressions that account for the variations of the outcome at household level, and thus to create the best-performing models[28]
- Considering missing predictions due to missing data as "positive" cases, i.e. assuming a case of child labour, proved to be the best approach to dealing with missing data.

Learnings

- Including child-level variables (e.g. age and gender) improves the models, whatever the statistical method used.
- Several models building on different sets of variables describing different areas (access to services, schooling, women empowerment, etc.) may display the same performance.
- It is possible to efficiently predict child labour in the cocoa sector, even when using no cocoa-related variables (see description of M3 in the next section).
- Variables whose association with the outcome is *not* statistically significant can still improve the performance of a risk model (i.e. it is not mandatory to remove them).
- **Multilevel logistic regression at HH-level** is a promising and flexible method for a risk-based approach.
- **A two-step strategy performs better** than direct predictions, while making it possible to reconcile several different operational challenges: e.g. narrowing the scope of households to be targeted, while obtaining predictions that capture the true cases of hazardous child labour with higher accuracy.
- **Tagging missing predictions as positive cases is an efficient measure for offsetting** data incompleteness.

[28] By definition, a multilevel regression accounting for the variation of the outcome at the household level is not interesting when the outcome is the status of the household. This is because, in such a case, no variation within the household is possible. On the contrary, when the dependent variable is the status of the child (used in the two-step strategy), variations occur (i.e. within the same household, child 1 may be in child labour = 1, child 2 out of child labour = 0, child 3 in child labour = 1, etc.).

More details on the models and results

The following table displays the performance (i.e. sensitivity, specificity, ROC area) of three models tested on the NORC dataset (see below), which use household- or child-level hazardous child labour status as outcomes (corresponding to either a direct or a two-stage prediction), logistic regressions or multilevel regressions accounting for the nested nature of the data at the household or community level, and a cut-off value arbitrarily set at the mean of each predicted value. These models are:

- **Model 1** is an **11-variable** model based on a mixture of low standard error and significantly contributing predictors at the **community level** (i.e. availability of casual daily labour, percentage of cocoa farms planted in the last 10 years, average number of classrooms per primary school), **household level** (i.e. type of dwelling, pesticide/herbicide use on the farm within 12 months, percentage of adults involved in agriculture, percentage of adults involved in cocoa, average age of children) and **child level** (i.e. age, gender and schooling status).
- **Model 2** is an **11-variable** model based on a mixture of low standard error and significantly contributing predictors at the **community level** (i.e. average size of the cocoa farms, percentage of cocoa farms planted in the last 10 years, access to improved source of water, distance to closest cocoa shed, tonnage of cocoa production the year before, distance to the closest sealed road, distance to the district capital) and **child-level** (i.e. age, gender and schooling status).
- **Model 3** is a **13-variable** model based on a mixture of low standard error and significantly contributing predictors at the **school environment level** (i.e. presence of preschool, number of years since the primary schools were created, number of occurrences of caning in primary schools, number of occurrences of verbal disciple in primary schools, average fees for primary schools, presence of grants for primary schools, average amount of grants for primary schools, number of primary schools with functional parent-teacher associations, average number of students/classroom in primary schools, distance to the closest primary school) and at the **child level** (i.e. age, gender and schooling status).

