scientific reports

Check for updates

OPEN

Uncovering individualised treatment effects for educational trials

ZhiMin Xiao[1]✉, Oliver Hauser[2,7], Charlie Kirkwood[3,7], Daniel Z. Li[4], Tamsin Ford[5] & Steve Higgins[6]

Large-scale Randomised Controlled Trials (RCTs) are widely regarded as "the gold standard" for testing the causal effects of school-based interventions. RCTs typically present the statistical significance of the average treatment effect (ATE), which captures the effect an intervention has had *on average* for a given *population*. However, key decisions in child health and education are often about *individuals* who may be very different from those averages. One way to identify heterogeneous treatment effects across different individuals, not captured by the ATE, is to conduct subgroup analyses. For example, free school meal (FSM) pupils as required for projects funded by the Education Endowment Foundation (EEF) in England. These subgroup analyses, as we demonstrate in 48 EEF-funded RCTs involving over 200,000 students, are usually not standardised across studies and offer flexible degrees of freedom to researchers, potentially leading to mixed, if not misleading, results. Here, we develop and deploy an alternative to ATE and subgroup analysis, a machine-learning and regression-based framework to predict individualised treatment effects (ITEs). ITEs could show where an intervention worked, for which individuals, and to what extent. Our findings have implications for decision-makers in fields like education, healthcare, law, and clinical practices concerning children and adolescents.

Keywords Causal inference, Data science, Evaluation, Free school meal pupils, RCT, Subgroup analysis

Decision makers in healthcare and education often rely on Randomised controlled trials (RCTs) to test the causal effects of psychosocial and educational interventions[1–6]. While the use of RCTs in public policy, healthcare, and education has generally been advocated for by academics with some critiques, such as Deaton and Cartwright[7] and Biesta[8], the specific techniques to evaluate RCTs have received less attention. Indeed, only a fraction (less than 1%) of Education Endowment Foundation's (EEF) £200 million budget has been allocated to the understanding and advancement of evaluation methods. Typically, RCTs are evaluated—and ultimately judged—by the statistical significance of the Average Treatment Effect (ATE) that they achieve for participants in the treatment arm. Interventions that do not have a statistically significant ATE are often discarded as *not meaningful* and, as a result, usually not implemented more widely.

However, while some interventions may not have an effect on average, they might still have a meaningful effect on some individuals for whom the treatment is beneficial. In addition, key decisions in fields like education, healthcare, law, and clinical practices are often about individuals who are not hypothetical averages, which means the evidence that supports the decisions can be inappropriate or even harmful. Indeed, even trials that produce positive and statistically significant ATEs that are further supported by meta-analyses can have unintended consequences for some individuals, as evidenced in recent trials on statin side effects[9] in medicine. In the largest trial conducted so far to evaluate the effects of school-based mindfulness training on risk of depression and wellbeing in early adolescence, the overall effects were found to be detrimental to adolescents in need of mental health support, despite no ATE and a subgroup of young people who improved their mindfulness and executive skills after the intervention demonstrating improved mental health outcomes[10]. These problems are compounded by the fact that ATEs from large-scale and rigorously conducted and evaluated RCTs often produce small effect sizes sitting within wide confidence intervals[11], evidence that is hardly actionable for decision makers. This also suggests, ATEs can be positive, even when most participants did not benefit[12]. Therefore, understanding and identifying for whom an intervention works is critical for policy-makers and society at large[12–14].

[1]School of Health and Social Care, University of Essex, Colchester CO4 3SQ, UK. [2]Department of Economics, University of Exeter, Exeter EX4 4PU, UK. [3]Department of Mathematics, University of Exeter, Exeter EX4 4QF, UK. [4]Durham University Business School, Durham DH1 3LB, UK. [5]Department of Psychiatry, University of Cambridge, Cambridge CB2 0AH, UK. [6]School of Education, Durham University, Durham DH1 1TA, UK. [7]Institute for Data Science and Artificial Intelligence, University of Exeter, Exeter EX4 4QF, UK. ✉email: zhimin.xiao@essex.ac.uk

Conventionally, researchers rely on subgroup analyses to identify individuals who benefit in trials. Subgroup analyses, in its various forms, usually come with several drawbacks. For instance, choosing which subgroups to analyse, and how to conduct these analyses, which affords multiple degrees of freedom, and risks researchers selectively reporting results that are supportive of their overall conclusion[15,16]. This is essentially "statistical malpractice"[17]. If this occurs, findings may be published that are not reproducible by other researchers, or they may not hold up in direct replication studies, which can shake and erode public trust in science[18]. As such, subgroup analyses are as controversial as they are important: researchers "are damned if they do, and damned if they don't"[17] include subgroup analyses in their research.

The EEF, as part of their funding scheme, requires that researchers conduct a subgroup analysis of the treatment for the socioeconomically disadvantaged group of Free School Meal (FSM) pupils. FSM status is one of the most frequently used variables to approximate socioeconomic status in the UK[19,20]. This reporting requirement by the EEF enables us to study current practices in subgroup analyses since it is not specified (or generally agreed upon) how such analyses ought to be conceived and conducted. Previous researchers have cautioned against potential overinterpretation of effects for subgroups like FSM students[21–24]. Although some challenges of subgroup analyses, such as lack of statistical power and unreliable estimation, can be partially alleviated by standardised reporting[21–24], there are insufficient statistical details in the guidance of key organisations on how to exactly estimate and interpret subgroup effects[25].

Building on recent insights from data science and machine learning, we propose an alternative to ATEs and subgroup analyses in conventional large-scale RCTs in education, which are an excellent area of research, because policy-makers are keen to improve educational practices and ultimately, equity in pupil attainment and wellbeing and, school-based RCTs, unlike many other policy areas, have become increasingly popular[26–29]. Our approach combines the pragmatism of subgroup analyses, by allowing researchers to learn what works for whom, with the security of relying on a robust and replicable methodology. On the one hand, it is flexible and adaptable to a large breadth of covariates in any given RCT, allowing researchers to study individualised treatment effects (ITEs) from "bottom-up"[30,31], thus avoiding false positives associated with less rigorous subgroup analyses and unrealistic assumptions often made in the estimation of conventional ATEs. On the other hand, by introducing a principled procedure a priori, it ensures that findings are as robust, reproducible, and transparent as possible. Thanks to recent development in data science and the increase in both quantity and quality of research data, it is easier than before to predict and compare treatment effects at individual and group levels, and increasingly possible to marshal and display telling details on any individual in a given trial, "for the perusal of that individual"[9], ultimately generating deeper insights into the causal treatment effects of tested interventions[32].

