Efficient targeting in childhood interventions

Alexander Paul² Dorthe Bleses³ Michael Rosholm⁴

Abstract

Using Danish administrative data, we address the issue of efficient targeting in childhood interventions. We define children to be in need of an intervention if they experience one or more socially undesirable outcomes in adulthood. Because interventions are very effective early in life, we then test if and to what extent indicators available at birth can predict these outcomes. We find fair to good levels of prediction accuracy for many outcomes, driven by a parsimonious set of predictors. We show that optimal weights for the construction of risk scores deviate from the weights typically used in targeted interventions.

1 We thank seminar participants at Aarhus University and three anonymous referees for useful comments and suggestions. This paper uses confidential administrative data from Denmark. The data can be obtained by filing a request directly to Statistics Denmark’s Division of Research Services (homepage: www.dst.dk/en/TilSalg/Forskningsservice). The authors are willing to assist in the process (Michael Rosholm, rom@econ.au.dk). An Online Appendix accompanies this paper. JEL Codes: I18, I28, I38.

2 Alexander Paul is a manager at E.CA Economics, Berlin, Germany, and an associate of the TrygFonden’s Centre for Child Research, Denmark.

3 Dorthe Bleses is a professor at the School of Communication and Culture, Aarhus University, and TrygFonden’s Centre for Child Research, Aarhus, Denmark.

4 Michael Rosholm is a professor of economics at the Department of Economics and Business Economics, Aarhus University, and an affiliate of TrygFonden’s Centre for Child Research, Aarhus, Denmark, and IZA, Bonn, Germany. Email: rom@econ.au.dk (corresponding author)

doi:10.3368/jhr.0320-10756R4

This open access article is distributed under the terms of the CC-BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0) and is freely available online at: http://jhr.uwpress.org
I Introduction

Many childhood interventions such as the famous Perry Preschool Project (Weikart 1967) target a selected subset of children rather than applying the intervention universally to the whole cohort. One rationale behind such targeting is the budget constraint. Even more important is that only certain children promise a sufficiently large return to justify the cost of the intervention. Children who already enjoy a beneficial environment in the absence of the intervention would benefit little or might even be harmed. Although the theoretical motivation for targeting interventions is evident, an important question in practice is which children should be targeted. Scholars have thus called for the development of “measures of risky family environments … that facilitate efficient targeting” (Heckman 2008, p. 314).

In this article, we approach the problem of efficient targeting from a long-term perspective. We define children as disadvantaged (in need of an intervention) if they are at risk of experiencing undesirable outcomes in their adult life, such as criminal behavior or high health care use. We exploit rich register data from Denmark to predict disadvantage, using only data available at the child’s birth. This constraint is motivated by recent studies in human development showing that childhood intervention programs are highly effective if administered very early in life (Heckman 2006; Cunha et al. 2010; Allen 2011). According to this literature, and with perfect information, children should be targeted immediately after birth, as in the prominent Carolina Abecedarian Project (Ramey et al. 1974). Children at high risk for adverse outcomes are likely to benefit greatly from early intervention.

Some prominent childhood interventions target children by means of a risk score. Only children that score sufficiently high are eligible for participation. A regular ingredient in defining the risk score is a measure of the family’s socioeconomic status (SES), such as household income or parental education. When combining all the indicators into a one-dimensional score, interventions tend to place a larger weight on family SES, but they generally do not motivate the exact weight given to
each component. As an example, the Carolina Abecedarian Project (Ramey et al. 1974, p. 65) constructed a “high risk index” that increased by one unit for each year missing from 12 years of parental schooling (separately for both the father and mother). It also increased by 4 units if family income fell short of 5,000 dollars and by one additional unit for each further 1,000-dollar reduction. Ramey et al. (1974) note that the weights they assigned to each component were based on a "best guess" (p. 10-11).

The unclear motivation behind the chosen weights suggests that risk scores might be substantially improved by answering the following basic questions: First, what relative weight should each indicator optimally receive? Does parental schooling matter more than income or vice versa? Second, do paternal and maternal characteristics matter in the same way, e.g. with respect to years of schooling? Third, is the relationship of the outcomes with the predictors non-linear and are there interactions among predictors? E.g., is each additional year of parental schooling equally important?

This study takes an econometric approach to address the problem of optimally selecting and weighting early indicators of disadvantage. We start by performing standard logit regressions to predict long-term outcomes using predictors measured at birth. Logit regression has the advantage that it allows for easy computation and interpretation of risk scores. However, it is not necessarily best at prediction, as interactions among predictors and functional forms must be explicitly specified and there are no safeguards against overfitting. We therefore also apply more sophisticated machine learning techniques that are known for good predictive power, in part because they implicitly allow for interaction terms among predictors and flexible functional form and prevent overfitting. We include indicators of family SES (income, education and occupation) and several other parental variables such as hours of work, health status and criminal activity as predictors. Some of these variables potentially correlate with quality time investments, which play an important role in human capital formation (e.g., Del Boca et al. 2014). At the child level, we include sex, birth order, county of residence at birth, and region of origin. We examine whether and which of the child
and parental variables can predict adverse outcomes in adulthood and derive optimal weights for the formation of composite risk scores.

The outcomes we consider are meant to capture the economic cost of different social dimensions, ranging from education and labor market outcomes to health and crime. The cost associated with these outcomes is not spread evenly across all members of society but can vary substantially. Indeed, it has been shown that a relatively small fraction of the population accounts for a sizable share of the total economic burden (Caspi et al. 2017; Richmond-Rakerd et al. 2020). We rank children by the outcome-specific cost they generate in adulthood and define the top 20% of the distribution as “at-risk” of the particular outcome. In the case of social benefits, for instance, the top 20% recipients in our sample account for 76% of total benefit receipt. We aim to predict which children are at risk and can potentially be targeted by an intervention. We also predict which children risk experiencing combinations of these outcomes.

We find that, first, predictions using register data available at birth are possible and often yield fair to good prediction accuracy. Predictions are most accurate for educational attainment, criminal behavior, placement in foster care and combinations of these outcomes, but are less accurate for health-related outcomes.

