

Predicting adult life outcomes from earlier signals: identifying those at risk

Report for the PMSU, version 2.2

Leon Feinstein and Ricardo Sabates

August 31, 2006



**Centre for Research on the Wider Benefits of Learning
Institute of Education, University of London**

Executive Summary

Overview

This report is about the identification of children at risk of high cost/harm adult outcomes. It summarises findings from the UK Birth Cohort data about the extent to which information about children and their family environments is predictive of later outcomes. The outcomes are those which tend to be associated with personal harm for them as young people and adults and social cost for those in their environment and wider society.

We find a potentially very high level of predictability in the extent to which we can identify those at risk using early childhood information about family context and child development. In our view these findings are a challenge to which current central and local government should respond with appropriate and measured policy in the interests of social inclusion, personal welfare and the wider economic and social development of the UK.

However, the relationship between childhood risk and high cost or high harm outcomes in adolescence or adulthood is not deterministic, mechanistic or inevitable. There are many steps on the pathway from risk to outcome. There are children at risk who do not experience harmful outcomes and there are children with low apparent or observable risk who do.

Therefore policy responses must allow for flexibility and change. Administrative data should always be augmented by local level, practitioner knowledge and the appropriate interventions should also be selected by local level practitioners who should work closely alongside local communities and agencies to avoid rigid tracking or excessive and unnecessary stigmatisation of vulnerable young people and their families.

Key general findings on forecast accuracy

In the findings reported the method applied is as follows:

1. We run a statistical regression to predict known adult outcomes on the basis of known childhood risk factors.
2. Using these results we create a prediction for each individual based on their level of childhood risk.
3. We select a group of children with the highest level of childhood risk such that this group comprises the same proportion of the sample as does the group that actually experience the adult outcome.
4. Thus, we have two groups; a group predicted to have the high harm, high cost outcome and a group that does in reality have the outcome. The analysis is based on a comparison of these two groups.

5. Because not all of the survey participants answer all the questions at all the stages of childhood, we create upper and lower bound estimates of forecast accuracy. The upper bound results are for those sample members for whom we have full information. The lower bound results are based also on cases without full information, giving a larger sample size but lower forecast accuracy.
6. We have undertaken this analysis for a great many outcomes of adult deprivation in three datasets.

Key findings are as follows:

1. Using only cases with full information, and using information to age 11 to predict multiple deprivation at age 23, 70.8% of those who go on to experience multiple deprivation at age 23 can be identified at age 11 (the true positive rate). 1.4% of those who do not experience adult deprivation are falsely predicted to do so (the false positive rate).
2. When cases without full information are also used, the true positive rate is 41.1%, with a false positive rate of 5.2%. The true rates are likely to be between these upper and lower bound estimates, but closer to the upper bound.
3. For the 1970 cohort, when only observations with full information are used, those with multiple deprivation in terms of 2 or more of 5 key outcomes at age 30 can be predicted in 87.1% of cases using data to age 10, with a false positive rate of 1.1%. When cases without full information are also used, the true positive rate is 43.7%, with a false positive rate of 8.1%.
4. Using only 5 constructs or pieces of information, assessed up to age 10, we can identify 35% of those who will experience 2 or more of 5 key high cost/harm outcomes at age 30, or, using a slightly different set of 5 constructs, 35% of those who will experience 9 or more from 30 diverse high cost/harm outcomes.
5. Using only 6 constructs assessed in childhood and also allowing for interactions between these constructs, sub-groups can be created that have probabilities that range between 17% and 77% for multiple deprivation in adulthood.
6. The probability of experiencing multiple adult deprivation defined in terms of experience of at least 5 of 32 outcomes is 27% for those at moderate to low risk (the 50% at lowest risk), 37% on average and 86% for those defined as at highest risk (the 5% most at risk) based on their childhood family circumstances to age 10.
7. For higher levels of adult deprivation, this ratio increases. Thus, the probabilities of experiencing 10 or more of the 32 outcomes are 1% for the low risk group, 12% on average and 51% for the high risk group. This highlights the very strong relationship between high childhood risk and

multiple adult deprivation. Although none of the 50% at lowest risk experienced 12 or more of these outcomes, 49% of those in the highest risk group did, a very stark difference.

A context for policy responses

In our view it would be irresponsible and socially and economically inefficient to ignore this very high level of capacity to identify early on those at risk of high cost, high harm outcomes. An intelligent system of policy implementation would both respond to these signals of risk and learn from the resulting evidence on implementation to improve its own identification capacity and policy responses.

However, it is also vital to remember that the issue of identification is distinct from that of causation. The fact that we can identify those at risk of high cost/harm outcomes using a small number of key measures does not at all mean that these measures are the causal mechanisms responsible for the outcomes observed or that the measures supply clear indications about what policy initiatives should be employed, when or by whom.