Results

An individualised approach to effect estimation

ATE has long been a quantity of interest in RCTs. However, "the response of the average patient to therapy is not necessarily the response of the patient being treated"[33]. Strong scientific interests in effect heterogeneity do exist, but detecting such variation is not always straightforward in increasingly complex designs. Many EEF trials are, for example, not ideally powered to detect the main effect due to challenges associated with sample size calculations[34] and resource constraints. As a result, as we will demonstrate later, current methods to identify individuals who benefited from an intervention via subgroup analyses are not always helpful. Unsurprisingly, estimates of effects for FSM students in EEF trials usually come with the caveat, "should be interpreted with caution".The alternative approach we propose here taps into the advancement of predictive algorithms in machine learning[35–38]. It focuses on the differences between two potential outcomes[38–40], often an observed factual and an unobservable counterfactual outcome[41,42]. We call the differences individualised treatment effects for a given individual in a trial that has taken place (see Fig. 1 for an illustration).

The evidence generated from the individualised approach is thus unique to specific individuals according to their observed characteristics. For each individual in an educational RCT that has taken place for instance, we observe the intervention arm assigned $T \in \mathcal{T}$, the intervention outcome $Y \in \mathbb{R}$, and the student's pre-intervention and school-level characteristics, all captured by an m-dimensional vector $\mathbf{X} \in \mathcal{X}$. (In line with published works[33,38,40,43], we use upper- and lower-case letters to denote random variables and their realised values respectively, and bold letters to represent vectors in the data.) For simplicity, we let $T = 1$ denote the treatment group and $T = 0$ the control or business-as-usual group, hence $T \in \{0, 1\}$. Without loss of generality, we assume a higher value of Y suggests a better outcome.

The ATE in a conventional RCT evaluation is defined as $\mathbb{E}(Y|T = 1) - \mathbb{E}(Y|T = 0)$. When calculating ITE, we have an observed factual outcome, and need to predict a *counterfactual* outcome, had the student been assigned to an alternative intervention arm. That is to say, for a student with covariates $\mathbf{x}$ in the treatment group $T = 1$, we observe the factual treatment outcome, denoted as $y_1(\mathbf{x})$. We utilise the observed data to predict the unobservable counterfactual outcome, denoted as $\hat{y}_0(\mathbf{x})$, and calculate ITEs as $y_1(\mathbf{x}) - \hat{y}_0(\mathbf{x})$.

A number of models can be deployed to predict potential outcomes. Given a pre-specified model $f(\mathbf{x}; T)$ and the observed covariates $\mathbf{x}$, the true data generation process for the student i is

$$Y_i = \underbrace{f(\mathbf{X}_i, T_i) + \xi_i(\mathbf{X}_i, \mathbf{U}_i, T_i)}_{\mathbb{E}[Y_i|\mathbf{X}_i, \mathbf{U}_i, T_i]} + \varepsilon_i,$$

where the first two terms represent the true conditional expectation, and the last term is the irreducible error around it[33]. Unfortunately, the true data generation process is unknown, even in the absence of ε_i, as f will always differ from the true process by $\xi_i(\mathbf{X}_i, \mathbf{U}_i, T_i)$, where $\mathbf{U}_i \in \mathcal{U}$ represents *unobserved* and *unobservable* covariates,

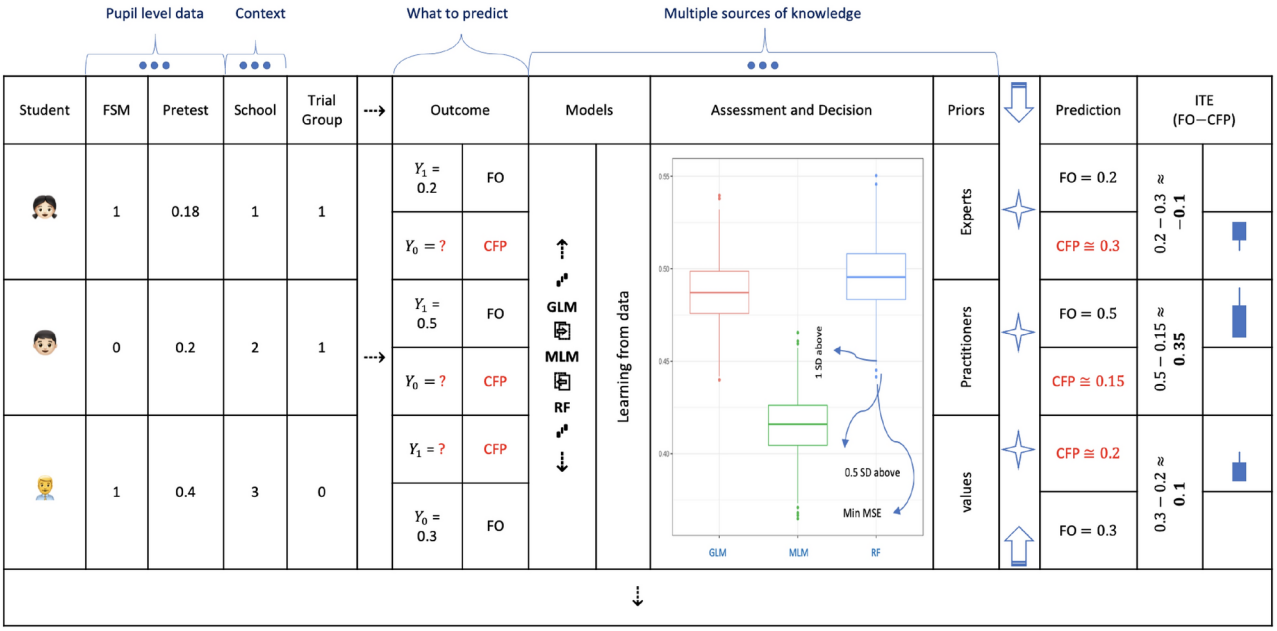

Fig. 1. ITEs through a graphical journey. With student- and school-level data, such as FSM status and Pretest for the former, and School ID for the latter, multiple models compete on the same data to predict counterfactual outcomes (CFP) for individual students. FO is the factual outcome observed under the treatment condition a student actually received in an intervention. The difference between FO and CFP is used to re-train the best performing model for the re-prediction of ITEs, which are then visualised as bar plots for unique individuals. Note that the generation process for ITEs is iterative and the arrows pointing to four directions represent the hidden procedures. There are also many more variables and even more observations, which are visualised as dots and dotted arrows.

such as culture, tradition, and value-based decisions. That is, we cannot gather data on all possible covariates in a study, and we also need to recognise that researchers bring different assumptions into the data generation process at collection, pre-processing, and analysis stages[44]. Yet, given sufficient covariates, these models *can* yield insightful information about ITEs.

