Balancing Trade-Offs in the Detection of Primary Schools at Risk

Savi, Alexander O.; Cornelisz, Ilja; Sjerps, Matthias J.; Greup, Steffen L.; Bres, Chris M.; Klaveren, Chris van

published in
Educational Measurement : Issues and Practice
2021

DOI (link to publisher)
10.1111/emip.12433

document version
Publisher's PDF, also known as Version of record

citation for published version (APA)
Savi, A. O., Cornelisz, I., Sjerps, M. J., Greup, S. L., Bres, C. M., & Klaveren, C. V. (2021). Balancing Trade-Offs in the Detection of Primary Schools at Risk. *Educational Measurement : Issues and Practice*, *40*(3), 110-124. https://doi.org/10.1111/emip.12433

Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Fall 2021, Vol. 40, No. 3, pp. 110–124

Balancing Trade-Offs in the Detection of Primary Schools at Risk

Alexander O. Savi and Ilja Cornelisz, *Amsterdam Center for Learning Analytics, VU University Amsterdam,* Matthias J. Sjerps and Steffen L. Greup, *Dutch Inspectorate of Education, Utrecht,* Chris M. Bres, *Dutch Inspectorate of Education, Utrecht, Dutch Human Environment and Transport Inspectorate, Utrecht,* and Chris van Klaveren, *Amsterdam Center for Learning Analytics, VU University Amsterdam*

Abstract: *The quality assurance and evaluation of primary schools requires early risk detection. This is a daunting task, not only because risks are typically rare and their origins complex, but also because governing institutions have limited resources and capacity and desire efficiency and proportionality. Many countries, including most Organisation for Economic Co-operation and Development countries, use inspections to detect schools at risk. In order to aid these efforts, we first evaluate various case-based prediction models, and then propose a principled exploit–explore procedure for organizing school inspections. We demonstrate these methods using data from the Dutch Inspectorate of Education, which monitors the regulatory compliance of roughly 6,000 primary schools. The approach has the potential to balance the benefits of prioritizing inspections of predicted high-risk schools on the one hand, with the benefits of verifying predicted risks and causal impact evaluations of school inspections on the other hand.*

Keywords: case-based prediction, exploration–exploitation trade-off, inspection policy, precision–recall trade-off, predictive modeling, primary education, school accountability

The pivotal importance of primary education is uncontroversial (UNESCO, 2017). While some countries struggle to provide the most basic access (UNESCO Institute for Statistics and UNICEF, 2015), other countries seek to maintain and improve the quality of their educational institutions. Once structural educational governance is established, most countries pursue to assure the quality of their educational institutions by means of accreditation, inspections, or both. These approaches are hard to universally define, are heavily influenced by national traditions (Grek, Lawn, Ozga, & Segerholm, 2013), and may overlap considerably. Nevertheless, accreditation generally functions to establish *a priori sufficient quality*, whereas inspections generally function to detect *a posteriori insufficient quality*. The early detection of schools at risk is key to such quality assurance, where schools at risk are defined as schools that fail to deliver the demanded quality. The current study targets this issue with a twofold aim. We first evaluate the predictive power of case-based predictions of primary schools at risk. Case-based predictions exploit statistically learned patterns in associations between characteristics of historical cases of schools at risk and schools not at risk. We then establish a principled procedure for prioritizing inspections on the basis of risk predictions. In the following sections, we delve into these aims in detail.

Risk-Based Inspection

Inspections, colloquially defined as the activities aimed at monitoring the quality of the institutions or activities of interest, are particularly well suited for early risk detection. Many countries that participate in the Organisation for Economic Co-operation and Development (OECD) perform regular educational inspections (Faubert, 2009). The United States is a notable exception, despite inspections being a topic of discussion (e.g., Berner, 2017; Ladd, 2010) and the use of similar "school quality reviews" in New York City and Massachusetts (Ladd, 2016). Meanwhile, in European countries, educational inspections are commonplace, with inspectorates of education united in, and supported by the Standing International Conference of Inspectorates (Grek et al., 2013; Grek & Lindgren, 2014). The importance of prediction is well recognized, with the United Kingdom's inspectorate having studied case-based models for predictions of risk (Sanders, Lawrence, Gibbons, & Calcraft, 2017) and the Dutch inspectorate employing a rule-based model for predictions of risk.

The inspection method of the Dutch Inspectorate of Education (hereafter inspectorate) is taken as an example to materialize the aims of this study. The inspectorate is, among other things, responsible for monitoring the quality of primary education. In a narrower sense, quality is defined by

the institutions' regulatory compliance, of which the inverse is termed risk. As such, schools at risk fail to deliver sufficient quality, substantiated by their regulatory noncompliance. This is not to say that risk detection is straightforward. Regulations are influenced by politics, evolve over the years, and are captured in extensive legislation. In order to put flesh on the bones of quality and compliance, the inspectorate captured Dutch legislation in the Inspection Framework Primary Education 2017 (available in English; Dutch Inspectorate of Education, 2019). This framework distinguishes five areas of interest, so-called *quality areas*, and stipulates the basic quality requirements, so-called *standards*. The quality areas comprise the educational process, school climate, learning outcomes, quality assurance and ambition, and financial management. Historically, the learning outcomes carry substantial weight in the inspection process.

In this study, we depart from the inspectorate's current risk-based inspection (RBI) method, which was adopted in 2007. In their RBI method, the *degree* of monitoring is determined by the *predicted* risk. A brief description of the RBI method clarifies the general procedure. Following current legislation, every school is inspected once every 4 years, although the degree and substance of such inspections depends on the predicted risk and may vary from school to school. On top of that, schools may be inspected on the basis of yearly risk predictions. These predictions have two stages. First, a set of indicators is used to predict schools at risk. Currently, the inspectorate uses *rule-based* predictions of schools at risk: hypothesized relations between possible risk factors and actual risk. The indicators and weights of the algorithm, as well as the thresholds, are determined by experts from the inspectorate. This rule-based algorithm has been in continuous development, and has substantially changed over the past decade. Ultimately, schools with a high-risk prediction are subjected to a desk analysis by an expert. If the expert analysis echoes the prediction, the school is inspected. The inspectorate's ultimate judgments are published and can have considerable impact on the judged schools. In this study, the inspectorate's historic judgments function as the ground truth of the regulatory compliance of primary schools.

The RBI method is rooted in a desire to increase the efficiency and proportionality of school visits and cope with (increasingly) limited capacity and resources. Indeed, submitting all schools to regular and rigorous risk assessments is often unfeasible due to resource capacity constraints. Notably, even if regular and rigorous assessments would be feasible, schools can be at risk for profoundly different reasons. It is precisely these characteristics that pose considerable challenges to the detection of schools at risk, and that obstruct efforts to warrant the demanded quality of education across primary schools. It is because of these challenges that the inspectorate's ultimate judgments are not expected to be without error. However, the judgments are arguably the best measurement of risk available, and provide a considerable benefit over countries that do not perform such rigorous measurements. Also, it is because of these challenges that the inspectorate desires predictions of risk, allowing a prioritization of inspections.

The RBI method creates additional challenges. Similar to the judgments, predictions of risk are generally not without error, and schools incorrectly receiving low risk predictions may go undetected when not inspected. Moreover, organizing inspections solely at schools with high-risk predictions obstructs the ability to verify the prediction model and evaluate the impact of school inspections. Therefore, we suggest a more principled approach for prioritizing inspections on the basis of obtained risk predictions. Rather than the common and often misguided practice of dichotomizing predictions using a threshold for risk and solely targeting the cases predicted to be at risk, we suggest an approach that not only exploits high-risk predictions by prioritizing the presumed high-risk schools, but also explores the remaining schools irrespective of their predicted risk. In this so-called exploitation–exploration trade-off, exploitation manifests the desire to detect all schools at risk, whereas exploration manifests the demand to maintain and improve the effectiveness of RBI schemes.

On top of this exploitation–exploration trade-off, the suggested approach selects risk predictions for rigorous follow-up verification by weighing a second trade-off: the desired degree of recall and precision. Precision and recall represent conflicting priorities of the inspectorate. Precision manifests the desire to increase the efficiency and proportionality of school visits and cope with limited financial resources by reducing the number of false positives (type I errors). Perfect precision would imply that all inspected schools are indeed at risk. Recall, on the other hand, manifests the political and societal demand to detect each and every school at risk by reducing the number of false negatives (type II errors). Perfect recall would imply that all schools at risk are inspected. In the following, we first discuss various conceptual considerations in the case-based approach to risk predictions and finally discuss the importance of explorations of risk predictions.

Case-Based Predictions

In the current study, we explore *case-based* predictions of risk. A case-based approach has multiple advantages over a rule-based approach. First, the prediction algorithm has a direct relationship with the outcome of interest: the inspectorate's (historical) judgments of schools. Second, it obviates the need to determine indicators and weights manually. Also, it may provide a more unified algorithm that disregards indicators that have no or little predictive value, diminishes the influence of highly correlated indicators, and takes into account complex interactions between various indicators. Of course, a case-based approach has disadvantages as well. For instance, some complex case-based approaches obscure the learned model (i.e., the learned relations between the indicators and the outcome). Such highly uninterpretable models are a *de facto* black box. On the other hand, the algorithms that are used for training such models are fully transparent, and arguably more so than some of the processes underlying rule-based approaches.

A case-based approach to risk prediction warrants two important remarks; one on the causes of risk and one on the definition of risk. First, a case-based approach capitalizes on possibly complex associations between the predictors and the inspectorate's risk judgments. Predictors are frequently interpreted as causal entities, but it cannot be overemphasized that correlation does not imply causation. Second, the developed models capitalize on the inspectorate's risk assessments of the past decade. That is, the models aim to predict the inspectorate's judgments of risk on the basis of past judgments. The prediction algorithm is therefore confined by the accuracy of the inspectorate's judgments, including possible idiosyncrasies such as human biases and varying decision processes. Moreover, it is confined by the historical risk patterns observed in the available data.