Rather, the measures proxy for a wide ranging set of features of circumstance, development and chance that are the context within which further development takes place. This development is neither deterministic nor inexorable and the degree of predictability does not indicate that the processes that lead to particular outcomes are mechanistic or amenable to obvious, centrally determined interventions.

The particular interventions or supports that will make a difference to the pathways that lead from social address and early development to adolescent and adult outcomes must be applied at a local level and depend on there being skilled and well-resourced local practitioners who are able to make more informed assessments of need and risk than are possible in the national survey data used here.

Just as the Department of Health depends on GPs and other practitioners to make decisions about diagnosis and treatment, so should those concerned with social policy have available to them a system of diagnosis and response that is informed by local level experience and skill. In fact, equivalent if more advanced systems for health care are being developed in the NHS and internationally¹.

The ethics and legality of data matching

A set of very important caveats must in our view also be taken into account in any discussion of appropriate policy responses to the findings in these data, namely issues about the actual data that can or ought to be collected in administrative terms, about the ethics and legality of access to data and about data linking and about the use of the data in the targeting of interventions.

¹ See for example, John Billings, Jennifer Dixon, Tod Mijanovich, David Wennberg, (2006). "Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients." *BMJ* 2006;333:327

The data used here are not available for use in policy-related exercises and are only for use in scientific research under strict ethical guidelines that ensure anonymity, i.e. that no individual, family or other institution can be identified in any way.

However, information is regularly gathered in schools, doctors' surgeries and elsewhere that in fact might be considerably more predictive for adult outcomes than that collected in the datasets investigated in this research study. Moreover, teachers, social workers and other practitioners routinely form assessments and perceptions of children that can be remarkably accurate about their level of risk. Thus, there is no information barrier to the application of the findings of this study to policy. The barriers are rather in terms of ethical, legal and practical issues that require detailed consideration and debate.

Accuracy is far from total

There will always be false positive and false negatives. This means that early identification cannot be a final or absolute marker of risk and that any system of policy intervention must build in the capability to undertake closer monitoring of risk before intervention is determined and to change assessments of risk in the light of new information, chance events and developmental and contextual shifts. Children change considerably during childhood as do family circumstances and this needs to be allowed for.

It is also to be remembered that the data used in this study were not developed for this purpose and had they been so then it is likely that different measures would have been collected at different ages using a different sample frame. Additionally, there are methodological issues not resolved in this report, in particular the treatment of missing values or missing cases.

Towards an intelligent system for preventative action

We propose in this report a system of policy delivery based on accurate information about individual, family and community context and development. We have shown that with the very rich data available in the UK Birth Cohort Studies it is possible to accurately identify children and families who would benefit from appropriate and effective intervention, were such interventions available. We have also shown that under reasonable assumptions, such interventions are likely to be cost-effective and we have set out a framework for assessing the cost-effectiveness of intervention.

In reality, the type of data and information available to policy-makers will be different to that of the Cohort Studies and so the information is presented as a guide to what might be possible rather than a definitive template. A great many other important longitudinal studies exist around the world and in the UK and each could also provide indications about the most useful topics of measurement. Many other researchers have investigated these issues and would have much to add based on their own analyses of these datasets.

At local level, using administrative data as well as local practitioner judgements such as those of teachers, medics, social workers and others, it would be possible to add to the forecast capability indicated here.

It would not be necessary to collect detailed data on all children. Rather, we propose a system of risk monitoring at which certain levels of risk would trigger greater monitoring and assessment, and ultimately, if judged appropriate, intervention. This is the same process as is followed in relation to medical practice.

The specific measures required

Our findings indicate that measurement of children's own achievements and teacher ratings, particularly from age 7 and beyond can be particularly predictive. Before this age, measurement of developmental health and social and family context will be more important.

The precise measures that are most relevant depend to a certain extent on the outcome of interest. Where these are related to mental health, for example, then earlier measures of mental health will be particularly relevant. For other outcomes such as drug use, worklessness or violence in the home, for example, then different measures may be more predictive. There will be some commonality in the core measures required but it is rarely possible to say that a particular risk factor will lead to a particular outcome. Important childhood risk factors indicate that some degree of multiple adult deprivation is likely and that the individual may struggle with a number of features of adult life. It is often hard to gauge precisely which features of adult deprivation are most likely to be experienced.

It seems likely, therefore, that the most useful framework for developmental measurement and assessment would start from birth with indicators of childhood health and development, together with measurement of family income, education, parenting skill and social ties to the neighbourhood or in terms of wider social and familial networks. As children mature, teacher ratings will become relevant and should be built in. In cases where these forms of measurement indicate high levels of risk, then developmental knowledge of the children's own physical and mental health, behaviour, attitudes and aspirations might be added, together with more detailed information about the family and social context, in order to inform decisions about intervention and support.