Summary of model characteristics and performance

				Household identification		Child identification		
Models	Regression	Prediction strategy	% households predicted with child labour	Sensitivity	% of true positive households identified	Sensitivity	Specificity	ROC
Model 1	Logistic	Direct	35%	64%	60%	74%	61%	0.66
	Logistic	Two-stage	47%	59%	73%	76%	53%	0.64
	Multilevel (ML) community-level	Direct	47%	65%	80%	83%	60%	0.7
	ML community-level	Two-stage	46%	65%	79%	83%	57%	0.7
	ML HH-level	Direct	NA**					
	ML HH-level	Two-stage	41%	92%	98%	98%	80%	0.89
Model 2	Logistic	Direct	31%	60%	57%	69%	58%	0.62
	Logistic	Two-stage	35%	60%	73%	76%	52%	0.64
	ML community-level	Direct	34%	73%	72%	75%	70%	0.72
	ML community-level	Two-stage	41%	67%	80%	84%	58%	0.71
	ML HH-level	Direct	NA**					
	ML HH-level	Two-stage	36%	94%	99%	99%	80%	0.9
Model 3	Logistic	Direct	34%	63%	56%	79%	61%	0.66
	Logistic	Two-stage	48%	59%	74%	78%	49%	0.63
	ML community-level	Direct	38%	72%	70%	68%	73%	0.71
	ML community-level	Two-stage	47%	68%	82%	85%	57%	0.71
	ML HH-level	Direct	NA**					
	ML HH-level	Two-stage	40%	94%	97%	98%	81%	0.89

***Running a multilevel regression accounting for the variations of the household hazardous child labour status at the household level is meaningless in this case, since there is by definition no possible variation in this outcome within the households.*

F. Assessing a risk model's capacity to identify initially "missed" children in child labour over time

Context

Every risk model entails a certain degree of error, which raises ethical concern about false negatives: what about those children *in* hazardous child labour who were "missed" by the model?

One way to address this concern is to use risk models as part of an *ongoing process.* This approach is especially relevant if a model includes time-sensitive predictors that are positively correlated to hazardous child labour, such as a child's age. Age is indeed known to increase the risk of child labour – therefore as a child grows older he or she is more likely to be flagged as being at risk by a risk model. Rerunning the same risk model on an annual basis, using up-to-date records, means that a child considered at *low risk* during the first assessment may be reclassified as at *high risk* in subsequent years.

A simulation based on the NORC data set was undertaken to test the theoretical capacity of a risk model to **identify initially unidentified children over time**. The time performance of a model is assessed by:

- The share of children initially missed and eventually flagged as at risk over time
- The speed at which a child is flagged as at risk

Stakeholders involved

This model was developed by ICI as part of its ongoing research activities.

Method

1. The Côte d'Ivoire NORC subset (2,734 observations) was randomly split into two equal subsamples balanced between the regions. The first one (1,371 observations) served to calibrate three risk models of hazardous child labour that were fine-tuned to yield different levels of accuracy (sensitivity = 96%, 76% and 67%, see legend of the figure below for more details about the models) and that included different sets of non-time-sensitive predictors, with the exception of the child's age. The rationale behind voluntarily varying the performance of the models was to test whether a higher initial performance makes it possible to capture children initially missed on a large scale and with greater speed.
2. For each model, two outcomes were modelled: the hazardous child labour status of the child and the hazardous child labour status of the household (i.e. whether there is at least one case of hazardous child labour predicted in the household). This was done in order to test whether modelling hazardous child labour at the household level would improve the performance of the model when it comes to recapturing missed children over time. In total, six models were produced.
3. In order to make a prediction, the second subsample of NORC was used to run a "first wave" of risk assessment, representing a year-one wave of identification.

4. Next, we incrementally added one year to the age of the child over the course of 11 simulated "years", while keeping the other predictors stable, and recomputed the risk scores for the children, taking into account their "updated" age.
5. For each simulated wave of assessment, the share of children in child labour missed by the model was reassessed, taken as the number and percentage of false negative children[29] since the first wave. If, in subsequent assessment waves, the model managed to identify children who had been previously missed, these newly identified children were deducted from the number of false negative children, such that the overall percentage of false negatives within a given cohort decreases in proportion to the true positives discovered over time.