As an illustration of the individualised approach, we focus on one EFF-funded dataset from a trial called Chess in Schools[45]. This project had 100 schools randomly assigned to either intervention or control, involving 4009 pupils. Intervention schools taught children how to play chess over a year, whereas control schools were business-as-usual. The primary outcome was Key Stage 2 Maths score—an important standardised test taken by UK pupils usually at age 11—1 year after the intervention.

We highlight the above project for the following reasons. First, it has a relatively large sample size and the highest possible security rating of five padlocks, which indicates high internal and external validity[46]. Second, at a total cost of £689,150, the project represents some intensive efforts to improve maths skills, yet the reported overall effect size of $0.01(-0.15, 0.16)$[45] might suggest that it was not a worthwhile investment, prompting the question whether this intervention did have some non-negligible and educationally meaningful effects for some students, given the report that 50% of the pupils said "they liked the chess lessons a lot" and "teachers were very positive about the intervention and its impact on pupils' skills and behaviour"[45]. Finally, the dataset has many observed covariates, which makes it suitable for this individualised approach. Any other RCT that fulfils the above criteria would make an ideal candidate for this approach.

Using generalised linear models (GLM), multilevel linear models (MLM), and random forests (RF) as candidate models, we made a chain of predictions for our ultimate quantity of interest, ITEs, of which two distributions are shown in Fig. 2A. For further details about model selection and discussions on the target of prediction, please see the Methods section.

With ITEs from the optimal model, namely, a re-trained RF, we visualised in Fig. 2B the proportions of FSM and Non-FSM students at different ITE thresholds, which were set according to what we knew from the literature and understood about the trial under investigation. For instance, either the overall effect size or that for FSM students reported by the original evaluation team is 0.01[45], and the average of all the EEF trials funded to date is about 0.05[11]. Other values in the figure reflect the distribution of our predicted ITEs. We found that 36% of the FSM students in the study benefited by at least 0.01 *sd* from the Chess intervention, whereas 27% of the Non-FSM students gained by the same amount. The pattern continues, such that more FSM students benefited from the intervention than Non-FSM students, up to 0.05 *sd*. A small fraction, equally made up of FSM and non-FSM students, benefited by 0.15 *sd*. On the negative side, fewer FSM students were worse off than their Non-FSM counterparts by 0.1 or 0.15 *sd*. Taken together, the individualised approach demonstrated that the intervention

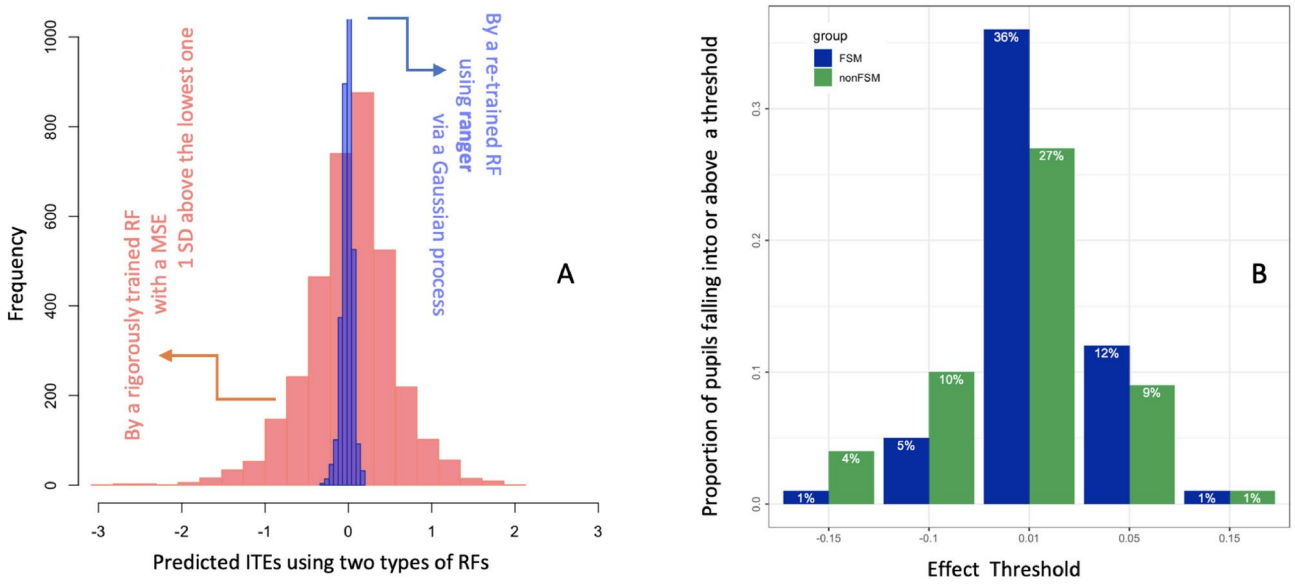

Fig. 2. ITEs for use. (**A**) Histograms of two distributions of ITEs from two types of RF, one predicting one potential outcome and the other re-trained to predict the differences of two, i.e., ITEs. (**B**) ITE thresholds used to visualise the proportions of FSM and Non-FSM students predicted to benefit (if positive) or lose (if negative) from the intervention by at least a chosen amount.

was more beneficial to FSM students than to Non-FSM students, amplifying the positive feedback in the form of qualitative data given by the pupils and teachers actually involved in the trial[45].

While the above statistics are telling of the benefits, it is important to note the uncertainties, either epistemic or aleatoric, surrounding those percentages. The point demonstrated above is not about an *absolute* reality of how many, *precisely*, FSM or Non-FSM students benefited from the intervention by how much, *exactly*. Different models, no matter how well calibrated they are via cross-validation or bootstrapping, could generate different distributions of ITEs that might result in different comparison statistics. The resultant statistics might also differ had we chosen different prior knowledge about the effect thresholds. The point we wanted to demonstrate is that significantly more FSM students than their Non-FSM counterparts benefited from the seemingly ineffective intervention as measured and reported using the official ATE, after we incorporated *some* uncertainties associated with model calibration and prior knowledge informed by the literature to date.