A case-based approach alone cannot constitute an entire RBI method, but complements risk assessments that target other aspects of risk. For one, predictions of risk can be embedded in an inspection policy such as described in the Inspection Framework (Dutch Inspectorate of Education, 2019). Moreover, so-called *signals*, such as parents' notifications of encountered problems at a school, can provide timely and possibly time-limited indications of risk that make limited sense in a historical context. Also, other methods of risk assessment exist that can add aspects of risk that are not as easily captured in historical (quantitative) data. For instance, banks are required to undergo (internal) stress tests, to help understand the effects of changes in economic circumstances. Likewise, stress tests for schools could help surface the effects of falling student or teacher numbers.

Finally, and not limited to the case-based approach, there are many dimensions to risk prediction. One may, for instance, distinguish the time of risk, the type of risk, the origin of risk, the chance of risk, and the consequence of risk. Predictions on multiple dimensions provide better clues for the ultimate risk assessment. A case-based approach such as exemplified in this study is primarily concerned with the time and chance of risk, rather than, for instance, the impact of a school at risk (such as if a school is located in a district with few alternative schools). In the studied case-based approach, we aim to predict whether the inspectorate would judge the school to be at risk in the subsequent school year, in order to prioritize inspections. Then, inspections are equipped to determine other dimensions of risk, such as the origins and consequences of risk.

Prediction, Classification, and Uncertainty

The term prediction has thus far been used colloquially, but demands a more thorough definition. The prediction of binary risk, with schools being either at risk or not at risk, is generally termed a classification problem. In this study, we make a distinction between *classification* and *prediction*. In classification, the models' risk scores are transformed into classes (i.e., *at risk* and *not at risk*), typically by means of a simple threshold. Such class assignment is only appropriate when classes can be sufficiently discriminated. Proper classification thus demands strong discrimination performance. Vice versa, when classes cannot be sufficiently discriminated, classification is misguided.

When class discrimination is insufficient and class assignment cannot be justified, prediction is more appropriate. In prediction, class probabilities may provide an estimate of the probability that a school is at risk. Class probabilities are only useful when probabilities are properly calibrated; when estimated class probabilities represent observed class proportions. If this is achieved, class probabilities provide the inspectorate with predictions of the actual chances of particular schools being at risk. Notably, while some models, such as logistic regression, return estimated class probabilities, others provide classification rankings based on risk scores. Nevertheless, several methods exist to calibrate the returned risk scores and obtain class probabilities.

Anticipating model performance results, class discrimination is expected to be poor by definition, or rather, *by lack of a definition*. For one, school quality legislation, the inspectorate's translation of that legislation into policy, and the ultimate implementation of that policy by different inspectors at different offices all carry significant degrees of freedom. Thus, although two schools may be identical, by chance alone the one may be judged to be at risk, whereas the other may not. This notorious challenge in harnessing a strict definition of schools at risk results in an inevitable poor signal-to-noise ratio and significant aleatory uncertainty (i.e., uncertainty due to inherent randomness).

On top of the aleatory uncertainty, there is a significant amount of epistemic uncertainty. For instance, the historical data that are used for training the prediction models are biased toward schools at risk, assuming that the current RBI method outperforms a random selection of schools. Such uncertainty can, in principle, be taken away with improved models, richer data, knowledge about future risk assessments, knowledge about causal risk factors, and so forth. However, in practice, it cannot be taken away easily. Therefore, to mitigate the aleatory and epistemic uncertainties, it is imperative to continuously verify the validity of the models, by exploring schools regardless of their presumed risk. We turn to this issue in the next section.

Exploring Risk Predictions

Ultimately, risk prediction creates an opportunity to prioritize inspections and prevent failure. To this end, we propose a principled procedure. Key in the suggested procedure is the rigorous verification of predictions. On the one hand, it ensures exploitation by prioritizing inspections of presumed high-risk schools: schools are ranked by their predicted risk. This ensures that the expected recall and precision are maximized and the desire to put the limited capacity to good use is respected. On the other hand, it introduces a random component to the inspections. This is fundamental for retaining accurate risk detection, for safeguarding the incentives that result from a nonzero probability of being inspected, and for enabling causal impact evaluations of school inspections. In this section, we discuss the importance of explorations of risk predictions.

First, RBI is susceptible to dramatic failure for a straightforward reason. It is tempting to disregard presumed low-risk schools, and solely target presumed high-risk schools, especially if inspection capacity is limited and proportionality is highly valued. It is similarly tempting to forget that predictions are based on models, and that perfect models only exist in imagination. Whatever RBI method is chosen, some schools predicted to be *at risk* will turn out to be *not at risk* (type I errors), and some schools predicted to be *not at risk* will turn out to be *at risk* (type II errors).

On top of that, prediction performance cannot be easily extrapolated to future observations. First, case-based prediction models are evaluated on the basis of a holdout set that necessarily includes only a portion of the available data. Performance on this set cannot simply be extrapolated to cases outside of it. Even more problematic is the fact that the definition of risk is not only subject to political and societal change, but also the mechanisms causing risk, and its relation with predictor variables, are neither necessarily constant. Indeed, changing inspection policies pose a threat to the long-term validity of prediction models, as metaphorically, the bow of prediction is spanned and targeted at a bull's eye that may have shifted as soon as the shot is fired. Prediction models thus need verification, such that type I errors (false positive predictions) and type II errors (false negative predictions) are not only detected once, but can be continuously detected.

Second, the problem of single-sided verification of predicted risk becomes excessively pronounced when prediction models capitalize on historic inspection judgments, especially if these are obtained through risk-based methods themselves. For one, it is difficult to rule out the existence of misclassifications in historic judgments, which, in turn, damages the accuracy of the prediction model. On top of that, the selection bias introduced by RBI methods (Jacobusse & Veenman, 2016; Sedee, 2019) paves the way for a self-fulfilling prophecy. Thus, explorations of presumed low-risk schools necessarily are a vital effort in any inspection policy that capitalizes on the discussed judgments. Put differently, the aim of visiting presumed high-risk schools is secondary, whereas the development of an accurate risk prediction model is primary; the former fails without the latter.

Third, inspectorate visits may provide an important incentive for schools to prioritize educational quality, similar to the effects of public disclosure of school performance (Canton & Webbink, 2004) and financial incentives (Barber, 2005). The current inspectorate's policy encourages the prioritization of educational quality by setting additional goals (termed quality standards) on top of the strictly legal requirements (Ehren, Leeuw, & Scheerens, 2005). Also, one may argue that the RBI method performs as an incentive, as schools are generally not lining up for inspection visits and schools that are assumed to perform well are less frequently inspected. Having said that, the RBI method also introduces a factor of predictability of the frequency of school inspections that impairs the incentive. Not inspecting presumed low-risk schools removes the incentive, which may have potentially unintentional negative effects. This dynamic can be neutralized by introducing a random component: exploration.

Fourth, it is desired that the inspectorate's presence and policies have a positive effect on school quality. However, for all we know, we can hardly tell. As de Wolf and Janssens (2007) substantiate: "(a) the findings are ambiguous, (b) the research methodology varies substantially and is not always appropriate for testing causal effects and (c) the findings appear to be closely linked to the research methodology used." de Wolf and Janssens rightly argue that experimental research designs are key to inspection policy evaluations. Thankfully, the Dutch four-yearly inspection commitment, in tandem with a random component in deciding on school visits (i.e., exploration), provides the desired conditions. In other words, with a sufficiently large set of schools randomly selected for judgment, an opportunity arises to learn whether the inspectorate's objectives indeed materialize.

In addition to exploration, other specialized methods exist that may target specific elements of the above challenges more efficiently. For instance, in the field of clinical prediction, statistical updating methods (e.g., Su, Jaki, Hickey, Buchan, & Sperrin, 2016) are specifically used to adapt prediction models to changing populations. And, in the field of machine learning, active learning is used to determine which unlabeled cases (schools for which there is no judgment available) must be labeled (i.e., judged) to increase the performance of the prediction model, also in case of class imbalance (e.g., Attenberg & Ertekin, 2013). The exploit and explore procedure suggested in this study is perfectly compatible with these approaches, and additionally addresses incentives and causal impact evaluations resulting from inspection schemes, thus having a more general benefit beyond verifying predictive power.

Table 1. Schools and Judgments across School Years

School Year	Schools	At Risk	Not At Risk	No Judgment
2011–2012	6,603	108	1,092	5,403
2012–2013	6,530	150	1,545	4,835
2013–2014	6,449	110	1,037	5,302
2014–2015	6,337	102	1,203	5,032

Note. For each *School Year* in the training data, the total number of observed *Schools*, the number of *At Risk* schools, the number of *Not At Risk* schools, and the number of schools that received *No Judgment*.

Summary and Overview

Two trade-offs influence the detection of primary schools at risk. Precision and recall balance the desire for efficiency and proportionality with the desire to detect all schools at risk. Case-based prediction models must optimize this trade-off. Then, exploitation and exploration balance the desire to harness the achieved precision–recall balance and the demand to continuously develop and evaluate the prediction models (such that those models remain optimized rather than deteriorate). Ultimately, the balance in exploitation and exploration determines the (long-term) effectiveness of the detection of schools at risk and the extent to which inspection effectiveness can be evaluated.

In the following sections, we evaluate case-based prediction models of Dutch primary schools at risk and suggest an exploit and explore procedure that takes the discussed trade-offs into account. In the Methods section, we describe the data used for training and testing the models. In the Results section, we evaluate both classification and prediction performance of the trained models, and introduce the exploit and explore procedure. We finally discuss exploitation and exploration in educational risk assessment, and the decisions that need to be considered when such a policy is implemented.

Methods

Data

Identifiers. Cases were identified by the school ID and school year. Table 1 gives the number of observed schools per school year and the received judgment (either *at risk*, *not at risk*, or *no judgment*).

