Performance of predictive algorithms in estimating the risk of being a zero-dose child in India, Mali and Nigeria

Arpita Biswas,1 John Tucker,2 Sebastian Bauhoff3

ABSTRACT

Introduction Many children in low-income and middle-income countries fail to receive any routine vaccinations. There is little evidence on how to effectively and efficiently identify and target such ‘zero-dose’ (ZD) children.

Methods We examined how well predictive algorithms can characterise a child’s risk of being ZD based on predictor variables that are available in routine administrative data. We applied supervised learning algorithms with three increasingly rich sets of predictors and multiple years of data from India, Mali and Nigeria. We assessed performance based on specificity, sensitivity and the F1 Score and investigated feature importance. We also examined how performance decays when the model is trained on older data. For data from India in 2015, we compared the inclusion and exclusion errors of the algorithmic approach with a simple geographical targeting approach based on district full-immunisation coverage.

Results Cost-sensitive Ridge classification correctly classifies most ZD children as being at high risk in most country-years (high specificity). Performance did not meaningfully increase when predictors were added beyond an initial sparse set of seven variables. Region and measures of contact with the health system (antenatal care and birth in a facility) had the highest feature importance. Model performance decreased in the time between the data on which the model was trained and the data to which it was applied (test data). The exclusion error of the algorithmic approach was about 9.1% lower than the exclusion error of the geographical approach. Furthermore, the algorithmic approach was able to detect ZD children across 176 more areas as compared with the geographical rule, for the same number of children targeted.

Interpretation Predictive algorithms applied to existing data can effectively identify ZD children and could be deployed at low cost to target interventions to reduce ZD prevalence and inequities in vaccination coverage.

INTRODUCTION

Incomplete or delayed childhood vaccinations are a persistent concern in many low-income and middle-income countries (LMICs) despite considerable public health attention over the last decades. Coverage rates have stagnated and prior to the COVID-19 pandemic, there were about 14 million ‘zero-dose’ (ZD) infants globally who did not even receive a single shot. This number may have increased to 17 million during the pandemic. As a result, vaccine-preventable illnesses remain a leading cause of child morbidity and mortality in many countries, causing an estimated 1.5 million deaths annually for children under the age of five, mostly in Africa and Asia and among populations that are already vulnerable and disadvantaged.

Reaching children at risk of becoming ZD has become a national and global priority. However, identifying and targeting at-risk children has been challenging in practice, and there is little evidence on what approaches perform best. Many immunisation programmes target interventions to geographical areas with low historical immunisation rates. For example, India’s Intensified Mission Indradhanush (IMI) 2.0 in 2019–2020 was deployed at the district...
level, prioritising districts that had less than 80% ‘full-immunisation coverage’ according to survey data. This coarse geographical targeting of entire districts may unnecessarily intervene on children in targeted districts who are not at risk (inclusion error). Similarly, this approach may miss children who are at-risk of being ZD in districts that are not targeted because their full-immunisation coverage is above the intervention threshold (exclusion error). The latter group includes children from vulnerable groups who are at risk but live in districts with high average coverage. In this way, coarse geographical targeting may not be cost-effective and may reinforce inequalities.

Advances in statistical methods and data science in combination with increasingly available routine and administrative data, provide new opportunities to identify and target at-risk children at the subdistrict or individual level. For example, satellite data can be used to identify households to be targeted in campaigns, while geospatial modelling can inform the placement of vaccination sites. Recent research suggests that narrow targeting is feasible and effective for vaccinations. For example, machine learning methods can relatively accurately predict the risk that Pakistani children default on follow-up immunisation, and identify families and municipalities that are vaccine-hesitant. However, it is unclear how well these algorithms perform at identifying ZD children (also relative to geographical targeting), whether they need to be retrained regularly, and what information (predictors) is needed to generate sufficiently accurate risk predictions.