The costs and benefits of policy intervention

One reason why it may be appropriate not to respond to these signals of risk would be if there were no appropriate interventions. Other work carried out as part of the Children and Young Persons Review, 2006, strongly indicates that such interventions do exist, provided that they are appropriately funded and targeted on those in need.

In order to clarify some of the costs and benefits of these findings, in Section 6 of this report we demonstrate how great a social saving could be made were

appropriate interventions to be carried out, drawing on the previous analysis of identification and targeting. We show that

1. Using prior information on risk of the outcome we can significantly improve the cost effectiveness of intervention by targeting intervention on those most at need.
2. Under reasonable assumptions about the social cost of high cost, high harm outcomes and about intervention effectiveness, our knowledge of individual risk is sufficient to provide a basis for the targeting of cost-effective, early intervention.
3. It is possible to use this information about the identification of those at risk together with the cost-benefit framework set out, to assess the likely cost-effectiveness of any specific proposed intervention to address any specific outcome or sets of outcomes. Use of the framework in this way would provide a common yardstick for intervention choice.

We would recommend use of this system of policy evaluation to assess the likely value of different policy initiatives that may be proposed.

Two issues that need greater reflection and analysis are the questions of stigma and moral hazard. Both present problems for this form of policy system.

1. Stigma. It is very important that early tracking and early intervention do not work so as to reify the problems they are designed to remedy. This would happen if the response to early signals of risk was to create artificial, rigid and exclusive categories. Early prevention will not work if individuals and families simply become categorised as problems. This will exacerbate many of the issues of social exclusion and disengagement that are in part responsible for the high cost, high harm outcomes observed.

However, neither can early intervention be fully effective at remedying the deepest problems of social exclusion if it is only ever voluntary. This has been a long-standing problem for those working in social services who have to make difficult judgements, for example, about when to take a child into care. What we are proposing here is that information be gathered and interpreted more effectively and earlier in the lives of children to make such dramatic and permanent choices less necessary rather than more so.

2. Moral hazard. It must also be a concern that where intervention is of a more positive kind, involving the extra expenditure of resources, then those at risk and others may experience or perceive a benefit from adding to risk rather than reducing it, in order to benefit from the additional resource. Related to this problem is the difficulty of ensuring that resources do indeed go to those most at risk. The inverse care law operates in all areas of social policy, as those with most resources of time, knowledge, social inclusion, income and ambition use those resources to ensure their access to public services, at the cost of those with fewer of such resources.

Again, these challenges can only be met by a system in which informed and skilled professionals are able to make the important judgements. Moreover, many of the elements of risk that we are proposing be observed would not make rational choices for anyone seeking extra resources and so the moral hazard problem may not be as great as sometimes supposed.

We already have in place a related system that may provide good evidence about what works in such terrain and about the risks and challenges of this type of identification and response. The SEN system is intended to determine the allocation of extra funding and intervention to support the learning of a relatively large proportion of the school population who have particular extra need. In this sense, the system rather mirrors the type of mechanism being proposed here in that there are many forms of SEN and many potential interventions to address these needs.

SEN is a particular risk indicator and the SEN system has its own challenges and difficulties. A broadening of the range of interventions that are triggered by a new set of risk indicators as proposed here would enable the formation of a system that i) identifies, ii) tracks and monitors, and, iii) supports and protects children and families so identified as having levels of developmental risk that are likely to lead to subsequent problems and the further inter-generational and social transmission of the difficulties.

Table of Contents

1.	RESEARCH FOCUS AND OBJECTIVES.....	1
1.1	BACKGROUND.....	1
1.2	METHODS.....	3
2.	THE PREDICTION DATA AND OUTCOMES.....	7
2.1	THEORETICAL FRAMEWORK.....	7
2.2	DOMAINS OF MEASUREMENT.....	8
2.3	PREDICTIVE MEASURES.....	12
2.4	OUTCOMES.....	13
3.	FORECASTING RISK 1: THE SIGNALLING POWER OF PRIOR DATA.....	16
3.1	METHOD.....	16
3.2	RESULTS: IDENTIFICATION USING ALL CONSTRUCTS.....	18
3.3	SUMMARY.....	20
4.	FORECASTING RISK 2: WHICH MEASURES?.....	22
4.1	METHOD.....	22
4.2	RESULTS 1: TRUE AND FALSE POSITIVES.....	26
4.3	RESULTS 2: ROC CURVE AREAS.....	29
4.4	RESULTS 3: CLASSIFICATION TREES.....	31
4.5	SUMMARY.....	35
5.	MULTIPLE OUTCOMES – THE FERGUSSON MODEL.....	38
6.	MODELS OF THE COSTS AND BENEFITS OF INTERVENTION.....	41
6.1	OVERVIEW AND SUMMARY OF SECTION.....	41
6.2	CALIBRATION OF COST-EFFECTIVENESS.....	42
6.3	SUMMARY.....	51

1. Research focus and objectives

The focus of this project is on assessing the predictive power of different types of indicators for predicting adolescent and adult outcomes related to social exclusion, associated with substantial social and personal costs.