Second, we find that logit regression performs well. Predictions generated by other machine learning methods are generally neither statistically nor economically significantly different from logit regression. The reason for this failure to outperform logit could be data limitations. Even though our dataset includes the full Danish birth cohorts and many predictors, it is possible that even richer data would allow more sophisticated models to play to their strengths.

Third, we find that updating the predictors with data from a few years after birth improves predictions very little. It seems as if further improvements are only possible with indicators of the child’s behavior and skills, which are much more costly to obtain than the variables generally available from register data.
Fourth, we find that indicators of parental SES are highly predictive. A parsimonious set of indicators consisting of sex, parental education and income yields predictions that are almost as accurate as those obtained using the full set of predictors. Knowledge of a child’s sex and a few variables related to socio-economic background may therefore be sufficient for effectively targeting children in childhood interventions. Our study thus provides support for the practice of including measures of parental SES as a key ingredient in the construction of risk scores.

Finally, we derive optimal weights for the formation of risk scores and find that they deviate in important ways from the weights typically used in childhood interventions. In the discussion, we also point out that, due to treatment effect heterogeneity, risk scores are best thought of as a tool that assists human experts in their decision on final treatment assignment, and we discuss the practical relevance of the risk scores, including their sensitivity to policy interventions.

A study closely related to ours is Caspi et al. (2017). The authors find that a small set of predictors consisting of SES, maltreatment indicator, IQ and self-control could accurately predict adverse outcomes for 1,037 New Zealanders at age 38. Predicting whether children experience combinations of multiple adverse outcomes works particularly well. They use predictors that are recorded up until age 11, which is too late for effective early interventions. In addition, obtaining measures of IQ or self-control for the whole population is relatively costly. In contrast, we only use indicators that are inexpensive to measure, available from the Danish registers, and available at birth. Chittleborough et al. (2016) also use only predictors from around birth. However, they study outcomes at age 5 (before schooling starts), thus missing substantial information on social burden that only a long-term perspective can offer. To predict high-school dropout in Denmark, S, ara et al. (2015) employ Danish register data but use predictors available after high school entry.

This paper also relates to other strands of the literature. First, we use machine-learning techniques to predict which children are most in need of help. A growing number of studies address similar “prediction policy problems” (Kleinberg et al. 2015) in various contexts, e.g. regional allocation of refugees (Bansak et al. 2018), shootings among at-risk youth (Chandler et al. 2011), food-safety
inspections (Glaeser et al. 2016), hip and knee replacements (Kleinberg et al. 2015) and judicial bail-or-release decisions (Kleinberg et al. 2017). Second, it loosely relates to the theoretical literature on optimal treatment assignment (Bhattacharya and Dupas 2012; Kitagawa and Tetenov 2018; Manski 2004). This literature typically uses experiments or observational studies to estimate covariate-specific heterogeneous treatment effects. We do not observe treatment effects associated with a particular intervention. Instead, we suggest that treatment should be assigned by human agents to children who are at risk of an adverse outcome and who thus have the potential to benefit from an appropriately designed intervention.

Finally, our paper adds to the discussion of targeted versus universal programs. Targeted programs are a response to limited resources, which is particularly important in the context of the Scandinavian welfare state that continually struggles with the Baumol cost disease: Since increases in productivity are lower in the public than in the private sector (in part due to its larger share of labor in production), but wages in the public and private sector increase at the same rate (due to, e.g., institutional arrangements and unions’ bargaining power), the welfare state will eventually be faced with the problem that the level of services offered cannot be sustained indefinitely. Targeting may provide a (temporary) solution to this problem. Moreover, targeting may avoid potentially negative effects on subgroups of the population (e.g., Havnes and Mogstad 2015; Cornelissen et al. 2018). At the same time, targeted programs might lead to stigmatization and are less effective when disadvantaged children are hard to identify.

The paper is structured as follows. Section II presents some organizational and theoretical considerations that motivate our analysis. Section III deals with the practical aspects of prediction, including the data, estimation, and the specific prediction methods. Section IV reports the results. Section V discusses our findings and concludes.
II Organizational and theoretical considerations

The organizational setup we have in mind is one where econometrically estimated weights are used to produce a risk assessment, identifying children who are at risk of experiencing adverse outcomes later in life. This risk assessment can subsequently be used by human experts to determine which at-risk children/families to offer assistance. Determining the appropriate type of assistance comes in a subsequent step and is at present exclusively in the hands of human experts. Such a setup is similar to practices in the criminal justice system, where justices combine data-driven risk assessments with their own expertise to make the final decision on pre-trial release, sentencing or parole (see Berk 2019; Ludwig and Mullainathan 2021, for overviews).

In Section A of the Online Appendix, we develop a simple model that demonstrates how prediction can help the policy-maker to assign treatment in a welfare-improving way. We summarize the key aspects of the model here. We assume that treatment has a positive and homogeneous effect on at-risk children and does not harm children that are falsely identified as at-risk. The issue of homogeneous treatment effects is further discussed in Section V. Given a certain fraction of the cohort to be targeted by the intervention, the policy-maker should maximize the number of correctly identified at-risk children out of all at-risk children (the true positive rate, TPR), as previously shown by Sansone (2019) using a similar model. This result is intuitive: As we assume that the intervention does no harm, we need not be concerned about false positives. Next, the policy-maker chooses what fraction of the cohort should receive the intervention. If prediction is of value, the marginal TPR, and thus also the marginal expected benefit from the intervention, will decline as more children are targeted. Improving upon random and uninformed treatment assignment, prediction enables the policy-maker to optimally administer the intervention to a selected fraction of the cohort for which the marginal expected benefit exceeds the marginal cost.
III Data and methods

III.A Data

III.A.1 Sample
Our sample consists of the full Danish birth cohorts from the years 1985, 1986 and 1987. This choice is motivated by the fact that most of the Danish register data is available from 1980 onwards, so a certain period after this year is required to construct powerful predictors, such as parental crime or parental hospital admissions. At the same time, children should not be born too recently, such that we can also observe their relevant adult outcomes, such as educational attainment or disposable income. Because we use parental predictors, we also impose the condition that both parents are known to us and have lived in Denmark continuously since 1980. The final sample contains 149,755 children.