Our two objectives are to examine how well predictive algorithms can characterise an individual child’s risk of being ZD, and whether this risk can be reliably calculated with data from existing routine administrative reporting systems. We examine these questions empirically on data from seven Demographic and Health Surveys (DHS) in India, Mali and Nigeria in the mid-2000s and 2010s. India and Nigeria accounted for the largest and second-largest shares, respectively, of ZD children globally prior to the COVID-19 pandemic while India and Mali may have experienced the largest increases in ZD children between 2019–2020. We create three sets of covariates, starting with information commonly available in routine data and adding contextual and individual information. We examine how well four supervised learning algorithms can identify at-risk children according to the algorithms’ sensitivity, specificity and accuracy. For the best-performing algorithm, cost-sensitive Ridge classification (RC), we further examine the relative contribution of the covariates to its performance. We further examined whether the best-performing model needs to be retrained by assessing how well models trained on older data can identify ZD children in the most recent data. Finally, we compare our predictions against a geographical targeting rule inspired by India’s IMI 2.0 that targets entire districts with less than 80% full immunisation coverage in the India 2015 DHS data, which contain district identifiers.

Methods
We examined seven DHS data sets across various country-years—India (2006, 2015 and 2020), Mali (2006 and 2018) and Nigeria (2008 and 2018). The DHS are generally based on nationally and regionally representative samples that are drawn using stratified two-stage cluster designs. The first stage is enumeration areas (usually geographical areas, such as villages or city blocks) and the second stage is households within each enumeration area. The primary respondents are women aged 15–49 years. The survey methods and questionnaires are largely standardised across countries and time, and the data are publicly and freely available. We complemented the survey data with selected geospatial covariates for each enumeration area that are available from DHS for the 2018 surveys in Mali and Nigeria, and the 2015 and 2020 surveys in India. (We used spatially harmonised region and state variables provided by the Integrated Public Use Microdata Series (IPUMS) DHS for all data sets except India 2020. We manually harmonised the India 2020 state names to align with those in the India 2015 data set. The India 2006 data set contains fewer states than the later data sets, because of the creation and merging of states.) We extracted information on immunisation status for children aged 12–23 months following the standard DHS approach, based on either vaccination cards or, if absent, the report of the child’s mother or caregiver. (We focus on the most recent birth, for which the DHS collects information about antenatal care visits and timing.) Following the operational definition used by Gavi, we defined a variable VS (vaccination status) for each child which is set to 1 if the child received the first dose of the diphtheria, tetanus toxoids and pertussis vaccine (DPT1), and set to 0 otherwise. We defined a child as ZD if their corresponding VS value is 0. Thus, VS=1 ZD. We used VS as the output variable of the prediction problem. Thus, a prediction of value 1 indicates that a child will receive at least the DPT1 vaccine and a prediction of value 0 indicates that a child will be ZD.

For the comparison of the algorithmic approach with geographical targeting, we constructed a district-level coverage measure similar to that of IMI’s ‘full-immunisation coverage’ Score. Specifically, we created a binary variable equal to 1 if a child received DPT1. We used survey weights to generate the district-level scores; unlike other DHS surveys, the 2015 and 2020 India DHS are representative at the district level and contain district identifiers. Under the geographical rule, a district will be targeted if this score is less than 80%.

We created three sets of predictor variables to capture characteristics that have been associated with ZD status (table 1, for summary statistics and detailed definitions see online supplemental appendix table 2). The sparse predictor set includes variables that we derived from the DHS but that are typically available in routine data systems such as the District Health Information System platform, Reproductive and Child Health, and Expanded Program on Immunization databases. This includes information...
on the place of residence (state and urban or rural location), whether the child was delivered at a healthcare facility and the number of antenatal visits of the mother. The second predictor set adds contextual information of the place of residence that is likely available in most settings from external sources, such as population density, travel time, and urbanicity and economic activity using average night-time luminosity as a proxy. These data contain values for 2015. The third predictor set further adds information from the DHS that are typically not available in routine data, such as the mother’s education and household wealth quintile. These predictors have previously been associated with vaccine take-up.15 16 Our base models do not include vaccination coverage rates for districts or regions, as these data usually require dedicated surveys (as, eg, in the case of India’s IMI 2.0). We add this predictor below when comparing the algorithmic and geographical targeting rules.