When acting on or interpreting the ITEs, we may incorporate non-statistical knowledge. For example, according to Fig. 3A, baseline measure k1m, or KS1 maths score, is the most important predictor of ITEs, which is followed by APS (KS1 Average Point Score) and FSP (Foundation Stage Profile total score), both extracted from the National Pupil Database. Given this information, educators and decision-makers who know the students and/or their schools best can intervene by focusing on what is practically possible and most important to do and when. However, the variables ranked in the importance plot are observed covariates in the data, it does not mean that no other factors that were not observed in the study are at least as important as k1m.

In other studies, the most important variable observed may not be educationally desirable[8]. Suppose corporal punishment, which is easy to measure, is one of the strongest predictors of ITEs, it does not mean that we should implement that policy. This is exactly when non-statistical knowledge and educational, social, and psychological theories have important roles to play in identifying variables to use for the prediction and helping interpret the ITEs then produced for policy use.

Suppose again decision-makers will listen to us and choose to intervene by investing more in maths education at KS1. They want to see by how much an individual would be (made) better or worse off, had the student's performance on the subject been different as a result of the investment. And what would the effect of the intervention be like if the student is eligible for FSM?

To answer the questions posed above, we visualised the effects for the individual under different hypothetical scenarios in Fig. 3B by following an individual conditional expectation (ICE) plot procedure, which, in the words of the authors[47], allowed us to "peek inside the black box" of the RF, and see how ITEs vary with the covariates that characterise each individual. With the assumption that the RF has successfully learned the relationships present in the dataset, we can for each individual vary each covariate one at a time through its full minimum-to-maximum range, and see how ITEs are expected to change for each individual "if they had this character instead".

Figure 3B shows the expected influence of KS1 maths score on ITEs at an individual level (each individual in the dataset has her/his own line). We can see that, in a general sense, ITEs tend to be *associated with* FSM and k1m. ITEs are higher at higher k1m scores. By also differentiating the lines based on the FSM status of an individual, we can observe an interaction effect between FSM and k1m, and it appears that FSM individuals with k1m scores between 10 and 20 would benefit more from the intervention than their Non-FSM counterparts.

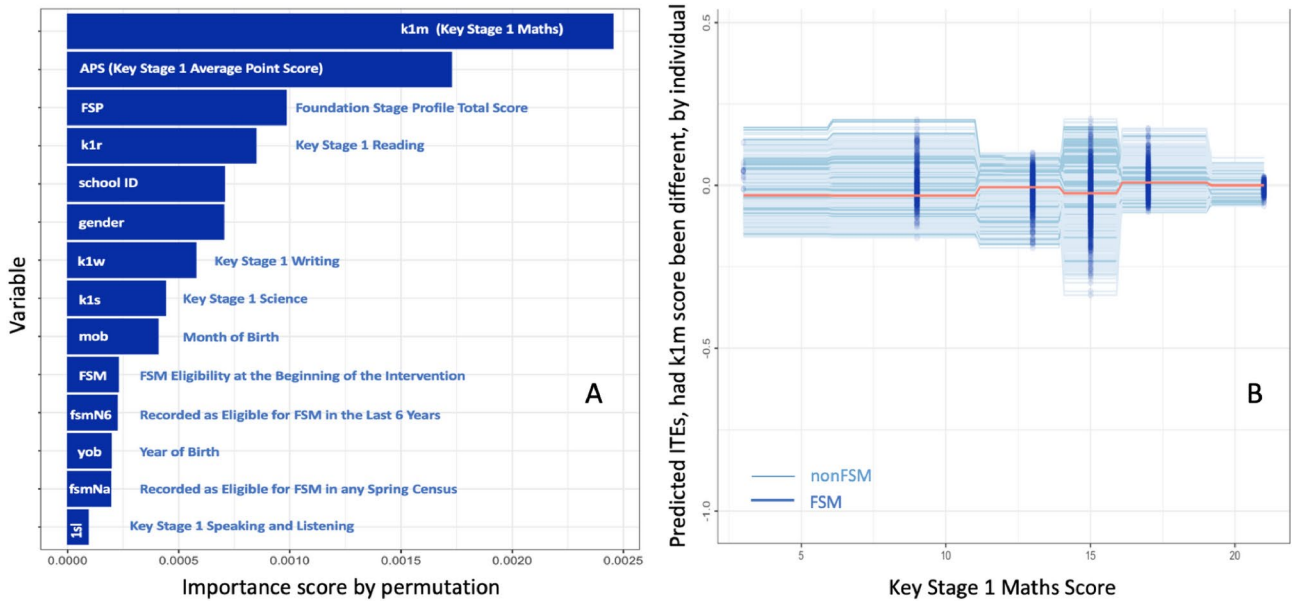

Fig. 3. ITEs and non-statistical knowledge. (**A**) The re-trained RF algorithm allows us to see the most important predictor of ITEs, which, in this case, is Key Stage 1 Maths. (**B**) Expected ITEs for different individuals, be they FSM or Non-FSM students, had their performances on Key Stage 1 Maths been different. The orange line represents an average student.

The highlighted words "associated with" imply that it would be ethically impractical to make a student eligible for FSM even it is practically possible to do so. Similar ethical and perhaps legal considerations apply when other covariates such as gender and ethnic backgrounds are involved, although an algorithm might suggest that those covariates would be important to predict ITEs.

Once again, we have shown that the individualised approach is a collaborative endeavour to evidence generation and use and it involves multiple stakeholders in education. To best serve the students we care about most, it invites academics from diverse backgrounds and people with domain knowledge and professional experiences to collaborate and co-produce evidence that is relevant to individuals and groups alike and actionable for policy-makers and practitioners.