Label. The "label" is the variable that the model aims to predict and comprises the inspectorate's historical judgments of Dutch primary schools. The inspectorate's judgments vary from *very poor* and *poor* to *average* and *good*. With the goal of predicting schools at risk in mind, we dichotimized the judgments, with poor and very poor schools labeled *at risk*, and average and good schools labeled *not at risk*. Table 1 gives the frequencies of observed values of the label for the different school years, and the numbers of missing label observations.

Two factors drive (missing) observations. First, the four-yearly inspection commitment drives the majority of observations, yet creates intermediate missing label observations. Second, the inspectorate's RBI method ensures that some schools are inspected more than once every 4 years, yet introduces a strong selection bias mechanism. In the training data, 4, 473 schools were judged once, 336 schools were judged twice, 62 schools were judged thrice, and 4 schools were judged four times.

Table 2. Number of Schools Receiving 1–4 At Risk Judgments in 1–4 School Years

No. of School Years	1	2	3	4
1	0			
2	3	0		
3	9	2	0	
4	275	71	11	1

Note. For each *No. of School Years* with observed *at risk* or *not at risk* judgments in the training data (rows), the number of schools that received an *at risk* judgment once (column *1*), twice (column *2*), thrice (column *3*), or four times (column *4*).

Table 2 provides an intuition for the frequencies of *at risk* judgments per school. It gives the number of schools for which judgments are available across a particular number of years (first column) and the times it was judged *at risk* (subsequent columns). It shows that 275 schools, the majority of schools that were judged *at risk* at least once, were judged every available school year, but only once *at risk*. On the other hand, one school was also judged every available school year, and four times *at risk*. To sum up, across the 6, 729 schools in the data, 372 schools provide 470 judgments of *at risk*.

Finally, the label is characterized by significant class imbalance, as evidenced in Table 1. One can safely assume severe class *prior* imbalance (i.e., base rate imbalance): the vast majority of schools in the Netherlands is fortunately not at risk. Also, the label is characterized by a strong class *sample* imbalance: among the schools that were judged by the inspectorate, the majority is not at risk. Class sample imbalance is arguably less pronounced due to the inspectorate's RBI method, assuming that it selects schools at risk with higher chance than schools not at risk.

Features. The "features" are the descriptive properties that function as the input for the model. The features included data on school-level performance (e.g., proportion of grade retention and grade skipping, average central exam score at the final school year), school-level demographics (e.g., proportion of student absence, changes in student numbers), general school properties (e.g., school denomination, school profile [such as *Montessori*]), general properties of school staff and boards (e.g., average teacher age, average teacher FTE), financial properties of school boards (e.g., financial solvency, housing costs ratio), school area-level demographics (e.g., school area urbanity, proportion of households with children in the school area), and school area-level geographics (e.g., proximity to alternative schools, latitude, longitude). With in total 137 features, ignoring dummy variables for missing observations, the data set is high-dimensional.

Figure 1 summarizes the missing feature observations. First, these missing observations can be expressed in the number of missing feature observations per school. Figure 1(a) provides this information, giving the proportion of missing feature observations for schools across school years. It shows that the vast majority of schools have fewer than 50% missing feature observations. Second, these missing observations can be expressed in the number of missing school observations per feature. Figure 1(b) provides this information, giving the proportion of missing observations across features. It shows that the majority of features have fewer than 25% missing school observations.

In order to take into account the possible effects of the mechanisms that drove missing feature observations, we included dummy variables for features, indicating 1 for a missing value and 0 otherwise, and replaced missing values with the mean (numeric variables) or mode (other variables) of the feature (also see Table 3).

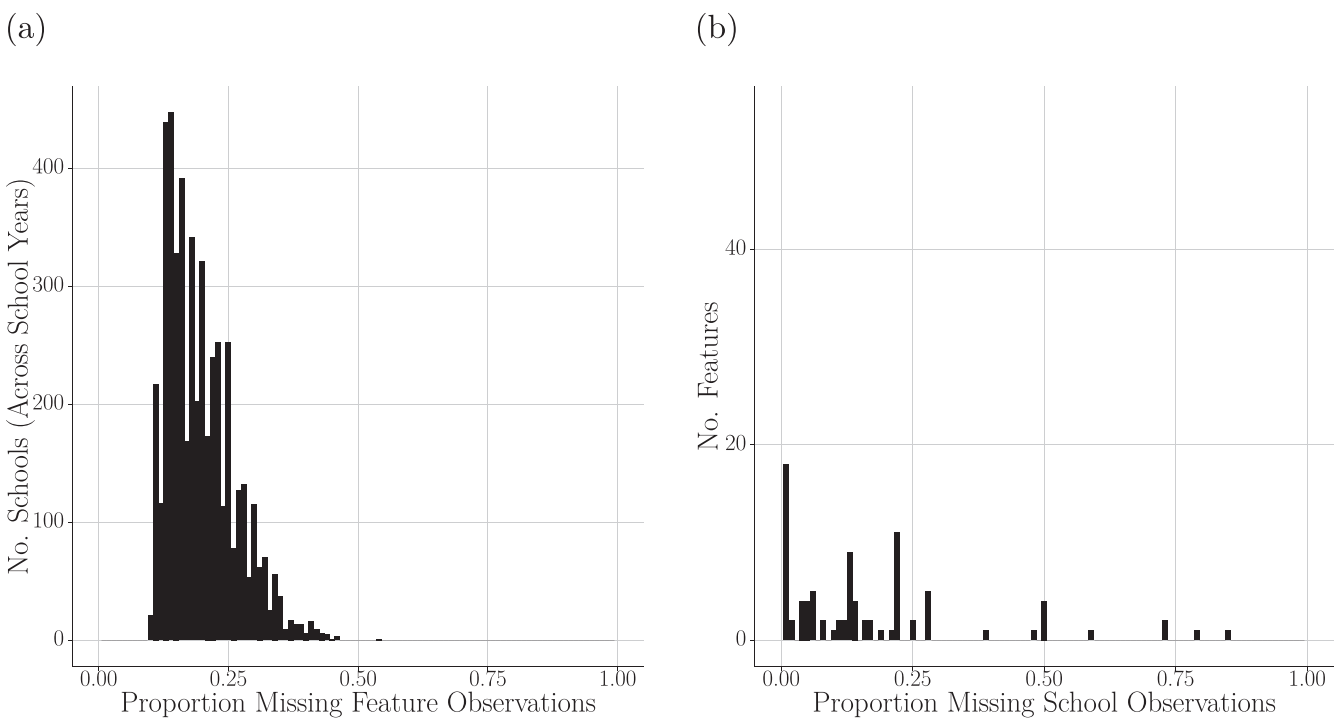

FIGURE 1. Missing feature observations by school (a) and by feature (b) *Note*. (a) The number of schools across school years (*y*-axis) and their proportions of missing feature observations (*x*-axis). Bin width = .01. (b) The number of features (*y*-axis) and their proportions of missing school observations across school years (*x*-axis). Bin width = .01.

Table 3. Feature Engineering

Transformation	Description
Add dummies	Add dummy variables for features with missing values.
Impute missings	Impute missing values. Numeric features were imputed with their mean value, whereas other features were imputed with their mode value. By using mean imputations the variance of the targeted features is decreased proportional to their degree of missingness, which reduces their possible importance in the model. Moreover, the imputed means are captured by the explicit missing dummy variables in the previous step.
Remove cases	Remove cases with a missing label value.
Remove features	Remove constant features and identical features.

Note. Transformations were performed in the given order.

Preprocessing. The data were preprocessed in three broad steps. First, we split the data into a training and test set. The cases for school years 2011–2012 to 2014–2015 were added to the training set and for school year 2015–2016 to the test set. The features of the cases at year t matched the labels of those cases at $t + 1$, such that the models were trained to predict the following rather than current year. We assumed that the associations between features at t and labels at $t + 1$ did not depend on the value of t in the training and test set. Second, for every school, only the first school year with an observed label was selected for the training regime. While the four-yearly evaluation cycle provides independent observations, the RBI method creates large dependencies between multiple-label observations for a single school within such a cycle (for instance, schools previously assessed to be at risk are visited the subsequent year). Finally, we performed various feature transformations and selections, described in Table 3.

Access and eligibility. The data come primarily from the same sources that provide data for the inspectorate's regular process of risk assessment, and partly from open sources, for instance, provided by the Dutch Ministry of Education, Culture and Science.[1] Access to the data was granted and facilitated by the inspectorate. The data were aggregated at the level of the schools.

Models

Training and validation. Four methods were evaluated using the R software (v4.0.2) and *caret* package (v6.0-68): logistic regression (`lr`; using the *stats* package), logistic regression via penalized maximum likelihood (`plr`; using the *glmnet* package, v4.0-2), support vector machine with linear kernel (`svm`; using the *e1071* package, v1.7-3), and stochastic gradient boosting (`gbm`; using the *gbm* package, v2.1.8). Parameters were tuned using caret's default grid search (with a grid size of three times the number of parameters). Resampling was done using 10-fold cross-validation, repeated five times. Error terms were averaged across folds and repetitions. Models were assessed on the basis of the obtained area under the precision–recall–gain curve (*PRG AUC*; Flach & Kull, 2015), and the model with the largest AUC was selected.

It deserves note that often, the receiver operating characteristic (*ROC*) curve is used to assess and select models. The ROC curve evaluates the trade-off between the obtained false positive rate ($\frac{FP}{FP+TN}$) and true positive rate ($\frac{TP}{TP+FN}$). The area under the ROC curve (*ROC AUC*) is then used to summarize this trade-off in a single number. However, with high class imbalance (i.e., positives $\ll$ negatives), the ROC curve is more informative of the identification of the majority class (schools *not at risk*) than of the identification of the minority class (schools *at risk*). Although Juba and Le (2019) show that larger data sets are necessary and sufficient for dealing with class imbalance, the number of Dutch primary schools is effectively fixed and there is limited to zero capacity for additional inspections.