### Data preprocessing and design of the training data set

We implemented two preprocessing steps. First, we one hot encoded all the categorical features, including the region identifier, by creating a new variable for every categorical value. For example, population density is replaced by three new binary (0/1) variables based on the degree of urbanisation, where 1 represents that the child belongs to that category.17 Similarly, we created 11 categories (zero and deciles) for the night light composite.18 This encoding ensures that we obtain meaningful predictions from the algorithms and avoid creating fictional ordinal relationships in the data.19 Second, we removed the children with missing information for at least one of the predictors. There were around 265 (out of 40 555) in the India 2020 data set, 232 (out of 46 209) children in the India 2015 data set, 60 (out of 1803) in the Mali 2018 data set, and 58 (out of 2350) in the Nigeria 2018 data set, with one or more missing values. Table 2 shows the analytical data sets.

We then created a training data set with a 70% sample of the overall data set that maintains the regional stratification used by DHS to account for heterogeneity across regions. Specifically, we drew a random sample of children within each state or regional stratum that is proportional to the total number of children in that stratum (proportional-to-size). After creating the training dataset, the remaining samples were then used for testing how the models perform on completely unseen data points.

### Model training

After the preprocessing, all the information of a child (data point) is converted to a vector form (with \( d \) values, one of each predictor). The outcome variable (also known as class label) corresponding to a data point is the binary (0/1) variable VS, where VS=1 denotes a non-ZD child and VS=0 denotes a ZD child. We used supervised learning algorithms to learn a model from the training data that predicts the outcome variable of an unknown data point in the test data, as correctly as possible (figure 1). Specifically, we created prediction models using four classical supervised learning algorithms to predict whether a child is at the risk of being ZD, namely, RC,20 Decision Tree,21 Nearest Neighbor 22 and Multi-Layer Perceptron.23 These predictive models output a predicted score between 0 and 1 for each data point (child), which is then classified into ZD or non-ZD using a threshold.
Figure 1 Input and output format of the prediction model

\[
predictor_1 \times predictor_2 \times \cdots \times predictor_d
\]

value \( x \in [0, 1] \), such that any child with a score of at least \( x \) is predicted as being non-ZD and a child with score less than \( x \) is predicted as ZD. We assume the threshold to be 0.5, unless specified otherwise.

Although these models have been well studied for binary classification tasks, the resulting predictions often incur low precision in the presence of class imbalance. Class imbalance occurs when the outcome variable appears in disproportionately higher frequency for one class (in our case, non-ZD) over the other class (ZD). In the presence of a large class imbalance, the models overfit to the majority class. For example, since only 6.7% of children are ZD (that is, have received at least one vaccination) in the India 2015 data set, a naïve model trained with this data set may tend to predict non-ZD for almost everyone to maximise the accuracy. That is because predicting non-ZD status for every single child would achieve a 93.3% accuracy rate. This high accuracy rate is misleading and particularly problematic in the context of our objective because misclassifying a ZD child as a non-ZD is a more serious error than misclassifying a non-ZD child as ZD (that is, exclusion error is more serious than inclusion error). This issue arises because supervised learning algorithms aim to minimise the average misclassification cost function. We accounted for this concern by using cost-sensitive learning, which modifies the cost function to associate different penalties on different types of misclassifications.

After using cost-sensitive analysis on the models, we tuned the hyperparameters of each model. For each method, we ran the model with different combinations of hyperparameters. For example, Ridge Classifier has a hyperparameter, called learning rate, that acts as a penalty on the sum of squares of the coefficients. We tested the performance of the model across various values of the hyperparameters to find the hyperparameter combination that performs the best. We used fivefold cross validation for hyperparameter tuning. Fivefold cross validation works by first randomly splitting the training data into five parts. In each round, four parts are used for training the model along with cost-sensitive learning, and the fifth part is used for validation, that is, testing how well the model performed. We then selected the best performing hyperparameters for training the final model.

Evaluation
We evaluated model performance based on specificity, sensitivity and F1 Score by comparing our classification of being ZD against the true ZD status of each child. Specificity denotes the fraction of individuals correctly classified as ZD children among all ZD children. In our application, specificity is the most important performance measure as the policy objective is to correctly identify as many ZD children as possible. Sensitivity is the fraction of correctly classified non-ZD children among all non-ZD children. A high sensitivity is not critical in our application, but desirable, for example, to avoid deploying costly interventions to children who are at low risk of ZD. The F1 Score is a commonly used performance metric for classification tasks, particularly when the data are imbalanced as is the case in our application. This metric is preferred over accuracy as it provides a more comprehensive evaluation of the model’s performance in class imbalance scenarios. Online supplemental appendix table A2 provides the mathematical expressions to compute these performance metrics.