Conventional approaches to subgroup analyses

To justify the need for an alternative approach to effect size estimation (re-)using data from large-scale RCTs, we must show how conventional ways of analysing the data pose substantial challenges. Following an established approach to subgroup analysis in the evaluation of EEF trials, we also employed, as specified in Supplementary Information (SI), an ordinary least squares (OLS) model, which has as covariates, a treatment-FSM interaction term plus pre-test, a baseline measure available in almost all EEF trials. For each outcome, we obtained the sample size used for the interaction test and the *p*-value associated with the interaction term. We then deployed a multilevel linear model (MLM) recommended by the EEF, also documented in SI, to estimate effect sizes within the subgroups of FSM and Non-FSM students using eefAnalytics, an R package specifically developed to estimate effect sizes for RCTs funded by the EEF. The two subgroup effect estimates were then compared with the *p*-values from the interaction tests, which indicate if the differences between the two separately estimated subgroup effect sizes are statistically significant.

In total, we examined 84 outcomes from 48 projects (see Table S1 for details), which are distinct and separate projects designed and evaluated by independent teams in different years. As we had access to the raw data, this analysis differs from a standard meta-analysis. Instead, we reported and calculated, for each outcome (see Figs. S1–S13), three effect size estimates, their associated sample sizes and 95% confidence intervals. The first, for reference only, is the overall effect size every EEF project reported for all the students involved in each trial. The other two are separate estimates produced by us for Non-FSM and FSM students.

The results of our re-analysis show that, only 6 out of the 84 outcomes are statistically significant for FSM students, which means, consistent with the literature on overall effect sizes of EEF studies[11], conventional regression analyses focusing on ATEs often produce results that are non-actionable, even when participants' lived experiences as reported in process evaluations show otherwise[45]. One reason, in addition to those given elsewhere[7,11], this may be the case is that we conducted the same analysis across all studies, thereby holding all analyses to the same standards. Arguably, of course, our analyses following some conventional approaches to subgroup analyses are not necessarily the only "correct" ones. However, this gives further credence to the need for alternatives to impact evaluation and better (re-)use of research data from trials we have invested so much in.

Discussion

In line with previous research, we proposed an individualised approach here to effect estimation, which employed three types of predictive algorithms to first predict how individuals would respond to different treatment options. We then quantified the differences in potential outcomes at individual level, which were subsequently used to re-train the best of the three candidate models to predict the ultimate quantity of interest, namely, ITEs for unique individuals. These highly individualised effect predictions can be utilised to evaluate an intervention that has taken place, thus making the best use of research data and offering an alternative to conventional ATE and subgroup analyses.

The ITEs can be further examined with other variables, such as FSM status, a variable used for subgroup analyses in EEF trials. While we focus on ITEs, it is still important to see if and how they converge with the ATE when aggregated, as schools and policy-makers in education and healthcare often need to consider participants in groups, be they classes, year groups, schools, or districts. We demonstrated that the results from the individualised approach could reproduce the original ATE *and* answer policy-relevant questions about subgroups of students, such as those from disadvantaged backgrounds who experience worse education.

In sum, we have shown that, while conventional approaches to effect size estimation in policy interventions often resort to aggregated measures of impact, such as ATE or conditional ATE[43,48], these evaluations take little note of individual characteristics that may alter individual responses to interventions. After all, an intervention that worked well *on average* may not be the best option for all, while a substantial benefit for some might be worth having, even if there was no ATE[49]. Any subgroup analysis, or heterogeneous treatment effects analysis, is one step closer towards a more sensitive estimation of individual responses to an intervention[50], but current practices are not standardised. Heterogeneous causal effect prediction using causal forests permits *statistical* inferences and quantifies their uncertainties, thanks to its asymptotic properties associated with a normality theory[40,43]. Nevertheless, it is worth pointing out that no method can lead to a truly individualised effect[51,52]: no individual can ever be in both the control and the treatment arms at the same time. However, the methods proposed here and elsewhere by others[38] can get closer to an ideal individualised effect. Future research will continue to develop consistent and efficient ways to construct these predictions, as the current process is rather "data hungry", often with half-half sampling split for feature space construction and effect estimation. As Wager & Athey note, however, the sampling split rule is arbitrary and "still in its infancy"[40].

As in any research, our individualised approach has its own limitations. First, while it focuses on evidence that is actionable, the prediction procedures could have been more dynamic (i.e., each individual may have their own best prediction algorithm, rather than *a* best performing algorithm for all individuals) and the uncertainties surrounding those ITEs are yet to be formalised and refined. Second, we could have simulated some data from an RCT where the intervention effect is zero and no relationships exist amongst covariates. If the predictive algorithms fail to identify individuals who benefit from the "trial", we would be much more confident about the approach. It would be even better if we can test the approach in multi-phase trials and/or deploy it to predict outcomes in longitudinal studies.

Finally, while our approach has shown to work effectively in education, it is worth emphasising that it can be applicable and relevant in any area of science and policy—from tax collection to medical trials to public health. The importance of individualising treatment effects is particularly critical in areas of rapid development, such as testing of drugs or vaccines in N-of-1 trials[53], or of policies that encourage certain behaviours for the public good[54]. Take, for example, the COVID-19 crisis: first, vaccine trials might benefit from the individualised approach, in that even a vaccine that only works for a relevant subgroup (e.g., patients with underlying health conditions) would be a much-needed advancement to battle the deadly disease; second, encouraging social distancing may take different forms and policy-makers would be well-advised to understand how different subgroups (e.g., the elderly, young people, and key workers) might respond to different messaging.

Methods

For the demonstration, the Chess in Schools dataset we constructed has 16 variables, including the Key Stage 2 Maths outcome measure and treatment indicator. Pre-treatment covariates consist of Key Stage 1 measures and pupil-level characteristics such as FSM status. To be consistent with the way most EEF evaluators dealt with missing data in their primary analyses, we also removed all the rows with missing data, which is less problematic to machine learning algorithms such as random forests than it is to inferences based on probability theories[55]. We ended up with a sample of 3514 complete cases, which is unsurprisingly smaller than 3,865 the evaluation team reported and 3695 for the interaction test in Fig. S1, as the variables used to construct the data are different.

To conduct model selection, we first randomly split, via bootstrapping, the observed RCT data into two disjoint training and testing subsets. The training set was used to train candidate models, and the testing set to assess their performances. In each bootstrapped re-sample, an outcome of interest was predicted by each model and for each student. An average error, namely mean squared error (MSE), in prediction across all the individuals in the testing set was produced. This process was repeated 1000 times to generate a distribution of the MSEs for each model. We then chose one type of model, taking into consideration its variation in prediction, that showed the *best* performance on average in the testing set for the final prediction of ITEs. Note that the highlighted word "best" here and elsewhere in the paper does not suggest the best of all possible models, it only means the best amongst the models deployed, i.e., GLM, MLM, and RF. The GLM for prediction here differs from the earlier OLS model for interaction test, as the former has more variables than the latter in order to predict well and is less concerned about collinearity and the coefficient of a particular variable. GLM and MLM are the primary evaluation models in almost all EEF trials, it is natural to compare them with RF, which have excellent performances in prediction[40,43,51,56,57].