When negatives prevail but true negatives are of limited interest, the precision–recall (*PR*) curve provides a more informative measure of performance than does the ROC curve (e.g., Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015). The precision–recall curve evaluates the trade-off between the obtained positive predictive value ($\frac{TP}{TP+FP}$) and true positive rate ($\frac{TP}{TP+FN}$). It is superior to other measures that are used in case of class imbalance (such as concentrated ROC curves and cost curves; Saito & Rehmsmeier, 2015) and the optimization of the ROC AUC does not necessarily improve the PR AUC (Davis & Goadrich, 2006). The PRG AUC that we used in this study is effectively obtained by plotting the PR curve in a different coordinate system, which solves some inherent issues of the PR AUC and is argued to result in better model selection (Flach & Kull, 2015).

Evaluation. Model performance was evaluated on the basis of different criteria. First, we evaluated *discrimination performance*, which is key to proper classification. For this we used rank-based criteria for data with class imbalance: precision–recall–gain curves (quantified by their AUC values) and precision and recall at k. Second, we evaluated *calibration performance*, which is key for proper prediction. For this, we used calibration-based criteria: reliability curves (quantified by two common skill scores: Brier scores and logarithmic loss). The criteria are all threshold independent, provide an opportunity to prioritize different aspects of model confusion (i.e., precision and recall), and evaluate different goals (i.e., classification and prediction). The criteria are summarized in Table 4.

Results

In this section, we discuss both model performance and model utilization. First, we evaluate discrimination performance and calibration performance using the criteria described in the Methods section. Second, we propose a principled exploit and explore procedure that allows one to prioritize inspections on the basis of predictions. This procedure

[1] https://duo.nl/open_onderwijsdata/

Table 4. Model Performance Measures for Discrimination and Calibration

Criterion	Description
Precision–recall–gain curve	The precision–recall–gain (PRG) curve evaluates the trade-off between the obtained positive predictive value ($\frac{TP}{TP+FP}$) and true positive rate ($\frac{TP}{TP+FN}$), where *T* and *F* denote *true/false* and *P* and *N* denote *positive/negative*. The baseline in a precision–recall analysis is the always-positive classifier, where recall = 1 and precision = π, such that only precision and recall values in the interval $[\pi, 1]$ need to be considered (Flach & Kull, 2015). In a precision–recall–gain analysis, these values are rescaled to the interval [0, 1], using a harmonic scale.
Precision–recall–gain AUC	The *area under the precision–recall–gain curve* (PRG AUC) summarizes the precision–recall trade-off in a single number. The higher the better, generally speaking. The PRG AUC is threshold independent.
Precision @ k	The precision at *k* evaluates the precision of the model at the *k* schools with the highest predicted risk. The limited resources of the inspectorate force the inspectorate to limit the amount of visited schools to some *k*. The precision at *k* is threshold independent.
Recall @ k	The recall at *k* evaluates the recall of the model at the *k* schools with the highest predicted risk. The limited resources of the inspectorate force the inspectorate to limit the amount of visited schools to some *k*. The recall at *k* is threshold independent.
Reliability curve	Reliability curves compare estimated class probabilities and observed class proportions. Probability calibration was performed using Platt scaling and isotonic regression. Calibration performance is summarized with skill scores.
Skill scores	Skill scores quantify calibration performance. The *Brier score* gives the mean squared difference between estimated class probabilities and the actual classes, whereas the *logarithmic loss* takes the natural logarithm rather than the square of the difference. The lower the better, generally speaking. The log loss provides a stronger penalization for large deviations than the Brier score.

exploits the desired level of precision and recall, yet maintains a similarly desired level of chance in picking schools for further risk assessment.

Model Performance

Discrimination performance. Discrimination performance signifies the extent to which a model can be used for classification. In case of a well discriminating model, the model consistently assigns higher risk scores to schools at risk than to schools not at risk. Similarly, in case of a poorly discriminating model, there is no such consistent difference. For the purpose of evaluating discrimination performance, only the predicted risk rankings are of interest, and we thus evaluated the models' default class probability estimates (without additional calibration) for cases in the test set.

Risk score distributions. To begin with, Figure 2 shows the distributions of risk scores for schools at risk and schools not at risk. Clearly, the achieved class discrimination suffers from a significant overlap between risk scores for both classes. As we alluded to in the introduction, perfect discrimination is hard to achieve even in simple problems, let alone in classifying schools at risk.

Precision–recall–gain curves. Precision–recall–gain (PRG) curves help capture the trade-off in precision and recall. The curves, and their associated areas under the curve (*PRG AUC*), provide global measures of classification performance. Figure 3 shows the PRG curves for all models. Table 4 provides an intuition for the precision and recall *gain* scales. Also, while in an ROC curve, the antidiagonal provides the baseline, in a PRG curve, the main diagonal provides the baseline. Table 5 summarizes the PRG AUC values for those curves. Although the AUC values obscure model performance idiosyncrasies by aggregating the curves into a single value, they do allow for a quick global comparison. Other than the ROC AUC, the PRG AUC does not allow for an intuitive understanding, but the same rule of thumb applies: a larger AUC generally indicates better model performance.

Table 5. Area Under the Curve (AUC) Values for the Precision–Recall–Gain (PRG) Curves

Model	PRG AUC
gbm	.710
lr	.731
plr	.812
svm	.757

Precision and recall at k. Although the risk score distributions and PRG curves give an important intuition of the overall classification performance, the inspectorate's responsibilities demand more specific performance measures. Precision and recall at *k*, as shown in Figure 4, capture both key aspects independently. Precision at *k* provides the number of correct predictions at the top *k* predicted schools divided by the total number of predicted schools at *k*. The top panel of Figure 4 shows the precision at *k* for the test set. When the top *k* schools are visited by the inspectorate, it provides the proportion of schools at risk in that sample of visited schools. A higher precision at *k* signifies a better proportionality. In the ideal scenario, precision remains 1 with every increase in *k*, until recall is 1.

Then, recall at *k* provides the number of correct predictions at the top *k* predicted schools divided by the total number of schools at risk. The bottom panel of Figure 4 shows the recall at *k* for the test set. It provides the number of schools that must be visited in order to detect a certain proportion of schools at risk. A higher recall at *k* signifies a better coverage of schools at risk. In the ideal scenario, recall increases with every increase in *k*, until recall is 1.

Taking these discrimination performance results together, classification by means of class assignment defeats the purpose of risk detection and is ill-advised. The reason is

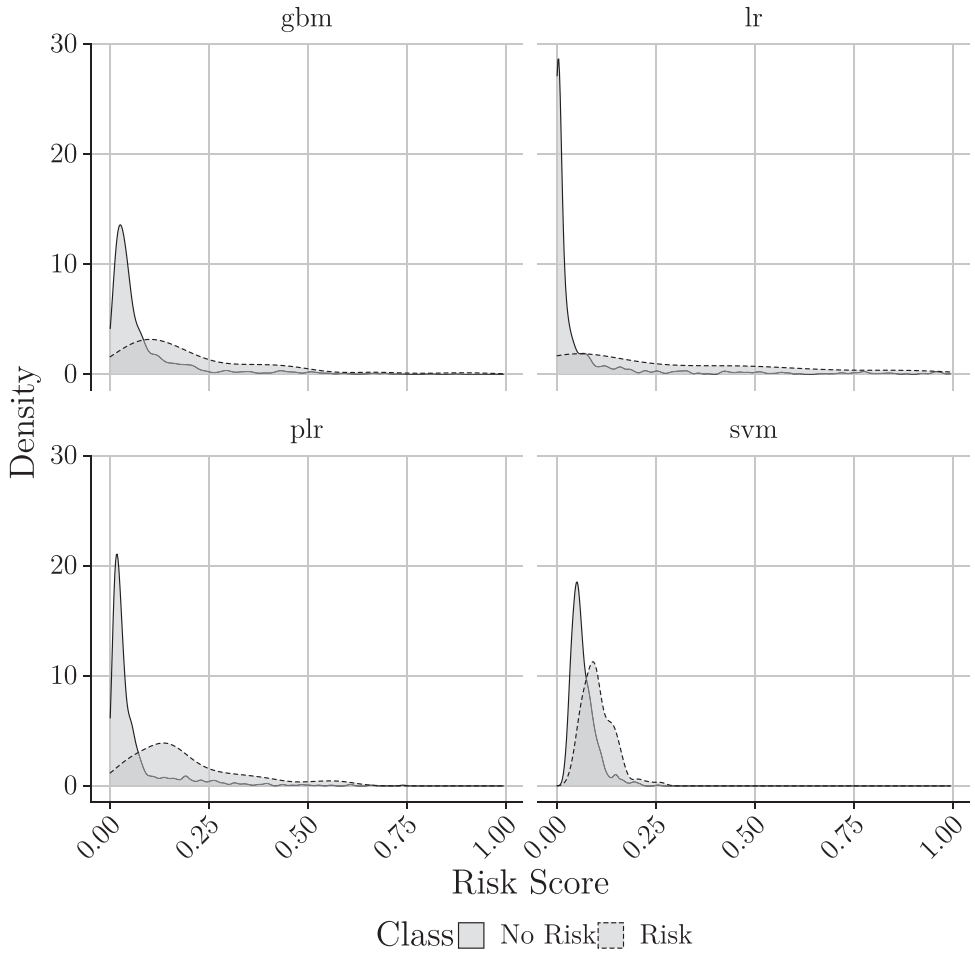

FIGURE 2. Density curves of the risk scores for the *at risk* and *not at risk* class. *Note*. The distributions reflect the achieved class discrimination for the models on the test set. Bin width = .01.