III.A.2 Outcomes
The age at which we measure the children’s outcomes ranges from 28 to 33 years, depending on the availability of most recent data. We focus on seven outcomes that capture the economic cost of different societal dimensions, ranging from education and labor market outcomes to health and crime. Reflecting the principle that a small fraction of the population accounts for a disproportionate share of the total economic burden associated with a certain outcome (Caspi et al. 2017; Richmond-Rakerd et al. 2020), we order children by the outcome-specific cost they generate in adulthood and define the top 20% of the distribution as “at-risk” of this outcome. While other values are also possible, we choose 20% because it was previously used by Caspi et al. (2017), who motivated this choice with the Pareto principle, named for economist Vilfredo Pareto, according to which 80% of consequences arise from 20% of causes (Bunkley 2008). If we instead used continuous measures as outcomes, our prediction methods would focus more on distinguishing between children within the low-burden group and thus become less accurate at identifying the high-burden children that we are most interested in.
Table 1 provides an overview of the eight outcomes considered in this study (see also Table G.1 in the Online Appendix for data sources and detailed descriptions). For five of the outcomes, those having the outcome account for 100% of the total burden to society. For example, children ever being criminally charged account for all of the societal cost associated with criminal charges, whereas children never charged do not cause any cost. For the three other outcomes, those having the outcome only account for part of the total burden to society, but their share is disproportionate.

Specifically, Table 1 shows that the top 20% social benefit recipients in our sample account for 76% of total benefit receipt. Similarly, the top 20% patients with the most hospital admissions account for 55% of all admissions. For income, which is a benefit rather than a burden to society, the pattern is reversed: the bottom 20% of the income distribution receive a disproportionately small share equal to 9% of total income.v

Table 2 shows that the various outcomes are not independently distributed across children but instead highly correlated. A consequence is that a disproportionately large share of the cohort will have combinations of several outcomes. By the same token, disproportionately many children will be entirely free of any adverse outcome. This is illustrated in Figure F.1 in the Online Appendix, which juxtaposes the actual distribution of the number of negative outcomes with a simulated one based on the assumption that outcomes were uncorrelated.

Figure F.2 in the Online Appendix reveals that children with combinations of outcomes not only account for a disproportionate share of the population but also for a disproportionate share of the total economic burden. In fact, this share becomes increasingly disproportionate as the number of adverse outcomes rises. Therefore, in addition to predicting each outcome individually, we also aim to predict if children have several adverse outcomes. Specifically, we search for a predictive algorithm that can distinguish children with 3+ outcomes versus those with 2 or fewer (and similarly for 4+ and 5+ outcomes). We also predict having 3+ outcomes as opposed to zero outcomes. This exercise implicitly assumes that we know a priori that a child will end up with either 0 or 3+ outcomes.vi
Finally, in addition to counting the number of adverse outcomes, we also construct a social burden (SB) indicator through confirmatory factor analysis assuming that a single factor underlies the seven individual outcomes. See Table G.2 in the Online Appendix for additional information.

**III.A.3 Predictors**
The predictors included at the level of the child are sex, county of residence at birth, region of origin, birth month and birth order. Predictors at the level of the parents are recorded separately for the mother and the father. They include: income, wealth, educational attainment, occupation, working hours, age, marital status, hospitalizations, placements, and criminal charges. All continuous predictors are turned into discrete categorical variables. Missing values are generally assigned to a separate category. We include a dummy for each category (except for the logit model, where we omit the baseline category) in the prediction analysis, resulting in a total of 184 dummy variables (159 in the logit model). See Tables G.3 and G.4 in the Online Appendix for more details and summary statistics. All of the child and parental predictors are well-known to be associated with outcomes in adulthood, as discussed in Section B of the Online Appendix.

In the main analysis, all predictors are measured at or before birth. In additional analyses, we extend the time frame to include the first few years after birth. This allows us to update and strengthen the parental predictors with more recent data and to include the child’s hospitalizations as an additional predictor.

**III.B Methods**

**III.B.1 True positive rate (TPR) and expected benefit**
A fully fledged benefit-cost analysis as sketched in Section II and described in detail in Section A of the Online Appendix would require information that depends on the context and is often not available, such as the monetary cost of the outcome as well as the impact of the intervention. For this reason, we do not attempt to find the optimal targeted fraction of the cohort that should receive the intervention. Instead, we examine whether the marginal expected benefit curve declines with the fraction treated. To do so, we examine the slope of the marginal TPR curve, which in turn
determines the slope of marginal expected benefit curve. If the marginal TPR curve - and thus the marginal expected benefit curve - is sloping downward, this would be evidence that data available at birth can be gainfully utilized to improve upon random and uninformed treatment assignment and thus increase social welfare.

To determine the shape of the marginal TPR curve, we proceed as follows. First, for a given targeted fraction, we maximize the true positive rate (TPR). The optimal mechanism assigns the intervention to those children with the highest probability of being at risk. We estimate this probability using machine learning techniques and categorize children with the highest estimated probability of experiencing the outcome as at risk, while collectively adding up to the chosen targeted fraction. We then compute the resulting TPR. Second, after performing this step for a grid of 20 targeted fractions (5%, 10%, …, 100%), we study how the marginal TPR changes with the targeted fraction and is sloping downward.

We also report the so-called area under the curve (AUC) as a summary measure of predictive accuracy across all possible values of targeted fractions. The AUC ranges from 0.5 to 1; in practice, AUC values are often interpreted as follows (Caspi et al. 2017): worthless (0.5–0.6), poor (0.6–0.7), fair (0.7–0.8), good (0.8–0.9), excellent (0.9–1.0). Appendix C.1 provides details on the derivation of the AUC in our setting.