We conducted three additional analyses. First, we examined whether the best-performing model needs to be retrained by assessing the performance of models trained on older data. For this purpose, we trained on the older data sets (Mali 2006, Nigeria 2008, India 2006 and India 2015) and tested prediction power on the newer data sets (Mali 2018, Nigeria 2018, India 2020). Second, we used the best-performing machine learning model for each individual country-year data set to identify the most important features in identifying ZD children. Third, we compared the performance of a simple geographical rule (target districts with less than 80% full immunisation coverage) to a prediction model that includes the full immunisation coverage as a predictor.

Human subjects
The institutional review board of the Harvard T H Chan School of Public Health reviewed the study protocol and determined that it does not represent ‘not human subjects’ research (IRB21-1521).

Patient and public involvement
This research did not have patient or public involvement.

RESULTS
Online supplemental appendix table A2 shows the summary statistics for the various predictors.

Performance of the prediction models
The cost-sensitive RC model either performed the best or nearly the best for all country-years and across all feature sets, in terms of specificity. Online supplemental appendix table A3 shows the performance of all the classifiers across various country-years. Using sparse predictor sets for India 2020, the specificity of cost-sensitive decision tree algorithm produced highest specificity of 0.5, while RC produced a specificity of 0.49, comparable to
district targeting. For Nigeria 2006, Nearest Neighbor obtains highest specificity. Except for these cases, we find that the RC algorithm consistently performed well across almost all the data sets.

In general, prioritising specificity worsens sensitivity, F1 Score and accuracy. This trade-off is apparent in online supplemental appendix figure A1 where we used RC and incrementally changed the cost of misclassifying ZD children as non-ZD using the India 2015 data set. The more we penalise this type of misclassification, the higher the specificity and lower the other metrics. The model gives a balanced performance over various metrics when the ZD misclassification penalty was set to a value equal to the fraction of non-ZD children and the non-ZD misclassification penalty was set to a value equal to the fraction of ZD children in the training data set.

Expanding the predictor set beyond the sparse feature set improved the performance on some data sets and did not improve much on others (table 3). For example, for the India 2015 data set, RC produced a specificity of 0.54 with only sparse feature set and the performance improved marginally to 0.61 after including additional features. Whereas for the India 2006 data set, RC produced a specificity of 0.78 with only sparse features, and the specificity slightly decreased to 0.77 after including additional features. This observation is consistent with the feature importance results reported below, since unequivocally the most important features are already contained in the sparse set, such as the place of residency (region), facility delivery information, and antenatal care frequency.

We further observe that the higher the class imbalance in a data set, the more difficult it is to improve the specificity of the predictions. For example, the unweighted fraction of ZD children in the India 2020 data set is only 6.6%, and thus, even after using a sufficiently high penalty for misclassification of ZD, RC could achieve a specificity of only around 0.50. On the other hand, for the Nigeria 2018 data set, where the fraction of ZD children is 30.6%, the best performing algorithm (RC) obtained a high specificity of around 0.78. Thus, the extent of class imbalance is a determining factor in the performance of the prediction algorithms.

Performance over time
We also find that models trained on relatively older data have a lower specificity if applied to test data collected a few years later. Table 4 shows the performance of models trained on the India 2006 and 2015 data when they are applied to test data from India 2015. The RC model trained on India 2015 data and tested on India 2020 data achieved a specificity of 0.43 with additional predictors, which is quite low as compared with the performance of the model trained on India 2020 (with a specificity of 0.52, as shown in row 1 of table 3). However, the performance drops even further when the model is trained on India 2006 data and tested on India 2020 data. We provide the comparison between new and old models for Mali and Nigeria in online supplemental table A4.