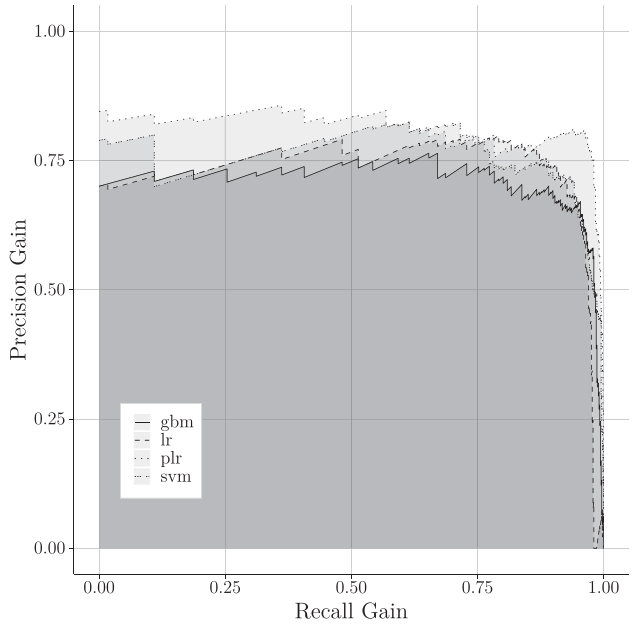

FIGURE 3. Precision–recall–gain curves. *Note*. A larger area under the curve signifies a better fit of the model with respect to its precision and recall.

easily observed. Class assignment makes sense if classes can be clearly discriminated and risk score rankings or class probabilities have little added value. However, the achieved class discrimination of the trained models is poor. We believe that the aleatory and epistemic uncertainty discussed in the introduction, strengthened by the currently achieved discrimination performance, renders classification inappropriate.

Calibration performance. Calibration performance signifies the extent to which a model can be used for prediction. In case of a poorly calibrated model, the estimated class probabilities and observed class proportions are *not* in agreement. Similarly, in case of a well-calibrated model, there *is* such agreement. Conveniently, a well-calibrated model can thus be said to predict the actual class, as it provides that, for instance, not visiting schools with a risk estimate of .1 would result in an approximate 10% error rate. As discussed in the introduction, proper prediction demands strong calibration performance. Since many predictive models return risk scores rather than class probabilities, we performed two calibration methods that aim to scale the obtained risk scores to actual probability estimates.

Following the model training phase, we calibrated the risk scores by applying either Platt scaling or isotonic regression to the model predictions for the entire training set, and used the obtained model to predict the classes of the test set. With Platt scaling, a logistic regression is performed that is robust

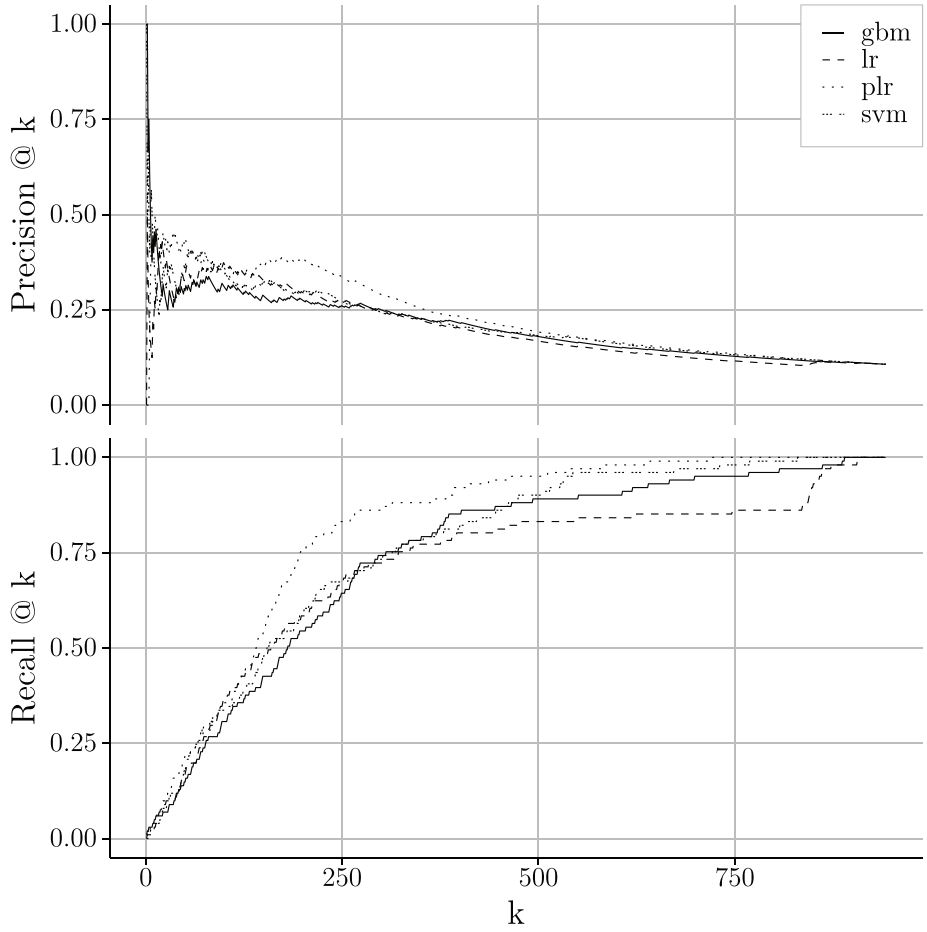

FIGURE 4. Precision and recall at the top *k* predicted schools at risk. *Note*. Evaluated on the test set. Precision at *k* is the number of correct predictions at the top *k* predicted schools divided by the total number of predicted schools at *k*. Recall at *k* is the number of correct predictions at the top *k* predicted schools divided by the total number of schools at risk.

to sigmoidal distortions of the risk scores. Likewise, isotonic regression is robust to any monotonic distortions of the risk scores. Unfortunately, also in this prediction rather than classification context, class imbalance is suspected to introduce bias toward the majority class, which is neither mitigated by the PRG AUC nor these calibrations (Wallace & Dahabreh, 2012). In the following, we evaluate the calibration performance of the obtained probability estimates.

Reliability curves. To begin with, Figure 5 shows the reliability curves for each calibration method, by comparing the binned probability estimates to the observed proportions of schools at risk in those bins. Close inspection of the reliability curves reveals an underestimation of risk in the low-risk bins and an overestimation of risk in the high-risk bins, for both Platt scaling and isotonic regression. The first (low-risk) bin is extremely well calibrated, whereas the high-risk bins are poorly calibrated. This is explained by the severe class imbalance (low-risk predictions are much more frequently observed than high-risk predictions) and illustrates the challenge to correctly predict schools at risk with high certainty (i.e., large probability).

Skill scores. Calibration performance (sometimes termed prediction skill) can be summarized with skill scores. Table 6 provides the Brier score and logarithmic loss for each type of calibration. The Brier score and log loss quantify the difference between the probability estimates and actual classes, with the log loss providing a stronger penalization for larger deviations. Smaller losses signify better calibration.

Table 6. Brier Scores and Logarithmic Loss for the Calibrated Models, Using Platt Scaling and Isotonic Regression

Model	Calibration	Brier Score	Log Loss
gbm	isotonic	.089	.317
gbm	none	.090	.303
gbm	platt	.100	.355
lr	isotonic	.101	.946
lr	none	.099	.521
lr	platt	.110	.420
plr	isotonic	.085	.309
plr	none	.083	.268
plr	platt	.094	.324
svm	isotonic	.100	.682
svm	none	.091	.312
svm	platt	.102	.373

Note. Where calibration = none, the package's default class probabilities (i.e., risk scores) were used.

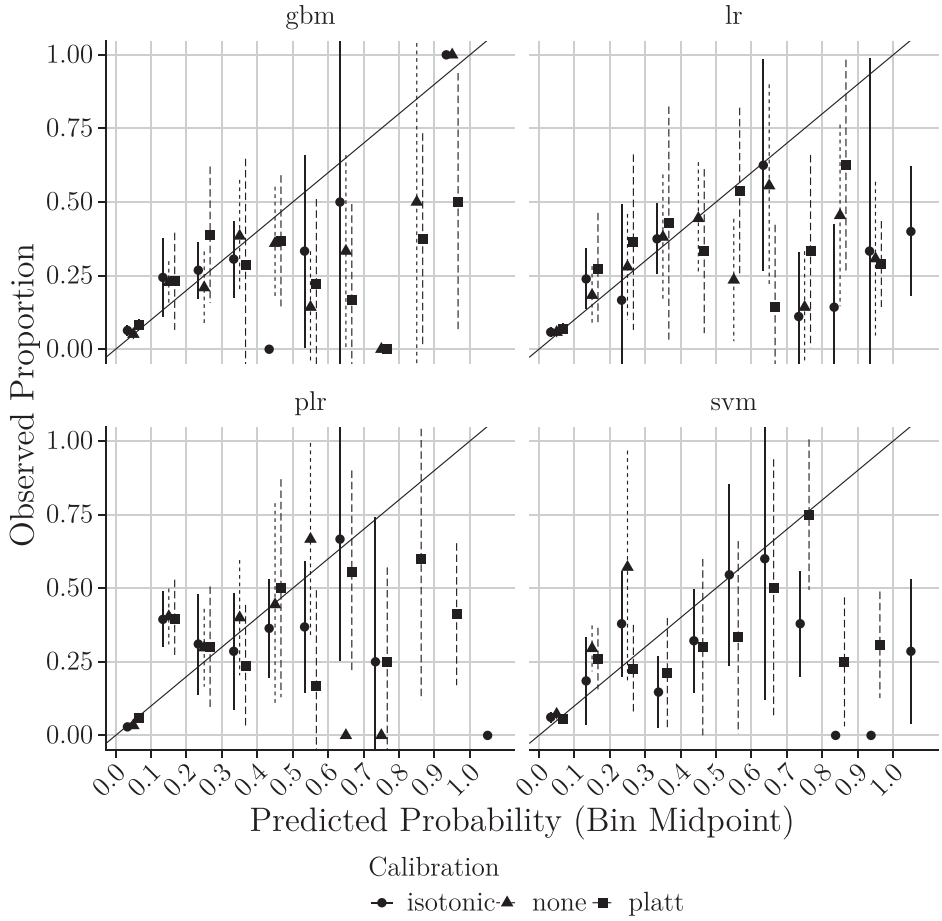

FIGURE 5. Reliability curves with 95% confidence intervals. *Note*. Platt scaling and isotonic regression are compared. Obtained probabilities are binned with bin width = .1. The diagonal reflects perfect calibration; the confidence intervals provide an intuition for the extent to which the point estimates may or may not deviate from the diagonal. Where calibration = none, the package's default class probabilities were used.