Based on the Chess in Schools data, MLM has an average MSE of 0.42 and a standard deviation (*sd*) of 0.02, and those of GLM and RF are 0.49 $(sd = 0.02)$ and 0.5 $(sd = 0.02)$, respectively. The distributions of the prediction errors for the three types of models employed are illustrated in Fig. S14. We should choose MLM when the target of prediction is post-test score at individual level. Amongst the 1000 sets of parameters for MLM, the 854th has the lowest out-of-sample prediction error, as reported in Table S2. That set of parameters should be used to make factual and counterfactual predictions for each individual. The factual and counterfactual datasets differ only in treatment status, where the factual values represent the real random assignment in the RCT, and their counterfactual values are the opposite of the actual random assignment. In other words, if a student was randomly assigned to the intervention group, the value for treatment indicator is 1, and its counterfactual value is 0, as if the student were assigned to the business-as-usual group.

The distribution of the 1000 prediction errors of MLM, as shown in Fig. S14, has some outliers. Choosing the best set of parameters with the lowest prediction error may produce biased results in the prediction. To take into consideration the variation in prediction errors, we examined three sets of parameters in MLM, one with the lowest prediction error (854th), the other two closest to half (792th) and one (805th) *sd* of the 1,000 MSEs above the set with the lowest error. Nevertheless, as reported in Table S2, the choice about which set to use for the prediction of post-test outcomes does not make much difference in the results. But when error distributions have extreme outliers in other studies, as shown in Figs. S15 and S16, the choice will make a substantial difference.

Given the data, we have one observed outcome, which we call Factual Outcome (FO), for each student. We can also use the best model to make two predictions for each individual, we call those predicted outcomes factual prediction (FP) and counterfactual prediction (CFP), with the former being the outcome predicted under the factual treatment condition a student actually allocated to, and the latter being the outcome predicted under the alternative treatment condition the student could have been allocated to. This means there are two ways to calculate ITEs, one being the difference between FO and CFP, and the other being FP and CFP. We assessed the performances of both approaches. Since we converted the outcome variable into *z*-scores, in either way, the ITEs computed are comparable to the reported effect sizes.

When ITEs were calculated as the differences between factual and counterfactual predictions (FP - CFP), they were constant or pre-determined by the chosen linear models. When we used the differences between the factual outcomes (FO) and counterfactual predictions (CFP) to calculate ITEs, the predictions reflected real-world uncertainties (the *sds* are much larger in the FO - CFP rows of Table S2).

For the above reasons, we employed RF to predict ITEs, as they were less deterministic and more responsive to individual differences than the other two types of linear models. However, when the target of prediction was a potential outcome, rather than the difference of two potential outcomes, the ranges of ITEs were too wide, as shown in Fig. 2A. This means the chosen RF have not learned enough about the data, particularly, the relationships between the differences we are interested in and other covariates. This empirically explains the need to focus on the effects rather than the prediction of potential outcomes alone for causal inference[40,43,58].

In order to reduce the variation in the prediction of ITEs, we re-trained the RF by fine-tuning the hyperparameters of the algorithm using its out-of-bag (OOB) error estimates. The speed of the R package, ranger[59], allowed us to conduct a random grid search on the three main hyperparameters: *mtry*—the number of variables sampled for splitting at each node; sample fraction – the proportion of the full dataset provided to each tree for training (through bootstrapping); and minimum node size – the minimum number of data points in each terminal node acting as a regulariser on each tree. To increase our confidence in selecting an optimal set of hyper-parameters in what is an inherently noisy system (due to the random sampling of the RF algorithm), we used heterogeneous Gaussian process regression[60] to smooth out the noise, and selected the hyperparameters that minimised the upper confidence bound on OOB error. As Fig. 2A shows, the ITEs from the re-trained RF have a narrower spread than those from the RF mentioned earlier.

Data availibility

As part of their funding scheme, the EEF requires all evaluation teams to submit their data to a central archive, which is managed by FFT Education and held by the ONS within their Secure Research Service. FFT provided us with 48 unique data extracts from large-scale RCTs of varied designs. The data were linked with the National Pupil Database in England, but de-identified at pupil and school levels. The 48 datasets that support the findings of this study are available from the EEF, but restrictions apply to the availability of these data. Please refer to this link about access to the data: https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/approvedresearcherscheme or contact FFT Education (via this link https://fft.org.uk/about-fft/) to request access to the data. Because the data ownership lies with the EEF/FFT/ONS, they make ultimate decisions on who to grant access.

Received: 24 November 2023; Accepted: 20 September 2024
Published online: 30 September 2024

References

1. Banerjee, A. *et al.* A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science* https://doi.org/10.1126/science.1260799 (2015).
2. Dillon, M. R., Kannan, H., Dean, J. T., Spelke, E. S. & Duflo, E. Cognitive science in the field: A preschool intervention durably enhances intuitive but not formal mathematics. *Science* **357**, 47–55. https://doi.org/10.1126/science.aal4724 (2017).
3. Higgins, S. *Improving Learning: Meta-analysis of Intervention Research in Education* (Cambridge University Press, 2018).
4. Rogers, T. & Feller, A. Reducing student absences at scale by targeting parents' misbeliefs. *Nat. Hum. Behav.* **2**, 335–342. https://doi.org/10.1038/s41562-018-0328-1 (2018).
5. Thaler, R. H. & Sunstein, C. R. *Nudge: Improving Decisions About Health, Wealth, and Happiness* (Yale University Press, 2008).
6. John, P. *et al.Nudge, Nudge, Think, Think: Experimenting with Ways to Change Civic Behaviour* (Bloomsbury Academic, 2011).