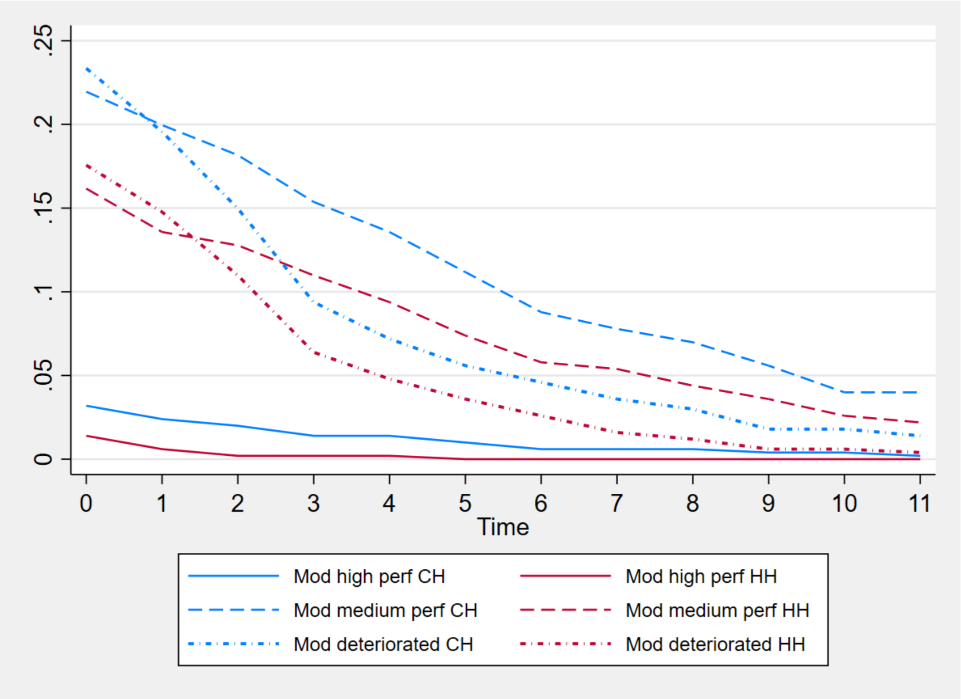
Results

- The initial share of false negatives ranged from 3% to 23%, with this share shifting to 0–11% after five years and 0–4% after 11 years. In other words, 80–100% of the false negatives from the first year were eventually identified.
- Multilevel models accounting for the *variation of the outcome at household level* performed better than multilevel models accounting for the variation in the outcome at the community level.
- Models using the hazardous child labour status of the *household* for the outcome performed better than models with the child's status as outcome.
- Multilevel models that accounted for the variation in the outcome at the household level, but whose performance was intentionally reduced by an unrealistic cut-off value initially performed worse than community-level models with a "reasonable" cut-off value, but then went on to rapidly outperform them over time.
- Models that initially performed better recaptured more false negatives, more rapidly than the other models that had performed worse (see graph below).

[29] The assumptions lying behind this approach were:

1. Child labour does not occur randomly (i.e. the correlations between the predictors and the outcome **within the same model** are stable over time for the same sample).
2. A change occurring in reality among the predictors positively correlated to the risk would increase the child's score and therefore make them more detectable by the model. This aspect of a changing environment that occurs in the reality can be neglected in the simulation, which is therefore **conservative**/underestimated (i.e. as it relies on one among many predictors that are positively correlated to hazardous child labour, the rate of children recaptured across the time can be higher, but certainly not lower).
3. A change occurring in the reality among the predictors negatively correlated to the risk (lowering the score and making the child less "visible" to the model): 1) could be balanced by the unaccounted effects of the positively correlated variables; 2) could mirror an ongoing shift of status of the child (in –> out) (even though we are not in a causal framework); 3) is not always possible in the short run (e.g. distance to the closest cocoa shed).

Figure 6: Performance of the six models over time



In the legend, "CH" and "HH" refer to the outcome of the model: the hazardous child labour status of the child (CH – blue lines) and of the household (HH – red lines). The "high-performance" models ("Mod high perf") are based on multilevel regressions accounting for the variation in the outcome at the household level, while the "medium-performance" ("Mod medium perf") models account for the variation in the outcome at the community level. "Deteriorated" models are high-performing models that were intentionally set with the wrong cut-off value, which exaggeratedly reduced the share of households flagged as at risk. While the high-performance models are clearly better (plain lines), the initial similarity between the other two categories ("medium" and "deteriorated") fades rapidly after about two years of assessment, with a much faster improvement for the "deteriorated" model (the slope of the dotted line).