Similar to the previously discussed PRG AUC values, the skill scores obscure possibly important model performance idiosyncrasies. That is, proper calibration may be more important for some parts of the curve than for others. For instance, overestimation in low-risk bins might be less of an issue if there is no capacity to inspect schools with low-risk predictions. Here, we do not evaluate such possible idiosyncrasies, as these are policy-specific.

Summarizing, accurate calibration may significantly aid prediction by directly indicating the expected hit and error rates when a follow-up risk assessment needs to be decided upon. However, taking these calibration performance results together, the currently achieved calibration performance renders strong prediction unfeasible. With neither definite classification nor accurate prediction, the risk scores nevertheless do provide predictive power. In the next section, we therefore discuss some fundamental considerations in model utilization.

Model Utilization

In the above, model performance is evaluated by the ability to discriminate schools at risk from schools not at risk, measured against the achieved precision and recall, and the ability to calibrate risk scores such that reliable probability estimates are obtained. Although these measures provide relevant details on model performance, they provide little guidance for using the risk scores to plan inspections. Necessarily, schools with a high presumed risk must be examined, a process called exploitation. On top of that, and irrespective of the scenario (classification or prediction), some form of exploration is fundamental to tracking the predictive power of the model, especially in the long run. Here, exploration is viewed as the examination of schools regardless of the presumed risk. Ultimately, an exploit–explore framework must be employed that both exploits probable risks and explores unforeseen ones. Therefore, we further develop the exploit and explore procedure that is suggested in the introduction. For the expository rather than evaluative purpose of this section, we consider a single model: the `plr` model. We consider its predictions for the training data, as the training data contain the vast majority of schools in the Netherlands, whereas the test data contain a limited subset. Moreover, for the expository purpose, we assume that these data reflect a cross section of the schools for a particular year.

In Figure 6, we suggest a principled approach. This approach does justice to the significant amount of prediction uncertainty discussed in the introduction and evidenced in this section, and provides important guidance when deciding upon an exploit–explore policy. It shows for different amounts

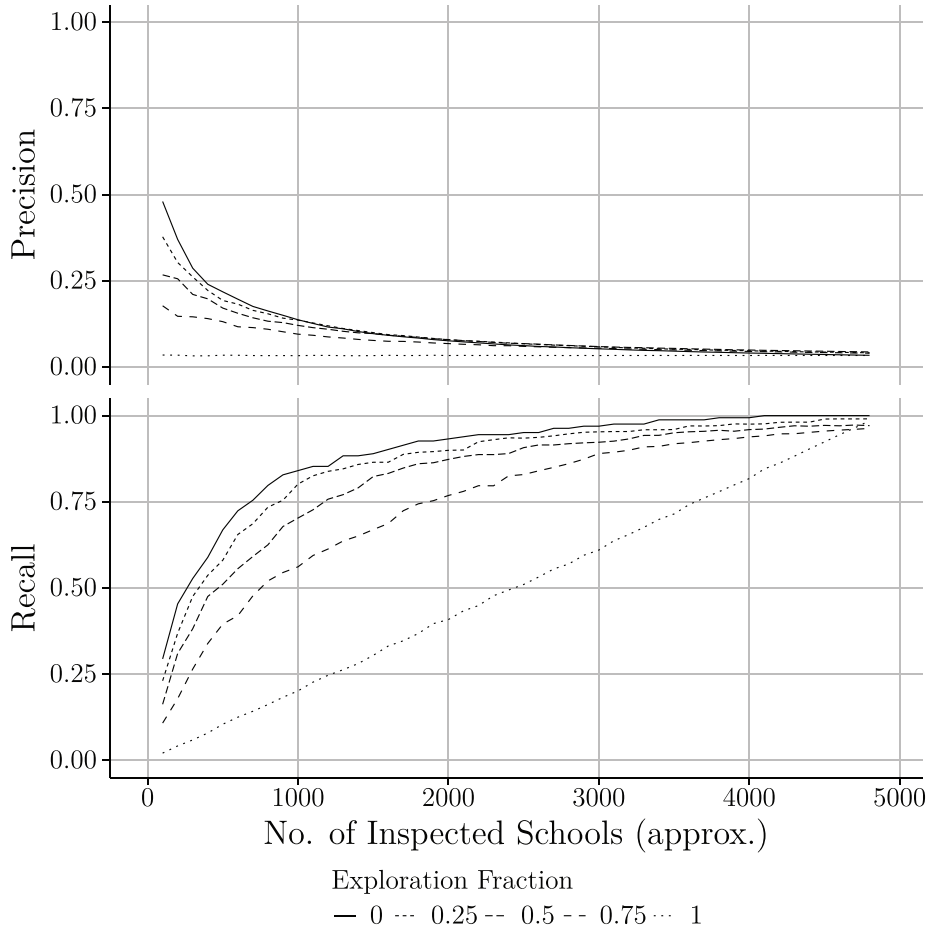

FIGURE 6. Precision and recall with exploration. *Note*. Precision (top panel, *y*-axis) and recall (bottom panel, *y*-axis) for different numbers of inspected schools (*x*-axis), given all (unique) schools in the training data. Predictions from the `plr` model were used. Schools are ordered by presumed risk. Line types distinguish the proportions of schools (of the number of inspected schools on the *x*-axis) that were randomly sampled from all schools in the training data. Each data point represents the average of 100 samples. An exploration fraction of 0 implies a fully exploitative strategy such that in that case, the *y*-axes reflect the precision (respectively, recall) at *k*. Exploration was performed on all predictions, such that schools may end up in both the exploited and explored set of schools (hence the approximation of the number of inspected schools).

of exploration the obtained precision and recall when a particular number of schools is inspected. For instance, with an exploration fraction of .25 and 1,000 inspected schools, the 750 schools with the highest presumed risk end up in the exploitation set and 250 randomly selected schools end up in the exploration set. As such, some schools may end up in both sets. The figure gives the expected precision and recall when the union of both sets is inspected. As can be seen in the figure, when 1,000 schools are inspected, recall may vary from .2 (with an exploration fraction of 1; all inspected schools are randomly chosen) to .85 (with an exploration fraction of 0; only the 1,000 schools with the highest presumed risk are inspected).

Figure 6 captures the key trade-offs in risk prediction discussed in the introduction. First, it demonstrates the trade-off in precision and recall; balancing the desire to spend the limited capacity and resources well (precision) and detecting all schools at risk (recall). Second, it demonstrates the trade-off in exploitation and exploration, balancing an optimal precision–recall trade-off (exploitation) and the imperative virtues of exploration. Finally, it demonstrates how these trade-offs are impacted by the number of inspected schools.

A number of considerations aid the formulation of a specific exploit and explore policy. The three factors that will shape most policies are the number of inspected schools, the desired recall, and the desired exploration fraction. The first depends on the available capacity for inspections, the second on the political or societal tolerance for schools at risk to go undetected, the third on the desire to maintain or improve the predictive validity of the prediction model, to provide an incentive to all schools, and to evaluate the causal effect of inspections. As shown in Figure 6, the consequence of a limited number of school inspections is a possibly unacceptable recall. Vice versa, the consequence of a low tolerance for schools at risk to go undetected is the necessity to guarantee sufficient inspection capacity. Exploration is key to maintain and improve the predictive validity of the models in the long term, and as such to increase recall and decrease the required number of inspections, but has the consequence of decreasing direct performance on both aspects.

In Figure 6, the slope of a line directly translates to the practical value of changes in either recall/precision or the number of inspected schools. Similarly, the proximity of different lines translates to the practical value of changes in

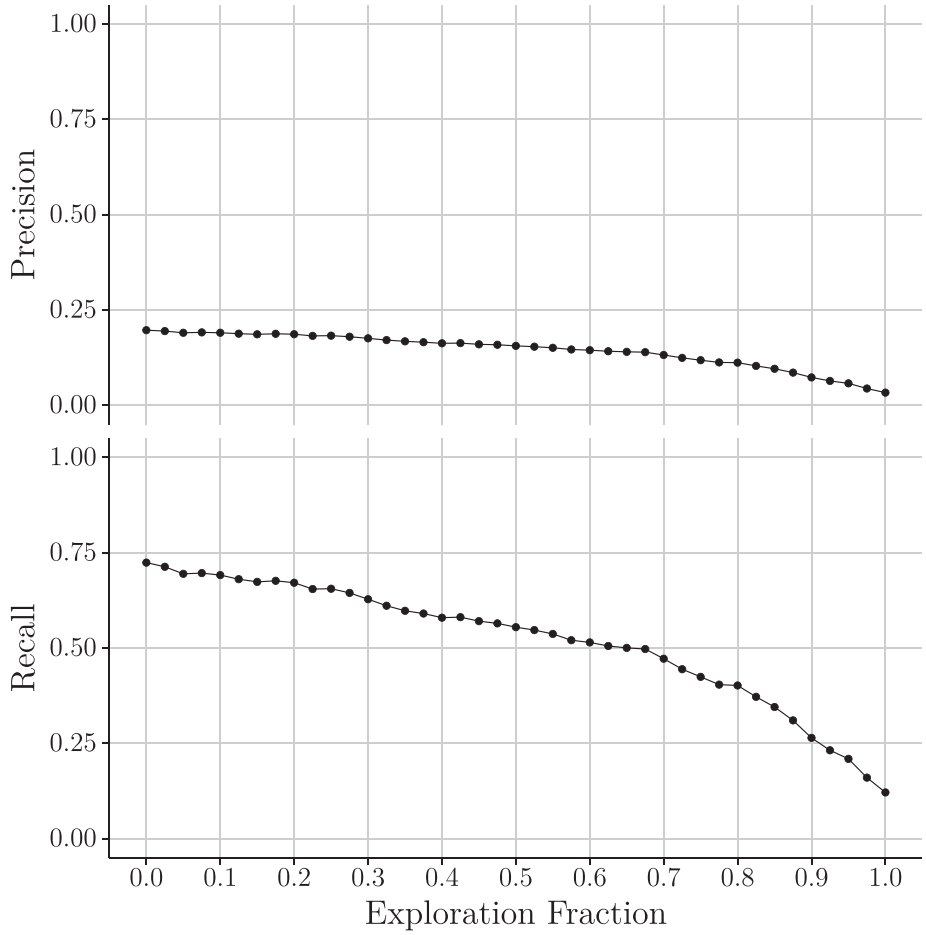

FIGURE 7. Precision and recall with exploration (600 inspected schools). *Note*. Precision (top panel, *y*-axis) and recall (bottom panel, *y*-axis) for different proportions of exploration (*x*-axis), given all (unique) schools in the training data. Predictions from the `plr` model were used. The number of inspected schools is fixed at approximately 600. Exploration was performed on all predictions, such that schools may end up in both the exploited and explored set of schools (hence the approximation of the number of inspected schools).