7. Deaton, A. & Cartwright, N. Understanding and misunderstanding randomized controlled trials. *Soc. Sci. Med.* **210**, 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005 (2018).
8. Biesta, G. J. J. Why 'What Works' Still Won't Work: From evidence-based education to value-based education. *Stud. Philos. Educ.* **29**, 491–503. https://doi.org/10.1007/s11217-010-9191-x (2010).
9. Howard, J. P., Wood, F. A. & Francis, D. P. Why do i get side effects? Personalized (N-of-1) trials for statin intolerance and the nocebo effect. *Harv. Data Sci. Rev.* https://doi.org/10.1162/99608f92.abc57f1b (2022).
10. Montero-Marin, J. *et al.* School-based mindfulness training in early adolescence: What works, for whom and how in the MYRIAD trial?. *Evid. Based Ment. Health* **25**, 117–124. https://doi.org/10.1136/ebmental-2022-300439 (2022).
11. Lortie-Forgues, H. & Inglis, M. Rigorous large-scale educational RCTs are often uninformative: Should we be concerned?. *Educ. Res.* **48**, 158–166. https://doi.org/10.3102/0013189X19832850 (2019).
12. Husain, M. Time for N-of-1 trials in clinical decision-making. *Brain* **144**, 1031–1032. https://doi.org/10.1093/brain/awab107 (2021).
13. Athey, S. Beyond prediction: Using big data for policy problems. *Science* **355**, 483–485. https://doi.org/10.1126/science.aal4321 (2017).
14. Mirza, R. D., Punja, S., Vohra, S. & Guyatt, G. The history and development of N-of-1 trials. *J. R. Soc. Med.* **110**, 330–340. https://doi.org/10.1177/0141076817721131 (2017).
15. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366. https://doi.org/10.1177/0956797611417632 (2011).
16. Xiao, Z., Kasim, A. & Higgins, S. Same difference? Understanding variation in the estimation of effect sizes from educational trials. *Int. J. Educ. Res.* **77**, 1–14. https://doi.org/10.1016/j.ijer.2016.02.001 (2016).
17. Petticrew, M. *et al.* Damned if you do, damned if you don't: Subgroup analysis and equity. *J. Epidemiol. Commun. Health* **66**, 95–98. https://doi.org/10.1136/jech.2010.121095 (2012).
18. Wingen, T., Berkessel, J. B. & Englich, B. No replication, No Trust? How low replicability influences trust in psychology. *Soc. Psychol. Personal. Sci.* https://doi.org/10.1177/1948550619877412 (2019).
19. Hobbs, G. & Vignoles, A. Is children's free school meal â€˜eligibility' a good proxy for family income?. *Br. Edu. Res. J.* **36**, 673–690. https://doi.org/10.1080/01411920903083111 (2010).
20. Strand, S. School effects and ethnic, gender and socio-economic gaps in educational achievement at age 11. *Oxf. Rev. Educ.* **40**, 223–245. https://doi.org/10.1080/03054985.2014.891980 (2014).
21. Assmann, S. F., Pocock, S. J., Enos, L. E. & Kasten, L. E. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *The Lancet* **355**, 1064–1069. https://doi.org/10.1016/S0140-6736(00)02039-0 (2000).
22. Lagakos, S. W. The challenge of subgroup analyses—Reporting without distorting. *N. Engl. J. Med.* **354**, 1667–1669. https://doi.org/10.1056/NEJMp068070 (2006).
23. Song, F. & Bachmann, M. Cumulative subgroup analysis to reduce waste. *BMC Med.* **14**, 1–8. https://doi.org/10.1186/s12916-016-0744-x (2016).
24. Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J. & Drazen, J. M. Statistics in medicine: Reporting of subgroup analyses in clinical trials. *N. Engl. J. Med.* **357**, 2189–2194. https://doi.org/10.1056/NEJMsr077003 (2007).
25. Wijn, S. R. W. *et al.* Guidance from key organisations on exploring, confirming and interpreting subgroup effects of medical treatments: A scoping review. *BMJ Open* **9**, e028751. https://doi.org/10.1136/bmjopen-2018-028751 (2019).
26. Connolly, P., Keenan, C. & Urbanska, K. The trials of evidence-based practice in education: A systematic review of randomised controlled trials in education research 1980–2016. *Educ. Res.* **60**, 276–291. https://doi.org/10.1080/00131881.2018.1493353 (2018).
27. Parker, K., Nunns, M. P., Xiao, Z., Ford, T. & Ukoumunne, O. C. Characteristics and practices of school-based cluster randomised controlled trials for improving health outcomes in pupils in the UK: A systematic review protocol. *BMJ Open* **11**, 1–17. https://doi.org/10.1136/bmjopen-2020-044143 (2021).
28. Parker, K., Nunns, M., Xiao, Z. M., Ford, T. & Ukoumunne, O. C. Characteristics and practices of school-based cluster randomised controlled trials for improving health outcomes in pupils in the United Kingdom: A methodological systematic review. *BMC Med. Res. Methodol.* **21**, 1–17. https://doi.org/10.1186/s12874-021-01348-0 (2021).
29. Parker, K., Nunns, M., Xiao, Z. M., Ford, T. & Ukoumunne, O. C. Intracluster correlation coefficients from school-based cluster randomized trials of interventions for improving health outcomes in pupils. *J. Clin. Epidemiol.* **158**, 18–26. https://doi.org/10.1016/j.jclinepi.2023.03.020 (2023).
30. Nguyen, T. L., Collins, G. S., Landais, P. & Le Manach, Y. Counterfactual clinical prediction models could help to infer individualized treatment effects in randomized controlled trialsâ€"An illustration with the International Stroke Trial. *J. Clin. Epidemiol.* **125**, 47–56. https://doi.org/10.1016/j.jclinepi.2020.05.022 (2020).
31. Efthimiou, O. *et al.* Measuring the performance of prediction models to personalize treatment choice. *Stat. Med.* https://doi.org/10.1002/sim.9665 (2023).
32. Horwitz, R. I. & Singer, B. Introduction. What works? And for whom?. *Soc. Sci. Med.* **210**, 22–25. https://doi.org/10.1016/j.socscimed.2018.05.013 (2018).
33. Kapelner, A. *et al.* Inference for the Effectiveness of Personalized Medicine with Software. arXiv:1404.7844 (2014)
34. Schulz, K. F. & Grimes, D. A. Sample size calculations in randomised trials: Mandatory and mystical. *Lancet* **365**, 1348–1353. https://doi.org/10.1016/S0140-6736(05)61034-3 (2005).
35. Breiman, L. Statistical modeling: The two cultures. *Stat. Sci.* **16**, 199–231. https://doi.org/10.1214/ss/1009213726 (2001).
36. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning: with Applications in R* 6th edn. (Springer, 2013).
37. van der Laan, M. J. & Rose, S. *Targeted learning in data science: Causal inference for complex longitudinal studies* (Springer, 2018).
38. Alaa, A. M. & van der Schaar, M. Validating causal inference models via influence functions. In *36th International Conference on Machine Learning, ICML 2019* (Long Beach, California, 2019).
39. Rubin, D. B. Estimating causal effects of treatment in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688–701 (1994).
40. Wager, S. & Athey, S. Estimation and inference of heterogeneous Treatment effects using random forests. *J. Am. Stat. Assoc.* **113**, 1228–1242. https://doi.org/10.1080/01621459.2017.1319839 (2018).
41. Hernán, M. A., Hsu, J. & Healy, B. A second chance to get causal inference right: A classification of data science tasks. *Chance* **32**, 42–49. https://doi.org/10.1080/09332480.2019.1579578 (2019).
42. Pearl, J. & Mackenzie, D. *The Book of Why: The New Science of Cause and Effect* (Penguin, 2019).
43. Athey, S. & Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7353–7360. https://doi.org/10.1073/pnas.1510489113 (2016).
44. Meng, X.-L. Dissecting multiple imputation from a multi-phase inference perspective: What happens when God's, imputer's and analyst's models are uncongenial?. *Stat. Sin.* **27**, 1485–1594. https://doi.org/10.5705/ss.2014.067 (2017).
45. Jerrim, J., Macmillan, L., Micklewright, J., Sawtell, M. & Wiggins, M. *Chess in Schools* (Tech. Rep, 2016).
46. Higgins, S. *et al.The Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit* (Tech. Rep, Education Endowment Foundation, 2015).
47. Goldstein, A., Kapelner, A., Bleich, J. & Pitkin, E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Stat.* **24**, 44–65. https://doi.org/10.1080/10618600.2014.907095 (2015).