Learnings

- Even though each model entails a certain degree of error, rerunning models on a regular basis means that most of the children from the same cohort who were initially missed by the model are eventually correctly identified, rather than neglected.
- Including the age of the child as a predictor in a risk model does improve its accuracy and is an asset when it comes to considering its sensitivity over time.
- The better the initial performance, the greater and faster the error recovery over time, in relation to the same cohort of children.
- The potential accuracy of the model determines its long-term capacity to identify children who were initially missed to a greater extent than having a well-defined cut-off – that is, a high-performing model with a wrong cut-off value recaptures more missed children more quickly than an average-performing model with a correct cut-off value.
- Even though risk models that include the age of the child demonstrate substantial error recoverability over time, it is important to take into account the potential time lapse between the first assessment and the eventual identification: for example, if a child who has been in

hazardous child labour since the age of 10 is not identified until the age of 16, the child is much more likely to have been exposed to multiple hazards for a longer period and may be far more difficult to provide with appropriate support to stop hazardous work. Therefore, the pace of recapture is of the utmost importance. In this respect, initially high-performing models (i.e. those that model the outcome at the most granular level and assess the outcome at the household level) are better.

- In line with the results of other studies, granular multilevel models using hazardous child labour household status were found to be the most recommendable and efficient approach.

Annexes

Annex 1: Key concepts for understanding the development, diversity and performance of risk models

Levels

Levels are a key notion for understanding how risk models are built and why some may outperform others. The following section outlines important distinctions to be made regarding the various aspects of these levels:

- **Levels at which the data is sourced** (e.g. child interviews providing information at the child level, community chiefs interviews providing information at the community level)
- **Levels at which the data is collected and managed** (is the data collected / made accessible by a cooperative or by an entity encompassing several cooperatives or communities?).
- The **level at which the outcome variable is measured**[30] **and at which the prediction is made**: for example, do we want to predict the child labour status of the child (is the child in child labour?) or the status of the household (is there at least one case of child labour within the household?). Even if the model's output is at a disaggregated level, subsequent processing of this output makes it possible to aggregate the prediction to higher levels, for example when using child status to predict household status.
- The **level(s) at which predictors are measured:**[31] one predictor in the risk model may describe the state of the community (e.g. access to grid electricity), whereas other predictors in the same model may describe the state of the household (e.g. total income generated last year or head of household's gender). Risk models generally include predictors measured at several different levels.
- **The hierarchical level(s) existing within a set of nested observations:** in a given sample, individuals may belong to a greater whole, for example farmers within communities/cooperatives or pupils within classrooms, which are in turn found within schools. When this is the case, observations are said to be "nested" within these "natural" levels. This notion is important in risk modelling, as it reflects the fact that observations may have different environments, which may have an influence (e.g. teachers in different classes may have different degrees of motivation or skills, while schools may have different financial resources at their disposal for rolling out a policy). In natural settings, one could argue that observations are always nested within a certain hierarchical level.
- The **level(s) of pooling of the observations** operated by the model (i.e. the hierarchical levels accounted for by the model): when the observations are nested within several

[30] Note that the levels of data measurement and data collection are to be distinguished: an outcome describing the state of the child is measured at the child level, but may be collected at another level, typically the household level (e.g. by surveying the head of the household).