We assess the relative predictive power of information gathered from different sources (teachers, parents, the children themselves through different methods) at different ages, and in relation to risks from different contexts (family, school, neighbourhoods and so on).

These risk and protective factors are not evidence of specific causal mechanisms (see below on causality) but are to be interpreted as signals about the likelihood of negative outcomes, the key question in hand being the extent to which it is possible to identify those sample members who will experience costly adult social exclusion.

1.1 Background

Under a model of progressive universalism in policy delivery, it is important to identify those with greatest need at the earliest possible opportunity and to provide appropriate support. If one recognises the objective of minimising the social costs that result from social exclusion and related outcomes of low productivity, social division and inter-generational inequity, then this requires identification of the relative risk of these outcomes for individuals, households and communities so as to target support. A great deal of intervention has been targeted at the community level but because of important within-community differences in need and access to resources, often the individuals and families most in need are missed by such an identification strategy. The ideal model would be one that recognised each of these units of analysis and used all in the identification strategy. The contribution of this research project, based as it is on individual level birth cohort data, is particularly but not exclusively focused on individual level targeting.

It is important to emphasise that this information is not by itself sufficient for an accurate assessment of the value of intervention or of the relative effects of interventions at different ages. That assessment would require knowledge in relation to three types of question:

1. what are the factors that indicate risk (signalling question)?
2. what are the related causal mechanisms (causal question)?
3. to what extent is each mechanism amenable to policy intervention (policy effectiveness question)?

The signalling question

This report focuses on the first of these questions. The key research objective has been to assess and describe the extent to which accurate predictions can be made during childhood about who will have high cost/ high harm outcomes later in adolescence and adulthood.

1. Research focus

The precise level of accurate identification required to make intervention cost-effective depends on the relative costs of intervention and failure. There is no a priori minimum proportion of accurate forecasts that is required to suggest that identification and intervention is cost-effective. One might assess the rate of accurate identification of children at risk (“hit rate” or “sensitivity”) for a particular outcome to be as low as, say, 30% but if the cost of failure to intervene and the success probability of intervention are both high then intervention may nonetheless be cost-effective. The required minimum hit rate depends on the trade-off of costs and can only meaningfully be identified on the basis of an assessment of this trade-off.

These cost/benefit trade-offs are modelled in this report under a range of assumptions about the social cost of outcomes, programme effectiveness and the costs of intervention. The empirical analysis of the identification of at-risk individuals is fed into the model to assess the cost-benefit ratios of specific interventions in different scenarios. This information is useful in providing a framework for the assessment of the social value of any specific intervention targeted at the prevention of the outcomes considered in this report. This model is described in more detail below.

The report also provides a description and application of a method (based on Receiver Operating Characteristics (ROC) curves) for assessing the appropriate level of universality or targeting of a specific intervention.

Risks of identification

There are costs associated with excessively rigid targeting schemes that risk stigmatising children and damaging their self-concepts, thus adding to their level of risk. It is also helpful to recognise that the assessment of true or accurate forecasts for individuals will change over childhood as children’s balance of risk and protection interacts with their own development to change the nature of the forecast for them. Thus an individual identified as high risk at one age will not necessarily be high risk at another age and there may be a substantial discontinuity of this kind and vice versa in terms of those becoming at risk.

The contribution of this study is to provide an accurate and broad assessment of the potential accurate identification of need for interventions of different types at different ages for different outcomes. Since the identification issues may be rather different for different outcomes, the breadth and interdependence of outcomes considered makes the study particularly innovative and informative.

The causality question

We have undertaken this work within a coherent framework of prediction that clarifies the relationship of the measures used to the outcomes described. In other words this is not a simple data-mining exercise because it is important to found the discussion of the predictive power of the different measures on

clear theoretical foundations that consider issues of causality and process, even if we do not resolve these issues.

The policy question

This project has not assessed the feasibility and effectiveness of interventions, instead providing information that can be used in a model to ascertain the cost-effectiveness of intervention under a range of assumptions about intervention effectiveness.