III.B.2 Estimation and prediction
Predicting the realization of binary outcome variables, as in this study, is known as a classification problem. We begin by estimating standard logit regressions. An advantage of logit regression is that predictor variables are assumed to combine linearly to form the risk score for each child. In combination with the exclusive use of dummy variables, the risk score thus becomes easy to construct and to interpret. A disadvantage of logit regression is that interactions among predictors must be explicitly specified. Another disadvantage is its implicit risk of overfitting and poor out-of-sample performance.
To address the issue of overfitting, we split the dataset into an 80% training dataset, on which we perform estimations, and a 20% test data set, on which we evaluate the model fit in terms of TPR and AUC. In an attempt to improve the predicted probabilities from the logit, we additionally employ three modern machine learning methods: (i) logit LASSO, (ii) random forest, and (iii) gradient boosting. See Section D.1 in the Online Appendix for details on these methods. The models have in common that they allow for different levels of model complexity that are governed by a vector of tuning parameters. A more complex model reduces the bias in representing the relationship between outcome and predictors, but comes at the risk of overfitting. Both overfitting and bias worsen out-of-sample model performance.

To find the optimal level of complexity, we tune parameters via 8-fold cross-validation on the training data (following the recommendations by Mullainathan and Spiess 2017). That is, we split the training data set into eight equally sized folds and set one of the folds aside. We then estimate the model for specified values of the tuning parameters on the remaining seven folds and evaluate the fit on the selected fold. After repeating this step for each of the other seven folds, we compute the average fit across all folds. The optimal parameter specification is the one that yields the highest average fit in the cross-validation procedure. We use this specification to re-estimate the model on the whole training data and evaluate its fit on the test data. Section D.2 of the Online Appendix reports the optimal tuning parameters.

IV Results

IV.A Individual outcomes

Based on predicted probabilities from logit regressions, panel A of Figure 1 shows the marginal increase in the true positive rate (TPR) for the education outcome as we increase the targeted fraction of the cohort to be treated. Similar graphs for all outcomes considered are shown in Figure F.3.
The graph provides clear evidence that informed treatment assignment based on predictions can substantially improve upon uninformed treatment assignment. For example, targeting 5% of the cohort will reach more than 15% of all children at risk of having only compulsory education compared to 5% under random, uninformed assignment. Thus, when the problem is to target a given fraction of the cohort, for example due to budget constraints, informed treatment assignment can undoubtedly increase welfare.

In panel B of Figure 1, we show the ROC plot along with corresponding AUC value to assess the overall accuracy of the education prediction. ROC plots for all outcomes are available in Figure F.4. Prediction works well for education, and also for criminal charges and foster care placement, with AUC values between 0.76 and 0.81. Health outcomes and income are less predictable by early-life indicators. These findings are consistent with earlier evidence that the child’s educational outcomes depends strongly on parental education, while intergenerational mobility is larger for income and health. As Landersø and Heckman (2017) pointed out, the Danish welfare state is characterized by large income redistribution through taxes and transfers in addition to wage compression, leading to higher intergenerational income mobility than educational mobility. Similarly, universal access to tax-financed medical care might explain why family background matters less for health outcomes.

**IV.B Combinations of outcomes**

Section III.A.2 demonstrated that outcomes are highly correlated within children. A small group of children accounts for a disproportionately large share of the total social burden. Predicting who these children are and targeting them for appropriately designed interventions promises large returns; Caspi et al. (2017) – using predictors throughout childhood – showed that prediction works even better for these “high-cost” children.

In the graph in the left panel in Figure 2, we predict whether children have 3+ outcomes versus 2 or fewer (and similarly for 4+ and 5+ outcomes). AUC values are high (0.75-0.81), suggesting that targeting high-cost children is indeed easier, even when using only predictors measured at birth. In the right panel, we repeat the analysis for whether children have 3+ outcomes versus 0 outcomes
(and similarly for 4+ and 5+ outcomes). AUC values are now even higher (0.80-0.87), implying that there is substantial variation between low-cost and high-cost children already at birth. Our estimates are similar to the ones by Caspi et al. (2017). Unfortunately, however, we do not know a priori which children will end up with either zero or 3+ outcomes (but not 1 or 2 outcomes), so this latter exercise is of limited practical relevance.

**IV.C Comparing methods**

Table G.7 in the Online Appendix compares AUC estimates from different methods and outcomes. For each method and outcome, we present two estimates of the AUC as explained in Section C.1 of the Online Appendix. The main result is that the directly optimized AUC, which we generally report in this study, yields similar results as the indirectly fine-tuned AUC (see Section C.2 of the Online Appendix a more detailed discussion).

Table G.7 also allows us to compare logit regression with machine learning techniques. The main result is that these more advanced machine learning techniques do not improve predictions in our case.

The quantitatively similar performance of the logit regression and the machine learning techniques may come as a surprise. After all, random forest and gradient boosting are tree-based approaches that fundamentally differ from regression-based approaches such as logit and LASSO logit (see Section D.1 of the Online Appendix). Moreover, random forest and gradient boosting offer enhanced flexibility with respect to interactions and functional form, which gives reason to expect a better predictive fit.

We take several steps to make sure the similar performance of the various methods is not caused by inadequate calibration. First, we compare the out-of-sample fit on the test data, which we have reported so far, with the in-sample fit on the training data. As we perform 8-fold cross-validation on the training sample, we compute the in-sample fit as the average across eight different fits, each belonging to the fold that is left out of the calibration and used for validation. Out-of-sample and in-
sample fit are very similar, alleviating concerns about over-fitting (Table G.8 in the Online Appendix). Second, we check whether the optimal tuning parameters of the machine learning methods are not at the edges of the tested parameter range, which would point to sub-optimal calibration. In Section D.2 of the Online Appendix, we report optimized tuning parameters for all combinations of outcomes and metrics, corresponding to more than 200 predictions. Most optimal parameters are indeed in the interior of the allowed range. Some are not, which is however not surprising in light of the large number of predictions made. Improper parameter tuning is hence unlikely to drive our findings.