Table 3: Performance of the algorithm (cost-sensitive Ridge classification algorithm), tested on 30% data

<table>
<thead>
<tr>
<th>Share of ZD children</th>
<th>Se</th>
<th>Spec</th>
<th>F1</th>
<th>Se</th>
<th>Spec</th>
<th>F1</th>
<th>Se</th>
<th>Spec</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>India 2020</td>
<td>0.066</td>
<td>0.73</td>
<td>0.79</td>
<td>0.47</td>
<td>0.82</td>
<td>0.51</td>
<td>0.73</td>
<td>0.82</td>
<td>0.51</td>
</tr>
<tr>
<td>India 2015</td>
<td>0.112</td>
<td>0.77</td>
<td>0.84</td>
<td>0.77</td>
<td>0.87</td>
<td>0.67</td>
<td>0.77</td>
<td>0.87</td>
<td>0.67</td>
</tr>
<tr>
<td>India 2006</td>
<td>0.215</td>
<td>0.69</td>
<td>0.78</td>
<td>0.54</td>
<td>0.74</td>
<td>0.59</td>
<td>0.54</td>
<td>0.74</td>
<td>0.59</td>
</tr>
<tr>
<td>Mali 2018</td>
<td>0.199</td>
<td>0.71</td>
<td>0.79</td>
<td>0.68</td>
<td>0.77</td>
<td>0.69</td>
<td>0.68</td>
<td>0.77</td>
<td>0.69</td>
</tr>
<tr>
<td>Mali 2006</td>
<td>0.171</td>
<td>0.71</td>
<td>0.79</td>
<td>0.71</td>
<td>0.77</td>
<td>0.72</td>
<td>0.71</td>
<td>0.77</td>
<td>0.72</td>
</tr>
<tr>
<td>Nigeria 2018</td>
<td>0.306</td>
<td>0.77</td>
<td>0.81</td>
<td>0.78</td>
<td>0.81</td>
<td>0.68</td>
<td>0.78</td>
<td>0.81</td>
<td>0.68</td>
</tr>
<tr>
<td>Nigeria 2008</td>
<td>0.502</td>
<td>0.77</td>
<td>0.81</td>
<td>0.77</td>
<td>0.77</td>
<td>0.77</td>
<td>0.77</td>
<td>0.77</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Notes: The share of ZD children is unweighted.

The contextual cluster-based information is not available for the India 2006, Mali 2006 and Nigeria 2008 data sets. Thus, the columns ‘adding survey-based information’ for those three data sets, do not contain cluster-based information. The complete table with results of all the classifiers is in online supplemental table A1.
Feature importance

Overall, the region of residence along with whether the child was delivered at a facility and whether the mother received antenatal care were the most important features across all data sets. Figure 2 shows that for the India 2015 data set, there are some regions positively correlated with the VS=1 ZD variable, that is, with a low risk of being ZD. For example, the state Sikkim received the highest positive importance (coefficient). On the other hand, the indicator that the mother received no antenatal care had the most negative coefficient, that is, this indicator was strongly correlated with being ZD. Another feature that has a negative coefficient is low night light, a proxy for low economic activity.18 The importance of each feature changes across different data sets. We provide the feature importance of models trained on different data sets using the RC model in online supplemental appendix figure A2. In all the feature importance graphs, the delivery of a child in a health facility is positively correlated with non-ZD and the absence of antenatal care of the mother is negatively correlated with non-ZD.

Comparison with district-level targeting

We compared the performance of the machine learning approach with a simple geographical rule that targets entire districts with less than 80% full-immunisation coverage. For this comparison, we examine a RC model trained using only the sparse predictors along with the continuous district-level DPT1 coverage rate as additional predictor. Note that the previous analyses did not include the coverage rate because only the India data include district identifiers. Moreover, the RC model is trained on 70% of the India 2015 data set and tested on the full India 2015 data set. Note that the previous analyses (in Table 3 and online supplemental table A3) were tested on 30% of the data rather than the full data set.

Figure 3 provides a graphical comparison of the RC and the geographical rule, applied on the entire data from

<table>
<thead>
<tr>
<th>Table 4</th>
<th>Performance on India 2020 (test data) when the model is trained using data sets from 2006 and 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Train data</strong></td>
<td><strong>Administrative data</strong></td>
</tr>
<tr>
<td></td>
<td>Sensitivity</td>
</tr>
<tr>
<td>India 2015</td>
<td>0.82</td>
</tr>
<tr>
<td>India 2006</td>
<td>0.80</td>
</tr>
</tbody>
</table>
India in 2015. Each point in the scatter plot represents a child, with blue colour denoting non-ZD (VS=1) and red colour denoting ZD (VS=0). The horizontal axis shows the predicted risk score of a child (a higher score implies more chance of receiving any vaccinations and less likely to be ZD) and the vertical axis shows the district-level DPT1 Immunisation Score, calculated from the source data using survey weights. The geographical approach targets all children in districts with less than 80% DPT1 coverage, that is, below the horizontal line at 80%. If the model predictions are used for targeting, all children on the left side of the vertical line at 0.5 will be targeted. Children in the bottom-left quadrant are predicted to have a high risk of ZD and are also targeted by the geographical rule, and the children in the top-right quadrant are predicted as low risk and are not targeted by the geographical rule. Children in the bottom-right quadrant are predicted as low risk of ZD but are targeted by the geographical approach.