48. Lamont, A. *et al.* Identification of predicted individual treatment effects in randomized clinical trials. *Stat. Methods Med. Res.* https://doi.org/10.1177/0962280215623981 (2016).
49. Schnell, P. M., Tang, Q., Offen, W. W. & Carlin, B. P. A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. *Biometrics* **72**, 1026–1036. https://doi.org/10.1111/biom.12522 (2016).
50. Seibold, H., Zeileis, A. & Hothorn, T. Individual treatment effect prediction for amyotrophic lateral sclerosis patients. *Stat. Methods Med. Res.* https://doi.org/10.1177/0962280217693034 (2017).
51. Athey, S. & Imbens, G. W. The state of applied econometrics: Causality and policy evaluation. *J. Econ. Perspect.* **31**, 3–32. https://doi.org/10.1257/jep.31.2.3 (2017).
52. Alaa, A. M. & van der Schaar, M. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *35th International Conference on Machine Learning, ICML 2018* (Stockholm, 2018).
53. Duan, N., Norman, D., Schmid, C., Sim, I. & Kravitz, R. L. Personalized data science and personalized (N-of-1) trials: Promising paradigms for individualized health care. *Harv. Data Sci. Rev.* https://doi.org/10.1162/99608f92.8439a336 (2022).
54. Ruggeri, K. *et al.* A synthesis of evidence for policy from behavioural science during COVID-19. *Nature (London)* **625**, 134–147. https://doi.org/10.1038/s41586-023-06840-9 (2024).
55. Shmueli, G. To explain or to predict?. *Stat. Sci.* **25**, 289–310. https://doi.org/10.1214/10-STS330 (2010).
56. Athey, S. & Imbens, G. W. Machine learning methods that economists should know about. *Ann. Rev. Econ.* **11**, 685–725. https://doi.org/10.1146/annurev-economics-080217-053433 (2019).
57. Xiao, Z. & Higgins, S. The power of noise and the art of prediction. *Int. J. Educ. Res.* **87**, 36–46. https://doi.org/10.1016/j.ijer.2017.10.006 (2018).
58. Foster, J. C., Taylor, J. M. & Ruberg, S. J. Subgroup identification from randomized clinical trial data. *Stat. Med.* **30**, 2867–2880. https://doi.org/10.1002/sim.4322 (2011).
59. Wright, M. N. & Ziegler, A. ranger: A Fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* https://doi.org/10.18637/jss.v077.i01 (2017).
60. Binois, M., Gramacy, R. B. & Ludkovski, M. Practical heteroscedastic Gaussian process modeling for large simulation experiments. *J. Comput. Graph. Stat.* **27**, 808–821. https://doi.org/10.1080/10618600.2018.1458625 (2018).

Acknowledgements

Part of this research was funded by a grant to Durham University from the EEF. We acknowledge the vision of the EEF in creating a data archive of their educational trials, which will enable further exploration and development; and we thank the staff at FFT Education Ltd for their data curation of this archive. The first author is grateful to the College of Social Sciences and International Studies and, the Institute for Data Science and Artificial Intelligence at the University of Exeter, for a strategic discretionary research grant and a University of Exeter-Alan Turing Institute Research Grant, without which, this collaborative project would not have been able to achieve where it is in terms of quality and reach. He is also thankful to Ben Neild and Lindsey Anderson of Exeter University for a HEFCE Catalyst Grant, which enabled him to teach causal inference covered in this paper to a group of teenagers who were considering data science as a profession or degree for higher education. This research has been presented at the Royal Statistical Society (RSS) in London, RSS Southwest Local Group in Plymouth, the Department of Computer Science at the University of Bath, the Q-Step Centre and Health Statistics Group at the University of Exeter. We appreciate the support and/or feedback on earlier drafts and/or talks we received from Alaidde Villanueva, Katherin Barg, Hugues Lortie-Forgues, Yinghui Wei, John Dennis, Obi Ukoumunne, William Henley, Justin Dillon, Lee Elliot Major, David Hall, Vivienne Baumfield, Oliver James, Susan Banducci, Andrei Zhirnov, Mark Kelson, Richard Everson, Jingyuan Mo, and Thomas King. Wherever possible, we have incorporated those thoughtful comments.

Author contributions

Z.M.X. and O.H. wrote the main manuscript text, Z.M.X and C.K. analysed the data, D.Z.L. refined the mathematical underpinning of the models, T.F. and S.H. provided resources and guidance. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-73714-z.

Correspondence and requests for materials should be addressed to Z.X.