To conclude, we believe that data limitations are the most plausible reason why machine learning methods in the present case do not outperform logit. Machine learning excels at prediction in big data. Our data are limited. First, the number of observations is dictated by the magnitude of the Danish birth cohorts. Although a sample size of 149,755 is quite large, one cannot rule out that the machine learning algorithms would outperform the logit regressions if applied to much larger datasets with millions of children. Second, the number of predictors (184 dummy variables, see Section III.A.3) is also limited. This is in part on purpose, as we aim to include predictors available at or before birth and attainable at low cost. However, as shown in Section B of the Online Appendix, all predictors have been previously shown to correlate with the outcomes and thus also with each other. Consequently, the algorithms do not have the chance to detect subtle patterns in the data as there are no new variables to explore. See McKenzie and Sansone (2019) for a previous paper in which machine learning methods are not superior likely because of data limitations.

**IV.D Post-birth predictors**

A trade-off exists between the effectiveness of the intervention, which some studies have shown to decline with age (Heckman 2006; Rosholm et al. 2021), and the accuracy of predictive targeting, which likely improves with more information collected later in life. The question is whether we can improve predictions by extending the time frame to a few years after birth. Adding years after birth
allows us to update parental predictors and to include the child’s hospitalizations as an additional predictor. We look at 1, 3 and 5 years after birth.

Figure 3 illustrates how the AUC changes as we extend the time window. As expected, the AUC increases with more recent predictors. The increase is most pronounced for foster care placement, which is also the outcome that occurs earliest in life and is probably more sensitive to changes in early-life family environment. Overall, however, the marginal improvements in predictive accuracy are rather modest. It seems as if the role that family background plays in shaping long-term outcomes is largely determined by factors set in place at birth. Targeting children already at birth appears to come at little cost of predictive accuracy relative to the large potential benefits of early versus late interventions. We discuss this finding further in Section V.

**IV.E Parsimonious model**

Our predictions are based on a rich set of variables. If a parsimonious model with a small number of predictors can generate predictions close to the full model, then the necessary data collection would be less costly and computing risk scores would be simpler and more transparent. We therefore examine whether all predictors contribute equally to the quality of the predictions or only few predictors drive our results.

As shown in detail in Section E of the Online Appendix, any two of the three SES predictors parental income, education and occupation (plus the child’s sex) generates around 90% of the full-model AUC. The best combination is that of income and education, suggesting that these two variables complement each other most, while occupation captures aspects of both and thus contributes the least independent variation. Adding occupation to both income and education does not yield great additional improvement.\(^{[\natural]}\)

The finding that a parsimonious model with the predictors sex, parental income and parental education performs almost as well as the full model suggests that collecting data on these few variables is sufficient for efficient targeting in practice. It also bolsters the way that targeting has
been operationalized in many prominent childhood interventions that used indicators of parental SES to define disadvantage. A separate but equally important question is how much weight to attach to the different values that an indicator can take on. We address this question next.

**IV.F Optimal weights**

Efficient targeting based on a one-dimensional risk score requires optimally weighing the selected predictors that contribute to the risk score. The Perry Preschool Project derived a “cultural deprivation rating” in which higher values indicated better outcomes (Weikart 1967, p. 3-4). It had three components: Paternal occupation entered the rating with 1 point if the father engaged in unskilled work and 4 points if he engaged in skilled work. Each of the parents’ average years of schooling entered with another point. Finally, “density in the home”, measured as number of rooms divided number of people, entered after multiplying by one half to give it “a 1/2 weight”. Before aggregating to the final rating, each component was additionally divided by its standard deviation to equate different distributions. Only children with a final rating below 11 were considered further for the experiment.

The Carolina Abecedarian Project (Ramey et al. 1974, p. 65) constructed a “high risk index”. The index increased by one point for each year missing from 12 years of schooling, for both the mother and father. Family income above 5,000 dollars left the risk score unchanged, while income below 5,000 dollars increased it by 4 points, and by another point for each additional 1,000-dollar step downwards. Additional points were assigned for, among others, the absence of the father and low parental I.Q. scores. Only children with an index value of 11 or higher participated in the experiment. Ramey et al. (1974) note that these “[w]eights were assigned to the various factors based upon our ‘best guess’ of their relative importance.” (p. 10-11).

In interventions such as the Perry Preschool and the Carolina Abecedarian Projects, the assigned weights often lack a clear motivation. Thus, at least three questions arise: First, what relative weight should each indicator optimally receive? Second, should indicators such as education enter differently for fathers and for mothers? Third, should indicators such as years of schooling enter
linearly? We do not have access to all the indicators used in these two or other studies. Instead, we will use a parsimonious set of predictors consisting of sex, income and education, which we showed to predict quite well in Section IV.E. Income and education are key ingredients in defining early disadvantage in many targeted interventions. Answering the above questions for these variables is thus highly policy-relevant.

We can address the problem of optimal weighting with the estimated coefficients from the logit regressions. Because our predictors are discretized and a dummy is included for each discrete value, the coefficient directly gives the weight associated with that value of the predictor, relative to the baseline value. For easier interpretation, we rescale all weights such that the lowest possible weight is 0 and the highest possible weight is 100. Note that risk score values are not interpretable as percentiles; their distribution depends on the distribution of predictor values in the population.

Figure 4 illustrates the computation of the weights for the educational attainment outcome. Other outcomes and the exact values of the weights are shown in Figure F.6 and Table G.9 in the Online Appendix, respectively. The baseline child – the leftmost circles in each column with a gray edge – relative to which the weights are defined is characterized as: female, master’s degree/PhD (both mother and father) and 10th income decile (both mother and father). We see that the baseline child has a score of 0 (=0+0+0+0+0). If the child’s father had only compulsory schooling instead of a master’s degree/PhD, her score would increase to 28 (=0+28+0+0+0+0). Even though the coefficients are estimated with some uncertainty, as indicated by the 95% confidence bands, it is reassuring to see that in general income and education show the expected positively monotone relationship with the outcome.