To compare the two targeting policies, we computed the inclusion and exclusion errors, along with specificity and sensitivity of the decisions against the true VS. Table 5 shows that the exclusion error of the geographical targeting rule is 58.53%. The exclusion error with the prediction model depends on the cut-off that is used to classify children as ZD. A cut-off of 0.423 would target the same number of children as the geographical approach and has an exclusion error of 53.2%, which is 9.1% lower than that of the geographical targeting rule. For the same number of targeted children, the algorithmic approach targets more districts. Online supplemental appendix figure A3 shows the change in targeted districts as the number of targeted children increases for the geographical and algorithmic approaches. The former increases almost linearly whereas the prediction approach is quickly increasing but at a decreasing rate. Moreover, this prediction model improves the inclusion error as compared with that of the geographical rule. Both approaches have high specificity, indicating a low number of misclassified ZD children (high inclusion error).

**DISCUSSION**

We applied supervised learning algorithms to seven data sets from three countries—India, Mali and Nigeria—to assess how well they predict whether a child is likely to be ‘zero-dose’. Our findings suggest that these algorithms perform well at identifying at-risk children at the individual level even with a sparse set of predictors that can be derived from existing administrative data. As an example, our preferred algorithm—cost-sensitive RC with a 0.5 cut-off threshold in table 3—correctly identifies about 54% of children who are ZD in survey data from India in 2015 using six predictors that are commonly contained in pregnancy or birth registers. Further, we found that the cut-off threshold of the RC classifier is 0.423 when we restrict the number of children targeted to be exactly 2012 (to match with the number of children targeted by the geographical rule). We then observed that the algorithmic approach has lower exclusion and
incorporating algorithms into existing data systems to algorithmically targeting specificity. These results suggest scope for a commendable performance compared with geographically, more than 50% of the ZD children are correctly identified as such, which we consider than 0.5 indicates that more than 50% of the ZD children are considered non-ZD instead of ZD, which is excessively low. Our approach seeks to achieve specificity and sensitivity, ensuring that neither measure is excessively low. Our approach seeks to achieve specificity values ranging from 0.47 to 0.8. A specificity greater than 0.5 indicates that more than 50% of the ZD children are correctly identified as such, which we consider a commendable performance compared with geographically targeting specificity. These results suggest scope for applying these methods to identifying ZD children and incorporating algorithms into existing data systems to generate risk scores for each individual child. This can inform vaccination campaigns and targeted outreach, for example, by directing community health workers to the household or deploying individualised text messages or phone calls. Overall, the region of residence, along with proxies of contact with the health system (the frequency of antenatal visits and whether the child is delivered at a healthcare facility) are the most important features to identify at-risk children. We also find that the algorithm’s performance decreases when the training and test data are collected from different years, suggesting that the algorithms need to be retrained with relatively recent data at regular intervals to ensure maximal performance; occasionally the algorithms themselves may need to be examined and replaced.

Our analysis also illustrates challenges and choices with applying supervised learning to identify ZD children in practice. First, we find that the algorithm’s performance decays when the training and test data are collected from different years, suggesting that they need to be retrained periodically with newer data sets to ensure maximal performance. Second, we observed that the data sets are highly class-imbalanced because there are relatively few ZD children. It is therefore crucial to use cost-sensitive training to penalise the misclassification of ZD to non-ZD at a higher cost than that of the misclassification of non-ZD to ZD. Otherwise, the model predicts every child as non-ZD, producing a specificity of zero because all the ZD children are misclassified. Relatedly, the analyst can choose the form of the penalty to match the desired sensitivity, specificity and F1 Score, but this choice involves trade-offs. Third, in using the risk scores, the analyst can