Nonetheless, the specification of this signalling or identification question does raise issues for intervention design. It is important that the policy response fit with and be responsive to the identification process. The identification of an individual at risk should not necessarily trigger instant costly intervention. A well-designed and intelligent system of risk management would recognise the need to monitor levels of risk over time and aim to intervene at the time of greatest effectiveness. The identification of this time and of the appropriate intervention may not be best judged from the perspective of administrative or quantitative data of the kind being used here but is most effective if drawing on the skill, judgement and experience of a lead professional who knows the child and child's contexts.

This extra layer between identification of risk and subsequent action is vital if stigmatising effects and badly targeted interventions are to be minimised

1.2 Methods

As discussed above the central focus for this research is the prediction of social exclusion outcomes.

Data

We draw on three datasets: the 1958 and 1970 UK birth cohorts, the Avon Longitudinal Study of Parents and Children (ALSPAC), and the Longitudinal Survey of Young people in England (LSYPE). Results from the LSYPE are not ready for this draft of the report.

National Child Development Study (a.k.a. 1958 Cohort)

The sampling frame for the NCDS was the population of all children born in Great Britain in the first week of March. Data was collected at birth and then again at ages 7, 11, 16, 23, 33 and 42. During childhood interviews were conducted with parents, teachers and medical officers and since 16, the sample members themselves have responded to extensive interviews and questionnaires.

British Cohort Study (a.k.a. 1970 Cohort)

The BCS has a similar structure to that of the NCDS with sampling in the first week of April 1970 and data collected at ages 0, 5, 10, 16, 26 and 30. Again, there has been extensive data collection by interview and questionnaire with

the sample members, parents, teachers, health visitors and medical officers. Testing of cognitive development has also been undertaken.

The Avon Longitudinal Study of Parents and Children

ALSPAC is an ongoing, geographically-based population birth cohort study. Pregnant women who were resident in the area that was Avon, South West England and expected to deliver their baby between 1/4/91 and 31/12/92 were eligible for inclusion. 14,451 women joined the study - approximately 85% of those eligible. 13,988 children were alive at 1 year. The characteristics of the ALSPAC cohort are generally representative of the 1991 UK census data for the same geographic area. Approximately 11,500 children (now aged 13-14) and their families are still participating in the study and these children are still largely representative of the original cohort, although there is an increased bias towards more affluent families and therefore under-representation of very deprived families and non-white families. All participating children are invited to every ALSPAC assessment clinic; the attendance rate is usually 60% - 65%. Data has also been matched in from the DfES PLASC dataset providing Key stage scores for sample children at ages 7 and 11.

The '58 and '70 cohorts have the strength that they have adult outcomes and also multiple measures through childhood from various sources. The ALSPAC dataset can currently only predict to Key stage 2 but given the extent to which we can predict adult outcomes from Key stage 2 scores, the link to adulthood could be modelled. Inclusion of these data would make the study more current and also provide a range of measures that were not collected in the earlier studies, and, measures at more ages during childhood.

We describe the measures used in more detail below.

Assessing the accuracy of identification

The simplest model takes the form of a linear prediction equation, with each relevant adult outcome treated as a separate variable. The sets of risk and protective factors are entered on the right hand side and the findings indicate the extent to which each set of risk and protective factors enhances predictive capability relative to false predictions.

We present information on the predictive power of each separate childhood construct assessed in the data used for this study. We also present information on the predictive power of measures grouped according to the age of the child at which the constructs were assessed and according to the domain of measurement.

There are a number of ways of assessing the predictive power of measures of risk and protection. It is common in statistical analysis to assess the quality of the prediction in terms of the proportion of variance in an outcome explained by the predicting measures, or in terms of the statistical significance of predictor variables. To focus on the issue of the identification of individuals at risk, we use an alternative approach.

1. Research focus

The focus of the analysis here is on the quality of the identification of those at risk of high cost/ high harm outcomes and so being potentially triggering an intervention or interventions. The measures of childhood circumstances and child development are used to predict later outcomes. This generates a statistical model that effectively weights each childhood construct in terms of its importance as a signal of the risk that the outcome will occur. These weights are then linked to the observed childhood data for sample members to generate for each sample member a probability that they will experience the high cost/harm outcome. Much of the analysis then focuses on this probability. The probability can be thought of as an index of the risk for each individual that they will experience the high cost/harm outcome. We will call this probability the risk propensity score.

For each individual in the data, the data and analysis provides an observation of whether or not they actually experienced the high cost/harm outcome (actuality) and a prediction based on their earlier childhood risk factors of their likelihood of experiencing the outcome (prediction).