What lessons can we draw about optimal weights in general? First, men tend to have higher risk of adverse outcomes except for health-related outcomes. Income and education seem in general to contribute equally to the risk scores, when measuring contribution as the difference between the lowest and the highest score within a predictor. That said, education contributes more to the risk score than income when predicting education and vice versa.
Second, maternal and paternal education affect the risk scores in a similar manner for most outcomes. An interesting observation with respect to parental income is that maternal income plays a small role as long as she is in the upper 80% of the distribution, but being in the bottom 20% causes the risk score to spike. The relationship is in general more linear for paternal income.

Third, especially values at the bottom of the education and income distribution substantially raise the risk score. For income, these are the bottom 30% of fathers and the bottom 20% of mothers. For education, this corresponds to the parent having only compulsory schooling or a vocational education. This finding implies that risk scores in which income or years of schooling enter linearly, like in the Perry Preschool Project and the Carolina Abecedarian Project, assign too little weight to children with parents in the bottom of the distribution. Of course, this might be driven by our specific choice of outcomes that also focus on the bottom of the respective distribution. Predictions of, for instance, average hospitalizations rather than top 20% hospitalizations might depend much more linearly on parental SES. However, as we argued above, it is precisely those children in the top 20% of the distribution that cause a disproportionate burden to the welfare state. Targeting these children effectively seems appropriate from an economic point of view.

V Discussion and conclusion

We use prediction methods to show that efficient targeting in childhood interventions is possible even if only variables available at birth are at the decision-maker’s disposal. This applies to interventions addressing a wide range of long-term outcomes, ranging from labor market outcomes to health and crime. We also find that predictions do not improve much by adding post-birth indicators. We demonstrate that a parsimonious set of variables consisting of sex, parental education and parental income predicts almost as well as the full set of predictors. Finally, we provide econometrically derived optimal weights for the formation of risk scores that differ from the weights typically used in the literature.
In our simple motivational model, where treatment effects were assumed to be homogeneous among at-risk children, risk scores could be directly used for assigning children to interventions: Children with higher risk scores should be selected for the intervention. In reality, however, treatment effects are likely heterogeneous and not necessarily correlated with risk (Athey 2017; Kleinberg et al. 2015; Ascarza 2018). This has implications for treatment assignment, as targeting high-risk children yields little value if they are not actually the ones benefiting from an intervention. For example, patients with a family history of severe genetically determined chronic disease might not respond to any type of treatment.

Given the reality of heterogeneous treatment effects, data-driven risk scores should not be used alone to assign treatment. In our view, risk scores better serve as a tool that aids human experts in choosing which children should be targeted for some intervention: Given a pool of children identified as at risk by means of algorithms, experts can employ their expertise to select the children deemed most suited for intervention. With several interventions available, experts can assign each child the most appropriate type of treatment. Of course, for some types of risks, e.g. crime and drug abuse where social interactions may be important determinants, intervention may more fruitfully be targeted to all children within, say, residential areas rather than to children.

The idea of combining algorithms with human judgment is, for example, used in criminal justice, where risk scores inform justices who then make the final decision on e.g. bail, sentencing or parole (Stevenson and Doleac 2021; Ludwig and Mullainathan 2021; Berk 2019). See also Gennatas et al. (2020) for an interesting discussion of how to integrate human expert assessments with machine learning algorithms.

Risk scores may also be useful in randomized controlled trials (RCTs) that test the effectiveness of a novel intervention using treatment and control groups. Administering an RCT to a selected group of children identified as at risk – rather than the full cohort – saves resources while targeting those children that presumably benefit the most from the intervention. This was the motivation behind constructing the risk scores used in the Perry Preschool Project or the Carolina Abecedarian Project.
Our finding that predictions do not improve much by adding post-birth indicators is contingent on using only predictors typically available in registers. Post-birth measures of child development or parenting style are likely to have predictive power as well. For example, Dale et al. (2023) augment Danish register data with measures of the child’s expressive vocabulary and parental investment (daily singing/reading) at age 16 to 30 months and show that adding these variables improves predictions of school grades at age 15. However, these variables are typically not available in register data and costly to obtain at large scale, which poses a problem if the goal is to identify at-risk children in the larger population. Though beyond the scope of this paper, predictions from a richer small sample could be informative about those predictors in the register data that are particularly valuable or suggest a different weighting over them.

So far, we have been deliberately unspecific about the nature of the intervention that our risk scores can be used for. This is because an intervention can take many forms depending on the context. At the broadest level, the intervention could consist of the provision of free or subsidized public child care to parents. A related type of intervention are center-based programs such as the Perry Preschool Project or the Carolina Abecedarian Project that go beyond basic child care by closely involving parents or offering health care. Finally, in a more narrow sense, interventions can also be thought of as specific programs aimed at improving the learning environment within the daycare center (e.g., Bleses, Højen, Dale, et al. 2018; Bleses, Højen, Justice, et al. 2018) or at home (e.g., Andersen and Nielsen 2016; Hjort et al. 2017). See Rosholm et al. (2021) for an overview of recent RCTs conducted in the context of the Danish welfare state, covering various interventions throughout childhood.

The setting of our study is the generous Danish welfare state, which raises the question whether our findings can be generalized to countries with a less generous welfare state, such as the United States. In particular, the Danish welfare state features a high overall spending level and universal access to child and health care, which could impact the predictive patterns uncovered in this study. Heckman and Landersø shed light on this question in a series of recent papers (Landersø and
Heckman 2017; Heckman and Landersø 2022). They show that family background, measured as parental education, is approximately as good a predictor of various child outcomes in both Denmark and the U.S. This applies to outcomes as diverse as in our study: birth weight, admission to neonatal care unit, sociability, language test scores, criminal convictions, educational attainment, labor market status and mortality. The authors conclude that “despite generous social policies, family influence on many child outcomes in Denmark is comparable to that in the U.S. Common forces are at work in both countries that are not easily mitigated by welfare state policies.” (Heckman and Landersø 2022).

A notable exception to the similarity between Denmark and the U.S. is income, which is more loosely related to parental characteristics in Denmark than in the U.S (Landersø and Heckman 2017). Heckman and Landersø attribute this to ex-post redistribution through the Danish tax and transfer system, not ex-ante equalization of individual skill acquisition. For income, the role of parental background will likely be more pronounced in countries with a less generous welfare state than Denmark. Overall, however, the research suggests that the predictive patterns uncovered in this study carry over to other countries, in particular the U.S.