the exploration fraction. For instance, the figure exposes the seemingly limited effect of an increase of the exploration fraction from 0 to .25 on the obtained precision and recall. The implication of this increase is, however, significant; it marks the difference between all the disadvantages of a purely exploitative strategy and all the advantages of exploration. Figure 7 highlights the effect for a fixed number of 600 inspected schools and a larger variety of exploration fractions. Indeed, both precision and recall seem to suffer relatively little from increases in exploration, although the decline becomes more pronounced when the exploration fraction exceeds .7. However, the optimal balance in the various trade-offs is not a matter of statistics; it is a matter of policy. We describe various such policies in the Discussion section.

Crucially, to evaluate the causal impact of inspections, one may desire to compare schools across the full range of presumed risks. Figure 8 shows the impact of not inspecting schools across the full range, thus also among the schools with high presumed risk. The figure is similar to Figure 6, but for each school randomly designated to be visited (the exploration set), another school is designated not to be visited (the control set). Schools in the control set are removed from the schools in the exploitation set. The experimentation fraction denotes the fraction of schools that is randomly visited for the purpose of causal impact evaluation, and for which a similar number of randomly selected schools is not visited. It is expected that the impact of experiments on the expected recall is larger for larger experimentation fractions. Moreover, as experiments directly influence the exploitation set, recall can even be seen to decline with large numbers of inspected schools and large experimentation fractions. Fortunately, with larger numbers of inspected schools, smaller fractions of schools are required for proper causal impact evaluations.

Discussion

To guarantee a desirable level of educational quality, early risk detection is key. In this study, we explored case-based predictions of primary schools at risk. We evaluated both classification and prediction performance of the trained models, and suggested a principled and actionable exploit and explore procedure for prioritizing inspections. The exploit and explore procedure targets various challenges substantiated in the introduction. First, it allows one to systematically deal with the significant prediction uncertainty encountered in this very problem, and must ultimately help limit the number of schools at risk that go undetected. Performance results show that schools at risk appear across the complete rank ordering of most trained models, signifying that schools at risk

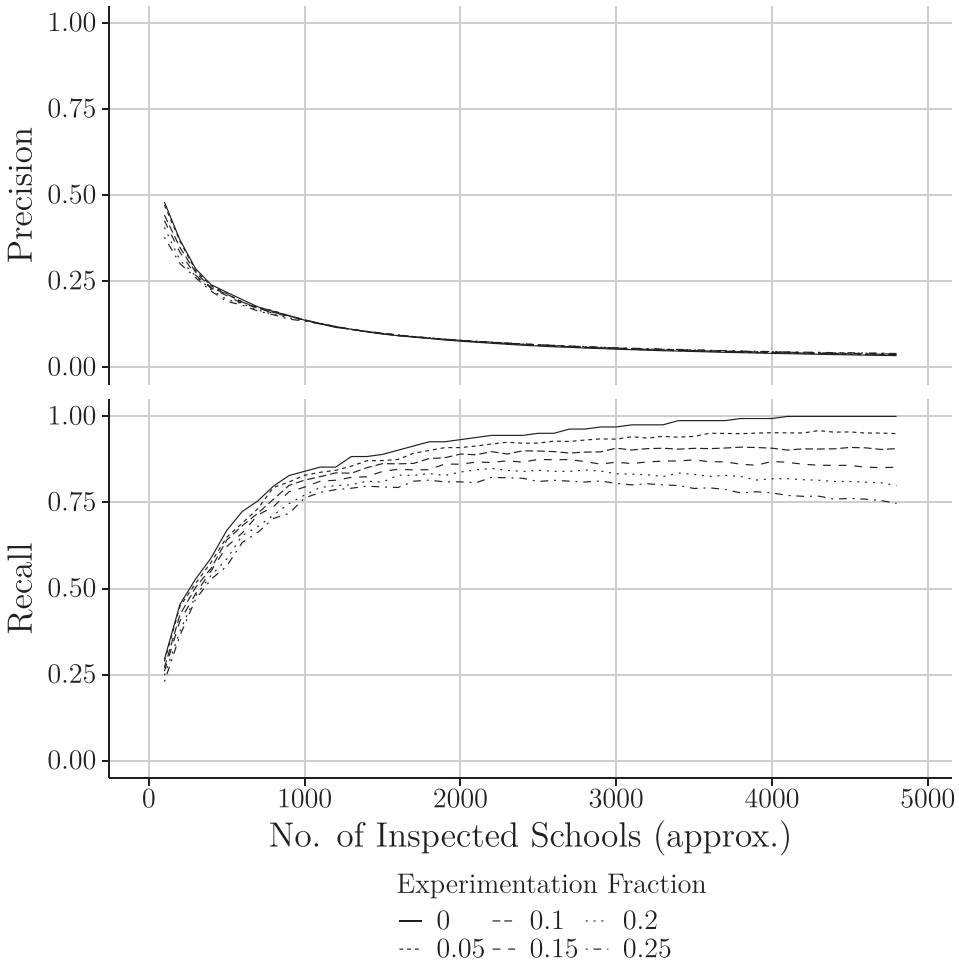

FIGURE 8. Precision and recall with experimentation. *Note*. Precision (top panel, *y*-axis) and recall (bottom panel, *y*-axis) for different numbers of inspected schools (*x*-axis), given all (unique) schools in the training data. Predictions from the `plr` model were used. Schools are ordered by presumed risk. Line types distinguish the proportions of schools (of the number of inspected schools on the *x*-axis) that were randomly sampled from all schools in the training data. Each data point represents the average of 100 samples. The experimentation fraction denotes the fraction of schools that is randomly visited for the purpose of causal impact evaluation, and for which a similar number of randomly selected schools are not visited. The experimentation set was randomly drawn from all predictions, such that schools may end up in both the exploited and experimental set of schools (hence the approximation of the number of inspected schools).

and schools not at risk are difficult to discriminate. On top of that, in such a situation, it is a matter of fairness to not solely exploit the model and only act upon presumed risks. Second, it provides a structural way of evaluating the effectiveness of school inspections, and the activities that schools implement conditionally on inspection to improve the quality of their education. Finally, having a random component in acting upon presumed risks may provide an additional accountability incentive for schools to prioritize educational quality.

In the following, we first discuss three principles that guided the development of the exploit and explore procedure (in addition to the exploration principle discussed in the introduction). We then discuss how these principles combine into different exploit and explore policies that help target the various challenges of risk detection. But before discussing exploit and explore policies in detail, it is useful to understand that such policies are beneficial beyond the application in this article. Risk detection is a universal problem, and the characteristics discussed in the introduction apply to many situations. First, failures (or other consequences) are typically rare and their origins complex. Second, resources and capacity to exhaustively determine all risks are often insufficient, and efficiency and proportionality are desired. When these conditions hold, exploit and explore policies help to prioritize investigations of risk and benefit from the previously discussed advantages of exploration. It follows that fields that may benefit from exploit and explore procedures span from health care to finance, and that the institutions that may benefit are not limited to inspectorates. More concretely, examples include the detection of risks of minimally invasive surgery (Stassen, Bemelman, & Meijerink, 2009), the detection of students at risk of adverse academic outcomes (Lakkaraju et al., 2015), or the detection of psychosocial risks at work (Weissbrodt, Arial, Graf, Iff, & Giauque, 2018).

Guiding Principles

To begin with, we evaluated the prediction models independent of a risk threshold. Indeed, the poor discrimination performance of the trained models provides one evident rationale for this choice. Nevertheless, the inspectorate's RBI method demands that only a selection of schools is inspected, and a policy prioritizing inspections must thus be established. The achieved calibration performance may provide

some guidance, as class probabilities can be used to select different proportions of schools with different risk estimates, but only when estimated class probabilities sufficiently match observed class proportions. In any case, a principled procedure must be used for school selection. The suggested exploit and explore framework can be used independent of risk thresholds or risk probabilities.

Second, we take into account the trade-off in recall and precision. As discussed previously, the precision–recall trade-off not only benefits model optimization for imbalanced problems (in comparison with the sensitivity–specificity trade-off), but precision and recall also represent two key conflicting priorities of the inspectorate. Typically, inspectorates must weigh their desire to spent their limited resources well (and obtain high precision), and their desire to prevent the misidentification of actual schools at risk (and obtain high recall). The suggested exploit and explore framework allows one to understand the expected impact of different policy decisions on the obtained precision and recall.

Finally, the exploit and explore framework must be actionable. That is, it must aid an inspectorate to act upon predictions in a principled manner. The Dutch inspectorate's current RBI method can be easily adapted to the suggested exploit and explore framework. As described in the introduction, risk predictions are currently used to prioritize desk research. Exploit and explore policies can augment such prioritization, such that a proportion of low risk predictions is acted upon, or a proportion of high-risk predictions is not acted upon. Naturally, to fully benefit from such a policy, this selection must contain a random component, and presumed risks must be thoroughly verified.