A very useful way of assessing the predictive capability of different measures of development and context is to compare the actual outcomes for individuals with what would be predicted for them using the risk propensity score derived from those measures of development and context. To do this it can be helpful to use a threshold in the risk propensity score, above which individuals will be said to be predicted to experience the outcome, below which the prediction will be that they will not experience the outcome.

This approach creates a two-way table of probabilities, as shown in Table 1.

Table 1: A tabular description of forecast accuracy for binary outcome.

		Predicted outcome	
		No	Yes
Actual outcome	No	True negative	False positive
	Yes	False negative	True positive

These probabilities are discussed in considerably more detail in the modelling section (section 6) towards the end of this report. Here we simply note that two proportions provide a very good indication of the predictive capability of the measures that were used to generate the predicted outcome, namely:

1. *The true positive proportion*, here defined as the proportion of those who in actuality experienced the outcome who are accurately predicted to do so by the risk propensity score;
2. *The false positive proportion*, i.e. the proportion of those who did not experience the outcome who are erroneously predicted to do so by the risk propensity score;

These two proportions provide an indication of the accuracy of the identification of those at risk of high cost/harm outcomes.

As Table 1 highlights, the usefulness of the prediction information provided by this study can be in assessing the relative likelihood for a range of outcomes of:

- i) True negatives - accurate identification of those not requiring extra support;
- ii) False positives - deadweight intervention;
- iii) False negatives - missed intervention needs resulting in failure to intervene;
- iv) True positives - accurately targeted interventions.

There are costs, savings and/or benefits associated with each type of assessment status. The measures are used to model the cost-benefit ratio for specific, targeted interventions in section 6.

It is important to note that as the threshold determining the prediction changes, there will be important implications for the proportions of true and false positives. A low threshold which creates a prediction that most individuals will be predicted to experience the outcome is likely to accurately identify a large proportion of those who eventually do have the outcome (a high true positive rate) but also to falsely suggest that the outcome will be experienced by a lot of those who did not in reality experience the outcome (a high false positive rate). Thus, the decision rule which sets this threshold and determines the proportion of those who are to be classified as at risk or positive predictions is a crucial factor in the model.

In the analysis of ROC curves, discussed below, this threshold can be allowed to vary, generating a whole distribution of different levels of true and false positives. This is very useful in determining not just the accuracy of prediction but also in assessing the most efficient level of targeting versus universality of specific interventions.

2. The prediction data and outcomes

There are a great many features of the background and development of children that may provide signals about likely future outcomes. In order to structure the analysis and to provide a foundation for interpreting the results we provide a suggested theoretical framework.

2.1 Theoretical framework

The general theoretical framework for considering these relationships is derived from the ecological model of Bronfenbrenner², together with more recent developing science in bio-medical, psychological social science fields. We emphasise the following key principals of human development, particularly during the period of childhood and adolescence:

- i) There are features of biology and temperament that are determined by inherited, genotypic characteristics and which have important implications for subsequent development and outcomes;
- ii) These genetic features interact with features of the environment that therefore also have important implications for development and outcomes;
- iii) There are many features of the environment that may impact on human development to influence the relative likelihood of high cost/harm outcomes for different children but key amongst these influences, particularly but not exclusively in the early years are those experienced in the family;
- iv) Other contexts also matter, increasingly through childhood. Particularly important other contexts are peer groups, schools, neighbourhood factors and wider social and community networks. As children move through childhood into adulthood, a wider array of contexts start to play an increasingly important role, particularly factors such as labour markets, work and college environments, close personal relationships and other interactions with adults in positions of authority and influence;
- v) These contexts interact in their influence in the development of children and young people;
- vi) The personal skills, attributes, characteristics, temperaments and self-concepts of children and young people also contribute to their success in negotiating the pathway through childhood into adulthood. Children have some degree of agency in their own development but many of the underlying attributes which support agency will also be strongly influenced by background and contextual features.
- vii) Thus development occurs through a multi-faceted and complex series of interactions between the developing child and the other agents and agencies with which the child forms relationships;
- viii) Pathways are not set in stone as a result of early development or contexts and there is always the possibility of discontinuity or plasticity which can change likely outcomes. However, disadvantage and background stress exerts continuous and

² E.g., Bronfenbrenner, U. (1979). *The ecology of human development*. Cambridge, MA: Harvard University Press.

compounding pressures on young people such that the discontinuities in development tend to privilege those from more advantaged backgrounds, i.e. with more supportive, nurturing and well-resourced environments.

These principles are intended to clarify the analysis undertaken on the signals of risk. It should be clear that not all of the factors that influence development can be measured in standard data collection exercises. Many of the important causal mechanisms are the result of very rapid intra-personal neurological and biological processes that cannot commonly be observed and which interact with features of the developing child and of social and family contexts in dynamic transactions which are not yet well understood. Therefore, it would be wrong to expect that these causal processes can be easily manipulated in a theoretically grounded science of policy delivery.