[31] Again, this is different from the level at which the predictors are collected.

hierarchical levels (see above), a model can **consider the sample as a single population** and use the mean of the outcome of interest for all the observations of the sample (complete pooling) **or** it can **consider the existence of hierarchical levels** in the sample and use the overall sample mean *and* the **means** of the outcome observed within the units at a given level (partial pooling), e.g. the prevalence of child labour within each community in the sample. The second approach allows for a more fine-grained estimation, as **it takes into account** the presence of **local particularities when making predictions** (see the figures below for an illustration). For example, rather than specifying a regression model such that it considers only the mean of the sample (complete pooling) and treats the variations at local levels (communities or households) as noise, it can be specified as a *multilevel* regression, so that it explicitly accounts for the fact that there is variation across communities or households.

Cut-off:

Definition

A risk model based on regression techniques produces a continuous score for each observation in a sample. When the outcome of interest is binary (e.g. whether or not a child is in child labour), since the predicted score ranges continuously from 0 to 1, we have to decide where to divide (cut) the scores, in order to translate them back into binary form for the prediction. For example, one may decide that when a score is below the cut-off value of 0.6, the case is predicted to be zero (e.g. not in child labour), while score above 0.6 indicates that the case is predicted to be one (e.g. in child labour).

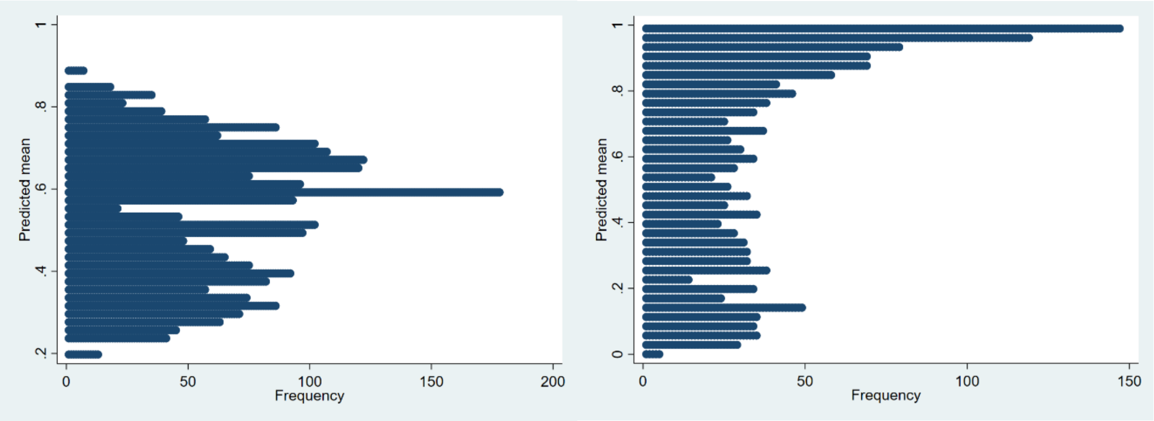
Setting the cut-off is key to achieving the desired balance between the sensitivity and specificity of the prediction.

Choosing a cut-off

The cut-off value may be defined on the basis of previous knowledge (e.g. known prevalence at national scale) or by using the mean or percentiles of the distribution of the predicted scores.

The way in which these scores are distributed among individuals or households may vary greatly (see figures below). Exploring and visualising the distribution of the predicted scores is helpful for informing the decision of where to define the cut-off value and thus meaningfully split the population into different categories of risk – as well as to fine-tune the models in order to obtain an exact proportion of the sample that is tagged as at risk. The **mean**, **median** and other **percentiles** are key parameters of the distribution which respectively inform us about what the most frequent score is (mean), at what point of the predicted scores the sample population is divided into two equal parts (median) and from what point we will find exactly, e.g., 25% of the sample population with the highest scores (in this example, the 75th percentile).