Exploit and Explore Policies

The discussed principles combine into an exploit and explore framework, allowing for different policies that balance the trade-offs in risk detection. However, before developing a policy, it is crucial to understand the interdependencies between the desired recall or precision, the degree of exploitation or exploration, and the number of predictions that are acted upon. Given that the prediction model outperforms a random selection of schools, exploitation (i.e., acting upon high-risk predictions) benefits both *immediate* precision and *immediate* recall, whereas exploration harms *immediate* precision and *immediate* recall. Here, *immediate* is emphasized, as in the long run, the prediction model may derail when a proper exploration policy is not in place. Indeed, evolving definitions of risk, evolving policies for determining risk, and evolving mechanisms underlying risk pose a threat to the prediction model. Second, precision benefits from acting upon a limited number of presumed high-risk schools, whereas recall benefits from acting upon as many schools as possible. It follows that while exploration harms immediate recall, this can be remediated by acting upon a larger number of predictions. For precision, it largely depends on the degree of exploitation and exploration.

The exploit and explore procedure provides two natural points of departure. First, the number of schools that can be acted upon can be a guideline, as inspection capacity does not allow all schools to be inspected on a yearly basis. In this case, the desired degree of exploitation and exploration ultimately determines the expected precision and recall. Second, the desired recall can be guiding, as the inspectorate is accountable for detecting schools at risk. In this case, the desired degree of exploitation and exploration determines the number of predictions that need to be acted upon. These points of departure are at odds, as with limited capacity, full recall cannot be guaranteed. Finally, the desired precision and desired degree of exploitation and exploration make limited sense as a point of departure, though both fulfill key roles that have already been discussed in length.

Regardless of one's point of departure, an exploit and explore policy requires one to set a degree of exploitation and exploration. For the sake of model evaluation and development, a fully exploratory strategy can be said to have the largest entropy and is suspected to provide the most (surprising) information. However, a fully exploratory strategy defeats the purpose of prediction. Therefore, a trade-off between exploitation and exploration must be set. One policy would be to exploit the top predictions of risk, and explore a number of remaining schools. Necessarily, exploration is random, although one may, for instance, create different bins for presumed risk severity, and alter the amount of exploration per bin. If calibration performance is high, probability estimates can be used to guide the exploit and explore policy. Regardless of the chosen policy, Sedee's (Sedee, 2019) tentative findings signal that there are optimal degrees of exploration in the exploit–explore trade-off.

For the sake of inspection evaluation, a different strategy must be used. Several challenges render inspection evaluation a particularly nontrivial objective. For one, proper evaluation demands experimental comparisons, and thus requires some schools to be acted upon and others not to be acted upon, independent of presumed risk. Thus, if the top predictions of risk are exploited, the inspection effectiveness with regard to those predicted schools cannot be compared to a control (i.e., similar schools that are not exploited). Indeed, even the verification of presumed risk by planning an inspection without taking further measures if a school is indeed at risk can be viewed as an intervention and potentially hampers the validity of the evaluation. Especially when RBI methods without exploration are used, school visits can signal expected risk and function similar to placebo treatments in clinical trials. Consequently, exploration has the additional benefit that this signal is removed, as each and every school may be visited, regardless of presumed risk. However, not acting upon presumed risks might be a step that the inspectorate is not willing to take. This can be partially resolved by resorting to a treatment as usual control and evaluating different types of measures. de Wolf and Janssens (2007) provide some suggestions, such as randomly varying the types or numbers of inspections.

Ultimately, an exploit and explore policy must help to improve prediction performance. If and when proper discrimination performance is achieved, the desired policy for the exploit and explore procedure may shift accordingly. For instance, one may choose to accept that full recall is difficult to achieve and focus on precision, by lowering the degree of exploration for cases with lower associated risk. On the other hand, if full recall is pursued, one may increase the degree of exploration for those cases. However, this comes at the cost of either the number of predictions that is acted upon, or the number of presumed high-risk schools that are judged to be at risk, without proper verification. Naturally, one may also alter the degree of verification depending on the presumed risk.

Conclusions

The ship of dichotomizing predictions of risk using thresholds, and only acting upon presumed risks, has sailed. The exploit and explore procedure introduced in this study provides an actionable alternative. This pragmatic approach demands one to deliberate the trade-offs in desired recall, desired precision, and available resources for acting upon presumed risks. In return, it provides a realistic understanding of the limits to RBI methods, returns appropriate data for improving future prediction performance, creates a strong incentive for schools to prioritize educational quality, and introduces an opportunity to evaluate the causal impact of inspections.

Acknowledgments

This research was supported by Dutch Inspectorate of Education. o.a.savi@gmail.com

References

Attenberg, J., & Ertekin, Ş. (2013). Class imbalance and active learning. In He, H. & Ma, Y., *Imbalanced Learning* (pp. 101–149). Hoboken, New Jersey: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118646106.ch6

Barber, M. (2005). The virtue of accountability: system redesign, inspection, and incentives in the era of informed professionalism. *Journal of Education*, *185*(1), 7–38. https://doi.org/10.1177/002205740518500102

Berner, A. (2017). *Would school inspections work in the United States?*. Baltimore, MD: Johns Hopkins Institute for Education Policy. https://jscholarship.library.jhu.edu/handle/1774.2/62992

Canton, E., & Webbink, D. (2004). Prestatieprikkels in het Nederlandse onderwijs. Wat kunnen we leren van recente buitenlandse ervaringen? (No. 49). Centraal Planbureau. Retrieved from https://www.cpb.nl/publicatie/prestatieprikkels-het-nederlandse-onderwijs-watkunnen-we-leren-van-recente-buitenlandse

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*. ACM Press. https://doi.org/10.1145/1143844.1143874

de Wolf, I. F., & Janssens, F. J. G. (2007). Effects and side effects of inspections and accountability in education: an overview of empirical studies. *Oxford Review of Education*, *33*(3), 379–396. https://doi.org/10.1080/03054980701366207

Dutch Inspectorate of Education. (2019). Inspection framework primary education 2017. Retrieved from https://english.onderwijsinspectie.nl/documents/publications/2017/06/21/inspectionframework-primary-educaton-2017

Ehren, M. C. M., Leeuw, F. L., & Scheerens, J. (2005). On the impact of the dutch educational supervision act. *American Journal of Evaluation*, *26*(1), 60–76. https://doi.org/10.1177/1098214004273182

Faubert, V. (2009). School evaluation: current practices in OECD countries and a literature review. Organisation for Economic Co-Operation; Development (OECD). https://doi.org/10.1787/218816547156

Flach, P. A., & Kull, M. (2015). Precision-recall-gain curves: PR analysis done right. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, Montreal, Canada: MIT Press.

Grek, S., Lawn, M., Ozga, J., & Segerholm, C. (2013). Governing by inspection? European inspectorates and the creation of a European education policy space. *Comparative Education*, *49*(4), 486–502. https://doi.org/10.1080/03050068.2013.787697

Grek, S., & Lindgren, J. (Eds.) (2014). *Governing by inspection*. New York: Routledge. https://doi.org/10.4324/9781315758091

Jacobusse, G., & Veenman, C. (2016). On selection bias with imbalanced classes, In Calders, T., Ceci, M. & Malerba, D., *Discovery science*. Berlin: Springer International Publishing. https://doi.org/10.1007/978-3-319-46307-0_21

Juba, B., & Le, H. S. (2019). Precision-recall versus accuracy and the role of large data sets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 4039–4048. https://doi.org/10.1609/aaai.v33i01.33014039

Ladd, H. F. (2010). Education inspectorate systems in New Zealand and the Netherlands. *Education Finance and Policy*, *5*(3), 378–392. https://doi.org/10.1162/edfp_a_00005

Ladd, H. F. (2010). Now is the time to experiment with inspections for school accountability. Retrieved from https://www.brookings.edu/blog/brown-center-chalkboard/2016/05/26/now-isthe-time-to-experiment-with-inspections-for-school-accountability/

Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., & Addison, K. L. (2015). A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. https://doi.org/10.1145/2783258.2788620

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Sanders, M., Lawrence, J., Gibbons, D., & Calcraft, P. (2017). Using data science in policy: a report by the Behavioural Insights Team. Behavioural Insights Team. Retrieved from https://www.bi.team/publications/using-data-science-in-policy/

Sedee, I. (2019). *Creating a bias in inspection data: exploring the medium- to long-term effects of data-driven risk-based regulation* [Master's thesis]. Delft University of Technology. http://resolver.tudelft.nl/uuid:4ee059f8-6923-4b9e-b807-08944027d2d5

Stassen, L. P. S., Bemelman, W. A., & Meijerink, J. (2009). Risks of minimally invasive surgery underestimated: a report of the Dutch health care inspectorate. *Surgical Endoscopy*, *24*(3), 495–498. https://doi.org/10.1007/s00464-009-0629-6.

Su, T.-L., Jaki, T., Hickey, G. L., Buchan, I., & Sperrin, M. (2016). A review of statistical updating methods for clinical prediction models. *Statistical Methods in Medical Research*, *27*(1), 185–197. https://doi.org/10.1177/0962280215626466

UNESCO. (2017). Education transforms lives. https://unesdoc.unesco.org/ark:/48223/pf0000247234

UNESCO Institute for Statistics and UNICEF. (2015). *Fixing the broken promise of education for all: findings from the global initiative on out-of-school children*. Montreal: UIS. https://doi.org/10.15220/978-92-9189-161-0-en

Wallace, B. C., & Dahabreh, I. J. (2012). Class probability estimates are unreliable for imbalanced data (and how to fix them). In *2012 IEEE 12th International Conference on Data Mining*. IEEE. https://doi.org/10.1109/icdm.2012.115

Weissbrodt, R., Arial, M., Graf, M., Iff, S., & Giauque, D. (2018). Preventing psychosocial risks at work: an evaluation study of labour inspectorate interventions. *Safety Science*, *110*, 355–362. https://doi.org/10.1016/j.ssci.2018.08.024