However, as a result of these causal processes there are relatively stable features of personal development and of social and family context which indicate the general trends in development and carry strong signals about likely outcomes. These signals are constructs which can be measured and modelled.

2.2 Domains of measurement

The features of contexts that are supportive of positive development can be classified in terms of their relationship to the developing child, namely: *Proximal processes* are the features of the interaction between the child and others in the family, typically the primary care giver. The proximal processes, such as home learning, are child-centred interactions, reciprocal, dynamic and transactional; *Distal family factors* refer here to observable features of the members of the family or of the general family context as a unit that impact and influence proximal processes, as well as having direct impacts on child development.

One might distinguish further between types of distal factors. Within this set of constructs there are family characteristics that are more directly related to the proximal zone of interaction with the child than are socio-economic and demographic factors. A factor such as mother's depression will impact on the child partly through observable, measurable parenting factors such as maternal warmth and responsiveness but also indirectly as a feature of the phenomenological world of the child. A more distal factor such as mother's age, for example, is not in any meaningful sense a part of the child's phenomenological world. Mother's age will tend to mean that certain features of the mother-child relationship will be a certain way, younger mothers tending to give less consistent care, for example, but that is not necessarily so and certainly won't be true in every case. Thus the connection is not a feature of the phenomenological world of the child in the same way as is mother's mental health. It is more distal. Age matters because of what it may tend to mean for the zone of interaction but it is not the age of the mother that matters in and of itself.

In two of the datasets used here (the 1958 and 1970 Birth Cohorts) there are not many measures through which it is possible to distinguish between socio-demographic distal factors and the more personal features of the family that impact more directly on the phenomenological world of the child and so we do not maintain this distinction. However, in the ALSPAC data there are more measures that can be distinguished in these terms, particularly in terms of the attitudes and well-being of the mother and so we construct an additional domain of measurement for this dataset, namely "*mother characteristics*"

One reason why this distinction is useful to maintain and experiment with is that the socio-demographic distal factors can be relatively easily recorded as part of a normal census type data collection, such as through the DWP, DfES or ONS. The other family characteristics such as mother's attitudes to school, locus of control and so on, like depression, are more personal and so harder to identify from census type information, although may be available through DH mechanisms. Therefore, the distinction is relevant to the question of how an expert system of policy implementation could identify children "in need" of support or other intervention before problems become entrenched and/or had severe consequences.

Nine domains of measurement

This framework provides us with 9 basic classifications of variables or domains of measurement, namely:

Family context

- 1) Distal family factors
- 2) Mother characteristics
- 3) Proximal family process

Other contexts

- 4) Pre-school factors (ALSPAC only)
- 5) Neighbourhood factors
- 6) School factors

Child factors

- 7) Features of the cognitive development of the child
- 8) Features of the affective and behavioural development of the child
- 9) Health

Most of the measures used to predict outcomes in adulthood and adolescence can be classified into one of these 9 domains. Other datasets also include data on a wider range of contexts in which case the list of domains can be extended. Moreover, some datasets might provide the capability to distinguish between distal and proximal features of contexts other than the family, but that is not the case with the datasets used in this study and so the proximal/distal distinction is only maintained here for the family context.

There are some complex issues in the classification of constructs that are not always straight-forward to resolve. For example, if a child attends an independent school is that a feature of the school context or more accurately

of the family context as it is indicative of family characteristics such as wealth and aspiration as well as of the likely characteristics of the school. Similarly, if a young person spends a lot of time in the company of peers engaged in activities that might be classified as anti-social, does this provide a signal about the peer group, properly categorised as part of the peer context, a separate domain of measurement, or is that more indicative of the affective and behavioural development of the young person.

These issues cannot be simply resolved and result from the obliqueness, multiplicity and inter-connectedness of the underlying causal mechanisms.

Our strategy in this report is to be transparent about the ways in which measures are classified so that inferences can be appropriately discussed.

Commensurability of measurement

The datasets used for this study are high quality, longitudinal datasets that are internationally renowned as being a major scientific resource of a kind not available in such depth and breadth anywhere else in the world. They are the result of long-term investments by scientific research councils, academics and sponsoring Government departments and make possible an in-depth and detailed analysis of the research questions of concern in this report.

However, that said, it is nonetheless important to recognise the limitations of these datasets and the implications of those limitations for inference in relation to the findings of this research.