Figure 7: Example of the different the distribution of the scores predicted by a risk model

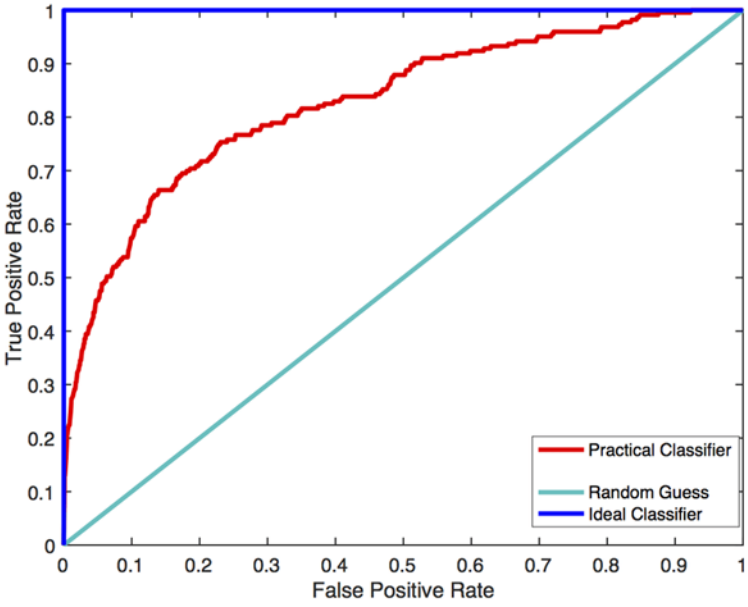


The bars represent the frequency of a given range of scores. On the basis of the graph on the left, it might make sense to select a cut-off value close to the median (about 0.55), allowing us to discriminate between two equally distributed "groups" (higher risk / lower risk). In the cases of the graph on the right, however, it would not make sense to adopt this approach, since the risk seems clustered in a much more limited fringe of the sample (above >0.7).

Receiver Operating Characteristic (ROC) curve and area under the ROC curve (AUC)

ROC curves are used to assess the accuracy of models that predict a **binary outcome**. In a ROC curve, the sensitivity (the true positive rate) of a risk model is plotted against the false positive rate (1 – specificity). A poorly performing model will have a ROC curve close to the diagonal, corresponding to a "random" prediction (green line in the figure below). A perfect model (achieving 100% sensitivity and 100% specificity at the same time, ROC area = 1) will see its ROC curve reach the upper left corner of the graph (blue line), meaning that it makes no false-positive predictions. **The closer the Area Under the Curve (AUC) to 1** (blue line in the figure), **the more accurate the model.**

Figure 8: Example of an ROC curve



Annex 2: Overview of the statistical methods mentioned in this report

The following section briefly presents key concepts that are useful for understanding this study and the case studies shared, but does not provide mathematical details about each method. Note that this is not an exhaustive overview and technical terms are avoided as much as possible.

Logistic regression

Logistic regression is used to **model binary variables** (taking on the values 0 or 1 – e.g. child *in* or *out* of hazardous child labour), with a non-linear function (inverse logit, or logit^{-1}) applied to its independent variables (predictors, in a risk modelling / prediction framework). When used to predict the probability of an event (e.g. child *in* CL), the **output** of a logistic regression is **continuous** and bounded from 0 to 1 (say, 0.55).

Logistic regressions use the mean of the predictors for the whole sample (**complete pooling**) to calculate their coefficients/weights (e.g. average distance to the road, average number of a primary schools in the communities). In this method, the **variations of the outcome of interest within the groups of the sample** (e.g. local variations at the cooperative or community levels) **are considered noise and** therefore **not modelled**.

That is, in a predictive/risk-modelling framework, a classical logistic regression will use **one series of predictor coefficients** and **one estimate of the variation** (standard error) of the outcome in the whole sample (global level).

Outcome of interest	**Predictors** (logit^{-1} of…)				
$Y =$	$X0_{\text{global}}$ +	$X1_{\text{global}}$ * a +	$X2_{\text{global}}$ * b +	$X3_{\text{global}}$ * c +	E_{global}

Simplified example of the parameters of a logistic regression model. $X0_{\text{global}}$ is the overall mean of the outcome of interest when $X1,2,3_{\text{global}} = 0$ (without the influence of $X1,2,3_{\text{global}}$). $X1,2,3_{\text{global}}$ are the respective means of the predictors in the whole sample, and a, b and c are the numeric values of the coefficients applied to $X1,2,3$ $_{\text{global}}$. E_{global} is the error of the model globally – that is, the mean difference between the predicted Y and the real/observed Y.