The question remains, of course, if risk scores should be applied in practice. There are a number of caveats. First, the Lucas critique obviously applies to our setting: The risk scores computed here are not policy-invariant. If, in reaction to a risk score, a predictor is successfully targeted by policy, then its predictive power will become attenuated; in the extreme, even though relevant, it might no longer show up as predictive at all. Similarly, the risk scores derived in this study could be affected by historical policies. For example, our finding that birth order has little predictive power could be due to interventions effectively targeting children based on this variable. Ignoring birth order in defining risk would then be a grave mistake. The sensitivity of machine learning algorithms to policy interventions that change the stability of the relation between predictors and outcomes has also been discussed in Athey (2017) and Hofman et al. (2021).
Second, as a related point, the predictors used in this study were recorded more than 30 years ago. To the extent that the relationship between at-birth predictors and long-term outcomes is different today, the risk scores we computed might no longer be optimal for cohorts born today. The justification for our approach is that long-term outcomes must be observed to be able to construct meaningful risk scores. The alternative to anchoring risk scores in long-term outcomes is to assign ad hoc weights, which most likely results in less efficient targeting. When training the model on one cohort and evaluating it on another cohort, we find no evidence for an unstable relation between predictors and outcomes in the short run (Figure F.8 in the Online Appendix).

Third, parents may try to manipulate the risk score in order to receive treatment for their child. In our setting, the potential for manipulation is small as data were directly taken from the official registers. Survey-based data are more likely to contain false records. Besides misreporting, parents could of course also directly lower their income to become eligible.

Finally, targeting raises ethical issues. Dare (2013) discusses several of these issues within the context of a predictive risk model for child maltreatment. The computation and publication of risk scores might be viewed as reducing the child to a number. Moreover, there is a large difference between a calculated risk score and a realized disadvantage. Intervening on the basis of a perceived risk raises questions concerning e.g. the rights of parents to raise their children as they see fit. Furthermore, if assignment of children identified as “high-risk” to an intervention is considered a stigma, then the positive effects of treatment may be counteracted by the negative effects of the stigma. Resistance to using data-based algorithms in social settings might also come from fears that algorithms are biased and tend to perpetuate existing inequalities (O’Neil 2016), even if the algorithm is actually designed to achieve the opposite. It is obvious that these concerns are very real and should be handled and/or taken into consideration. In the context of child maltreatment, Dare (2013) concludes that the potential gains outweigh any ethical reservations but in our context it is perhaps not so clear cut. See Kleinberg et al. (2018) and Holm (2019) for other discussions of the ethical aspects of predictive risk models. Discussing these issues more deeply is beyond the scope
of this paper, which aims to demonstrate the extent to which targeting is practically feasible. We thus hope our work will serve as a basis for further discussion, both inside and outside of academia.
References


__________. 2012. Invest in early childhood development: Reduce deficits, strengthen the economy. The Heckman Equation.


Şara, Nicolae-Bogdan, Rasmus Halland, Christian Igel, and Stephen Alstrup. 2015. “High-school dropout prediction using machine learning a Danish large-scale study.” In Proceedings. ESANN 2015: 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, edited by Michel Verleysen, 319–324.


Figure 1
Prediction of educational attainment outcome
Notes: Panel A: The solid line is based on predictions from logit regressions and shows incremental changes in the true positive rate as the targeted fraction of the cohort to be treated is increased in steps of 5%. The dashed line shows corresponding constant increase under uninformed treatment. The gray area represents 95% bootstrap confidence intervals based on 2,000 bootstrap samples from the test data, keeping the prediction function fixed. Panel B: Receiver operating curve (ROC) and area under the curve (AUC) for predictions from logit regression. The brackets contain the 95% confidence interval based on 2,000 bootstrap samples from the test data, keeping the prediction function fixed.
Figure 2  
Prediction – Combinations of outcomes  
Notes: Receiver operating curve (ROC) and area under the curve (AUC) for predictions from logit regression.
Figure 3
Prediction with post-birth predictors
Notes: Improvement in AUC when updating predictors 1, 3 and 5 years after birth and adding the child’s hospitalizations as another predictor. All results based on logit regressions.
Figure 4
Optimal risk scores for educational attainment outcome
Notes: This figure illustrates how to construct optimal risk scores for the outcome "Education (compulsory schooling only)". The baseline child is indicated by the leftmost circles with a gray edge and has the following characteristics: female, master’s degree/PhD (mother/father) and 10th income decile (mother/father). The score of the baseline child is the sum of the points in the leftmost circles. For children with other characteristics, the score is obtained by adding the points indicated to the right of the baseline child. The characteristics are given in the following order: master’s degree/PhD, bachelor/vocational bachelor, short cycle higher education, high school, vocational education and training, compulsory schooling (mother/father) and 10th decile, 9th decile, ..., 1st decile (mother/father). All figures are based on coefficients from logit regressions that have been rescaled such that the minimum possible score is 0 and the maximum possible score is 100. Vertical bars represent 95% confidence bands based on robust standard errors.
Table 1: Overview of outcomes

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Prevalence (%)</th>
<th>Fraction of total burden (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Education</td>
<td>12.7</td>
<td>100.0</td>
</tr>
<tr>
<td>Compulsory education only in 2018</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Income</td>
<td>20.0</td>
<td>9.7</td>
</tr>
<tr>
<td>Bottom 20% equivalised disposable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>family income 2016-2017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Social benefits</td>
<td>19.9</td>
<td>76.2</td>
</tr>
<tr>
<td>Top 20% social benefit recipients</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(week 14/2006 - week 01/2019)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Criminal charges</td>
<td>19.2</td>
<td>100.0</td>
</tr>
<tr>
<td>Ever criminally charged throughout</td>
<td></td>
<td></td>
</tr>
<tr>
<td>life (as of 2017)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Placement</td>
<td>10.1</td>
<td>100.0</td>
</tr>
<tr>
<td>Ever in placement or preventive measure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(as a child)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Hospitalizations</td>
<td>19.3</td>
<td>55.4</td>
</tr>
<tr>
<td>Top 20% hospital admissions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>throughout last 10 years (2006-2016)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Psychiatric condition</td>
<td>16.4</td>
<td>100.0</td>
</tr>
<tr>
<td>Whether diagnosed with a psychiatric condition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(as of 2016/2017)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Social burden (SB) indicator</td>
<td>20.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Top 20% score of the single common factor of all 7 outcomes extracted through confirmatory factor analysis</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Note that income is a benefit rather than a burden to society. The table shows that the bottom 20% of the income distribution receive a disproportionately small share of total income.