<table>
<thead>
<tr>
<th>Targeting rule</th>
<th>Number of children targeted</th>
<th>Number of districts targeted</th>
<th>Exclusion error (%)</th>
<th>Inclusion error (%)</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geographical rule</td>
<td>2012</td>
<td>93</td>
<td>58.53</td>
<td>67.99</td>
<td>0.89</td>
<td>0.41</td>
<td>0.91</td>
</tr>
<tr>
<td>RC (cut-off at 0.3)</td>
<td>230</td>
<td>71</td>
<td>90.85</td>
<td>38.26</td>
<td>0.99</td>
<td>0.09</td>
<td>0.94</td>
</tr>
<tr>
<td>RC (cut-off at 0.4)</td>
<td>1470</td>
<td>232</td>
<td>60.20</td>
<td>57.96</td>
<td>0.93</td>
<td>0.40</td>
<td>0.93</td>
</tr>
<tr>
<td>RC (cut-off at 0.423)</td>
<td>2012</td>
<td>269</td>
<td>53.2</td>
<td>63.87</td>
<td>0.89</td>
<td>0.47</td>
<td>0.91</td>
</tr>
<tr>
<td>RC (cut-off at 0.5)</td>
<td>4610</td>
<td>401</td>
<td>27.37</td>
<td>75.53</td>
<td>0.72</td>
<td>0.73</td>
<td>0.82</td>
</tr>
<tr>
<td>RC (cut-off at 0.6)</td>
<td>8664</td>
<td>561</td>
<td>8.44</td>
<td>83.59</td>
<td>0.41</td>
<td>0.92</td>
<td>0.58</td>
</tr>
<tr>
<td>RC (cut-off at 0.7)</td>
<td>12534</td>
<td>640</td>
<td>0.77</td>
<td>87.71</td>
<td>0.10</td>
<td>0.99</td>
<td>0.18</td>
</tr>
</tbody>
</table>

*RC is trained on 70% of India 2015 and the results are on the entire Indian 2015 data set. Also, we evaluate RC with cut-off at 0.423 since it yields the same number of children targeted with the geographical rule. The cost-sensitive RC uses the sparse covariate set with total vaccination coverage (used by IMI) as an additional predictor.

IMI, Intensified Mission Indradhanush; RC, Ridge Classification.

Our aim was to optimise specificity while still maintaining a reasonably high sensitivity. In scenarios where the number of ZD children is substantially lower than non-ZD children, a simplistic machine learning model that classifies all instances as non-ZD could yield a high accuracy and sensitivity of 1, but an undesirable specificity of 0. This extreme situation necessitates a more nuanced approach. On the other hand, simply maximising specificity could lead to a model that classifies all instances as ZD, achieving a specificity of 1 but rendering poor overall accuracy. To achieve a balanced prediction scenario for ZD children, we employ a cost-sensitive model. These models are designed to strike a favourable trade-off between sensitivity and specificity, ensuring that neither measure is excessively low. Our approach seeks to achieve specificity values ranging from 0.47 to 0.8. A specificity greater than 0.5 indicates that more than 50% of the ZD children are correctly identified as such, which we consider a commendable performance compared with geographically targeting specificity. These results suggest scope for applying these methods to identifying ZD children and incorporating algorithms into existing data systems to

Table 5  Comparing prediction model against geographical rule in terms of the number of districts covered (out of 640 districts), inclusion error, exclusion error and other performance metrics