Linear regression

Linear regression is used to **summarise how the average value of a continuous outcome** (e.g. number of working hours performed by the child) **varies** according to a (set of) predictor(s). The **output** of a logistic regression is **continuous**. Therefore, when used to **predict** an outcome, linear regressions display continuous estimates.

Linear regressions use the means of the outcome of interest and of the independent variables in the whole sample (**complete pooling**) to calculate the coefficients/weights of each predictor (e.g. distance to the road, presence of a primary school in the community). In this method, the **variations of the outcome of interest within the groups of the sample** (e.g. local variations at the cooperative or community levels) **are considered noise and** therefore **not modelled**.

That is, in a predictive/risk-modelling framework, a classical linear regression will use **one series of predictor coefficients** and **one estimate of the variation** (standard error) of the outcome in the whole sample (global level).

Outcome of interest	Predictors				
$Y =$	$X0_{\mathrm{global}}$ +	$X1_{\mathrm{global}} * a$ +	$X2_{\mathrm{global}} * b$ +	$X3_{\mathrm{global}} * c$ +	E_{global}

Simplified example of the parameters of a linear regression model. $X0_{\mathrm{global}}$ *is the overall mean of the outcome of interest when* $X1,2,3_{\mathrm{global}} = 0$ *(without the influence of* $X1,2,3_{\mathrm{global}}$*).* $X1,2,3_{\mathrm{global}}$ *are the respective means of the predictors in the whole sample, and a, b and c are the numeric values of the coefficients applied to* $X1,2,3_{\mathrm{global}}$ *.* E_{global} *is the error of the model globally – that is, the mean difference between the predicted Y and the real/observed Y (contrast with multilevel regression below).*

Multilevel regression

Multilevel regressions can use, among other things, linear or logistic functions (see above). Compared to classical logistic or linear regressions, multilevel regressions model the outcome of interest by considering **simultaneously the information available at global level *and* the information available at group level(s)** (**variations *within*** the cooperatives and/or communities)**.** Therefore, multilevel regressions account for the structured/nested nature of the data: e.g. children within *households*, children interviewed by *enumerators*, children within *households* and benefitting from *different projects*.

That is, in a predictive/risk-modelling framework, a multilevel regression accounting for the variations in a child's hazardous child labour status at the community level, within 20 communities, will not use 1 but rather **21 series of parameters** (1 series for the global level and 20 for the communities) and **21 estimates of the variation** (standard errors) of the outcome among the target population and its sub-levels.

The fine-grained capture of these variations is the key feature of multilevel regressions. When there is very little variation across the groups, then the multilevel regression reduces to a classical (linear or logistic) regression. But when variations occur, the model's performance improves greatly. Likewise, when standard errors are high, the predicted value for a group is "pushed" toward the global (sample) mean. But when error is smaller, the prediction is "pulled" toward the mean of this specific group, thus creating **estimates sensitive to the specific environment and initial value of the outcome of each group**. This is why using multilevel regressions and using predictors with low standard error improve the performance of the risk models.

Outcome of interest	Predictors				
Y =	$X0_{global}$ +	$X1_{global}$ * a +	$X2_{global}$ * b +	$X3_{global}$ * c +	E_{global}
	$X0_{community\ 1}$ +	$X1_{community\ 1}$ * a +	$X2_{community\ 1}$ * b +	$X3_{community\ 1}$ * c +	$E_{community\ 1}$
	$X0_{community\ 2}$ +	$X1_{community\ 2}$ * a +	$X2_{community\ 2}$ * b +	$X3_{community\ 2}$ * c +	$E_{community\ 2}$
	$X0_{community\ 3}$ +	$X1_{community\ 3}$ * a +	$X2_{community\ 3}$ * b +	$X3_{community\ 3}$ * c +	$E_{community\ 3}$
	$X0_{community\ 20}$ +	$X1_{community\ 20}$ * a +	$X2_{community\ 20}$ * b +	$X3_{community\ 20}$ * c +	$E_{community\ 20}$