Table 2: Correlations

<table>
<thead>
<tr>
<th></th>
<th>Education</th>
<th>Income</th>
<th>Social benefits</th>
<th>Criminal charges</th>
<th>Placement</th>
<th>Hospitalizations</th>
<th>Psychiatric condition</th>
<th>SB indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td>0.49</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social benefits</td>
<td>0.64</td>
<td>0.58</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Criminal charges</td>
<td>0.44</td>
<td>0.30</td>
<td>0.37</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Placement</td>
<td>0.57</td>
<td>0.40</td>
<td>0.59</td>
<td>0.41</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospitalizations</td>
<td>0.30</td>
<td>0.16</td>
<td>0.35</td>
<td>0.26</td>
<td>0.30</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychiatric condition</td>
<td>0.47</td>
<td>0.46</td>
<td>0.55</td>
<td>0.28</td>
<td>0.49</td>
<td>0.35</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>SB indicator</td>
<td>0.81</td>
<td>0.72</td>
<td>0.90</td>
<td>0.54</td>
<td>0.74</td>
<td>0.42</td>
<td>0.70</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Notes: Reported correlations are polychoric correlations that assume a normal distribution of the underlying latent variables.
Kline and Walters (2016) show that children participating in Head Start do not benefit from the program if they would have otherwise attended another preschool, but they benefit significantly if the alternative would have been being cared for at home. As another example, Cornelissen et al. (2018) find that children with immigrant ancestry benefit most from attending child care because their alternative care arrangements are of relatively low quality. Similarly, Havnes and Mogstad (2015) study the effect of child care attendance on earnings in adulthood and show that children from low-income families gain substantially, whereas children from upper-class families suffer a loss in earnings.

Heckman (2012) states: "The highest rate of return in early childhood development comes from investing as early as possible, from birth through age five, in disadvantaged families."

The Perry Preschool Project targeted African-American children with low IQ from families that performed poorly on a cultural deprivation scale based on paternal occupation, parental education and density in the home (persons per room) (Weikart 1967). The Carolina Abecedarian Project constructed a high risk index based on parental education and income and 10 additional minor indicators (Ramey et al. 1974). The Early Training Project considered housing, parental education and parental occupation (Klaus and Gray 1968). Eligibility for Head Start is mainly determined by parental income (U.S. Department of Health & Human Services 2021).

See also Gennatas et al. (2020) for an interesting discussion of how to integrate human expert assessments with machine learning algorithms.

A particularly relevant study in this context is by Richmond-Rakerd et al. (2020), who also use Danish administrative data. They show that adult health, crime and social welfare are unequally distributed across children and correlated within children. Our definition of outcomes varies slightly from theirs: we consider the number of hospitalizations rather than the number of hospital days and criminal charges rather than criminal convictions. Richmond-Rakerd et al. (2020) do not attempt to predict outcomes with information available at birth.

We include these results for the purpose of comparison with Caspi et al. (2017).

See also Varian (2014) for an introduction to machine learning from an economist’s perspective.

The coefficient estimates of the corresponding logit regressions can be found in Tables G.5 and G.6 in the Online Appendix.

Hertz et al. (2008) estimate that the intergenerational correlation for education in Denmark is 0.30. Andrade and Thomsen (2018) find values in the range between 0.35 and 0.39. In contrast, the intergenerational correlation for income is typically much smaller. For gross income including public transfers Landersø and Heckman (2017) estimate it to be 0.21, while Andersen (2021), studying total income before deductions and taxes, finds values between 0.05-0.06 (maternal income) and 0.13-0.21 (paternal income). Our measure of disposable family income reflects the progressivity of the Danish tax system, so these estimates are probably too large. Intergenerational correlations for health tend to be even smaller, at least with respect to fathers. Andersen (2021) finds values between 0.11-0.12 (paternal health) and 0.13-0.14 (maternal health).

Andersen (2021) finds rank-rank slopes for intergenerational health outcomes in Denmark that are only half the size of those found by Halliday et al. (2018) for the U.S., a country with a considerable fraction of uninsured people.

In certain contexts, particularly in the United States, only sex and income might be available, but not education. Targeting only on sex and income yields a substantially lower AUC than when either occupation or education are also included. In this case, however, extending the time frame leads to larger improvements in performance for a number of outcomes, suggesting that education, which is highly correlated with permanent income, can partially be substituted for by post-birth income (see Figure F.5 in the Online Appendix in analogy to Figure 3). However, the gap to the full model is far from fully closed.

In tree-based approaches such as random forest and gradient boosting, computing risk scores is not as straightforward as in logit. One way to quantify the importance of predictors is to measure the so-called gain, that is, the contribution of each predictor to the model based on the total gain of this predictor’s splits. However, note that these methods’ built-in randomness with respect to the predictors and observations used in each step of the algorithm tends to attribute importance somewhat evenly across predictors. That said, Figure F.7 in the Online Appendix confirms that SES-related predictors such as income and education are most important in our gradient boosting model. Table G.10 in the Online Appendix lists the top 5 most important individual predictors in gradient boosting; it shows that having a parent – in particular a mother – at the bottom of the education distribution (compulsory education only) or income distribution (1st decile) has the most important contribution to the prediction in the gradient boosting algorithm, similar to the risk scores based on the logit regression.

See Hellerstein and Imbens (1999) for an example of combining large and small datasets.

Consistent with this result, we also find relatively low predictive accuracy for the income outcome in this study.