<table>
<thead>
<tr>
<th>Targeting rule</th>
<th>Number of children targeted</th>
<th>Number of districts targeted</th>
<th>Exclusion error (%)</th>
<th>Inclusion error (%)</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geographical rule</td>
<td>2012</td>
<td>93</td>
<td>58.53</td>
<td>67.99</td>
<td>0.89</td>
<td>0.41</td>
<td>0.91</td>
</tr>
<tr>
<td>RC (cut-off at 0.3)</td>
<td>230</td>
<td>71</td>
<td>90.85</td>
<td>38.26</td>
<td>0.99</td>
<td>0.09</td>
<td>0.94</td>
</tr>
<tr>
<td>RC (cut-off at 0.4)</td>
<td>1470</td>
<td>232</td>
<td>60.20</td>
<td>57.96</td>
<td>0.93</td>
<td>0.40</td>
<td>0.93</td>
</tr>
<tr>
<td>RC (cut-off at 0.423)*</td>
<td>2012</td>
<td>269</td>
<td>53.2</td>
<td>63.87</td>
<td>0.89</td>
<td>0.47</td>
<td>0.91</td>
</tr>
<tr>
<td>RC (cut-off at 0.5)</td>
<td>4610</td>
<td>401</td>
<td>27.37</td>
<td>75.53</td>
<td>0.72</td>
<td>0.73</td>
<td>0.82</td>
</tr>
<tr>
<td>RC (cut-off at 0.6)</td>
<td>8664</td>
<td>561</td>
<td>8.44</td>
<td>83.59</td>
<td>0.41</td>
<td>0.92</td>
<td>0.58</td>
</tr>
<tr>
<td>RC (cut-off at 0.7)</td>
<td>12534</td>
<td>640</td>
<td>0.77</td>
<td>87.71</td>
<td>0.10</td>
<td>0.99</td>
<td>0.18</td>
</tr>
</tbody>
</table>

*RC is trained on 70% of India 2015 and the results are on the entire Indian 2015 data set. Also, we evaluate RC with cut-off at 0.423 since it yields the same number of children targeted with the geographical rule. The cost-sensitive RC uses the sparse covariate set with total vaccination coverage (used by IMI) as an additional predictor.

IMI, Intensified Mission Indradhanush; RC, Ridge Classification.
also choose the threshold at which a child with a (continuous) risk score is classified as non-ZD or ZD. This leads to a trade-off between the exclusion and inclusion errors. With regards to reaching ZD children, the threshold should be set to reduce the exclusion error, that is, the probability of erroneously classifying a ZD child as being non-ZD. Finally, the algorithm with a sparse predictor set and full immunisation coverage still performs better, and the few predictors are available in routine data. Thus, even in our application, it would be valuable and low-cost to choose the algorithmic approach.

Our analysis has several limitations. First, we examined performance on seven data sets from three countries, and most surveys predate the COVID-19 pandemic. The prevalence and patterns of ZD may have changed since the data were collected, and the data are not representative of all LMICs. Relatedly, the DHS sampling frames may have excluded communities with high number of ZD children, for example, informal settlements that developed after the sampling frame was established. Model performance could be different if these communities were included in the data. However, in practice a larger challenge may be the fundamental lack of data that is required for risk prediction. Second, we could only compare the exclusion and inclusion errors of the algorithmically informed and geographical targeting, but not the costs of reaching a ZD child under each approach and for different interventions.

Reaching children at risk of not receiving routine vaccinations is critical to improve vaccination coverage and reduce inequity. Our analysis illustrates that tools from data science can facilitate identifying and targeting ZD children. Because the required data are often already available in routine administrative data, algorithmic approaches are feasible and could be implemented at low cost (possibly in combination with other approaches), and could help improve vaccination coverage and reduce inequities. Further testing of this approach on routine data systems is needed to assess whether algorithmic approaches perform well in real-life settings.

Acknowledgements The authors thank Gustavo Correa and Agnes Watsemba for helpful comments.

Contributors AB: conceptualisation, software, formal analysis, visualisation, writing of the original draft, writing of the review and editing, supervision. JT: software, formal analysis, writing of the review and editing. SB: conceptualisation, writing original draft, writing of the review and editing, supervision, project administration. All authors attest they meet the ICMJE criteria for authorship. SB: writing original draft, writing of the review and editing, supervision, project administration. JT: conceptualisation, software, formal analysis, writing of the original draft, writing of the review and editing. SB: conceptualisation, software, formal analysis, writing of the review and editing. JT: conceptualisation, software, formal analysis, writing of the original draft, writing of the review and editing. SB: conceptualisation, software, formal analysis, writing of the review and editing. JT: conceptualisation, software, formal analysis, writing of the original draft, writing of the review and editing. SB: conceptualisation, software, formal analysis, writing of the review and editing.

Funding This work was supported by GAVI, the Vaccine Alliance.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. The data are available at www.dhsprogram.com and www.idhsdata.org/dhls. The analysis code can be found on GitHub at https://github.com/arpita-biswas/zero-dose-child-vaccination-targetting.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iD Sebastian Bauhoff http://orcid.org/0000-0001-6017-4991

REFERENCES