Simplified example of the parameters of a multilevel regression model accounting for the variations of the outcome of interest within 20 communities. $X0_{global}$ is the overall mean of the outcome of interest when $X1,2,3_{global} = 0$ (without the influence of $X1,2,3_{global}$). $X1,2,3_{global}$ are the respective means of the predictors at global (sample) level, and a, b and c are the numeric values of the coefficients applied to $X1,2,3_{global.}$ E_{global} is the error of the model globally, that is the mean difference between the predicted Y and the real/observed Y. $X0_{community\ 1,2,3,...\ 20}$ is the mean of the outcome of interest within community 1,2,…. 20 when $X1,2,3_{community\ 1,2,3,...\ 20} = 0$ (without the influence of $X1,2,3_{community\ 1,2,3,...\ 20}$). $X1,2,3_{community\ 1,2,...\ 20}$ are the respective means of the predictors within community 1,2,…. 20, and a, b and c are the numeric values of the coefficients applied to $X1,2,3_{community\ 1,2,...\ 20.}$ $E_{community\ 1,2,3,...\ 20}$ is the error of the model for each community, that is the mean difference between the predicted Y for each community and the real/observed Y in each community.

Extreme Gradient Boosting

Extreme Gradient Boosting is a machine-learning algorithm used to solve problems of classification (e.g. child *in* or *out* of child labour), ranking (e.g. children most exposed to long working hours *vs* children least exposed) or regression (e.g. number of hazardous tasks undertaken), by finding the best decision tree model for the data. This model is found by training the algorithm on a part of the dataset and iteratively assembling many high-error models fitted to many subsets of the data (weak learners) into a final high-performing model. The most difficult part of the dataset to predict is given more weight, and errors are used to indicate to the model the direction of needed improvements (gradients), which will help the model learn from its past errors.

Once the learning process is over, the model can predict new observations and returns the sensitivity/specificity (confusion matrix), as well as the overall accuracy of the model.

NB: no cut-off definition exercise is needed with this method.

Random Forest Classifier

Random Forest Classifier is a machine-learning algorithm used to solve problems of classification (e.g. child *in* or *out* of child labour). It is called "Forest" because it is comprised of numerous decision tree models created by randomly selecting sub-samples of a data set during the learning ("calibration") phase of the risk model construction. Each model's accuracy is assessed during the learning phase, and prediction is based on the "votes" of the most accurate decision trees, obtained by averaging all the predictions. In this method, there is therefore not one, but numerous averaged models, a fact that makes random forests more difficult to interpret than simple decision trees (which can be converted into rules).

Once the learning process is over, the model can predict new observations and returns the sensitivity/specificity (confusion matrix), as well as the overall accuracy of the model.

NB: no cut-off definition exercise is needed with this method.

Annex 3: Overview of components of a risk model

Components of a risk model				
Unit of observation / prediction	*Outcome*	*Predictors*	*Hierarchical levels of observations (unit of variation)*	*Statistical method or machine-learning technique*
Examples:				
• Child • Household • Community • Cooperative • Region	• Child in/out of hazardous child labour (binary variable) • Severity of hazardous child labour (on a scale from 0 to 10) • Prevalence of child labour within the community (%, i.e. continuous from 0 to 1)	• Age, gender (child level) • Income, mother's education, # children aged 5–17 (household level) • Distance to primary school, access to improved source of water (community level) • Management maturity (cooperative level)	• All pooled together • Region • Household	• Logistic regression • Linear regression • Latent class model • Multilevel regression • Random Forest Classifier • Extreme Gradient Boosting