One particularly important issue is that for a number of reasons comparability across ages within studies and between studies must be undertaken with considerable care. The measures were collected as part of three separate studies that were in differing amounts and at different times, differently intended to meet the different and changing research objectives of bio-medical science, social science and policy. The 1958 Cohort Study (NCDS) was originally intended to support analysis of perinatal mortality and was initially developed on the basis of a research design appropriate for that study. The 1970 Cohort Study (1970 Cohort) suffered from the co-existence of a teachers' strike with the age 16 data collection that led to big problems of missing data. The ALSPAC study differs in having greater geographical concentration than the other studies, but lacks national coverage. It has more regular data collection. The different datasets have used different forms of data collection at different ages of the study children. The main focus of the 1958 and 1970 cohorts was initially on health and development and there are few available measures of the details of family life.

The precise forms of data collection in each study by age are set out in Table 2.

Table 2 here

Because of the differences in focus of the study at different ages the quality of measurement across domains and ages varies quite considerably such that there is not commensurability of measurement across the 9 domains. Measurement of the neighbourhood context is particularly weak, there being few easily available measures collected in the data and the relevant constructs being less well understood and possibly harder to measure in individual level data than those for the family context. Thus scientific comparison of the relative importance of these different constructs as indicators of development must be undertaken with caution. However, the purpose of the analysis presented here is to assess the differences in relative predictive power of the different measures. This comparison is informative of the capability of the different available measures at the different ages. So long as this is seen as informative in relation to the current state of knowledge rather than definitive then in our view inference will be appropriate.

Accessibility of measurement

It is also important to emphasise that the capability of survey agencies to collect data as part of a scientific study may be different in important ways to that of Government agencies engaged in policy. Government-led data collection may, subject to legal and administrative issues, have good access to administrative data routinely collected by education, social services and health departments and agencies, although there remain barriers to the linkage of such data and issues of anonymity, ethics and access to the data.

The legislative power of Government and its ability to regulate and create frameworks for assessment and monitoring may change the context within which important constructs are to be measured such that the data collected as part of the policy framework may be more accurate and suffer less from problems of attrition and missing values than do the data analysed in this study. To the extent that is the case, the results from this study provide a lower bound or under-estimate of the true predictive power that might be generated in a policy-led system of data collection.

However, it may also be the case that for some sources of data collection such as parents and children the access to information may be easier for a disinterested survey design company than it is for government agencies, particularly if there may be punitive or otherwise unwanted repercussions resulting from particular assessments. This may result in mis-measurement or missing values, perhaps particularly for the most important cases if the data collection exercise is not handled sensitively and if the policy system that the data collection informs is seen as a threat rather than an opportunity. Balance is called for here as if the data collection exercise is seen as leading to potential benefits then that may also lead to moral hazard and systematically biased measurement error.

The situation is different in relation to the assessments of children and families by teachers, doctors, health visitors, social workers and other front-line professionals and service providers. Here again there will be important ethical and administrative issues but measurement accuracy may be less

biased than is the case for data from parents and other family members. Indeed these professionals may have access to very important information and judgements that were not available through the data collection of these national surveys.

To summarise, there is a very important and delicate relationship between the data collection element of a targeted policy implementation strategy for high cost/harm individuals and families and the policy mechanisms triggered by that data collection. There are important issues about how the data is analysed, by whom and with what results that will impact not just on the success of the policy implementation but also on the quality of the data collected. It is likely that a good degree of experimentation and evolution of the system will be required and the information/policy relationship will be a key element of the design and evolution of this policy system.

As far as the contribution and focus of this report is concerned, the key issue is the relative predictive power of information from these different sources. Is the information from teachers as predictive and valuable as that from parents, for example? How important is information gained from questions to children as compared to information from, say, tests of children's development that may be routinely collected in school and medical settings? Preliminary answers to questions such as these are provided in this report.

Nonetheless, it is to be hoped that subsequent development of local data collection will improve the quality of data available for analyses such as those presented here and that will improve our understanding of the predictive importance of information from different sources.

2.3 Predictive measures

Table 3 provides summary statistics for all the variable and constructs used to assess risk, protection and development for the sample members and which are to be tested as signals of likely outcomes in the analysis described in subsequent chapters. Table 3 provides these by age and by the domains of measurement set out above.

Table 3 here

In order to clarify the issue of the commensurability of measurement across domains and ages in the different datasets, Table 4 reports the number of constructs assessed at each age in each domain.

Table 4 here

It can be seen, for example, that although the 1970 cohort has assessments of 12 different distal factors at age 5, at age 16 only 1 is assessed. Partly this is a restriction of the data collection, partly of what has been coded to date. It should be borne in mind when comparing the predictive power of distal factors at the two ages in this dataset. Other important points to note are that:

- Mother characteristics and features of pre-school contexts have not been included for the 1958 and 1970 Cohorts;
- The 1970 Cohort at age 16 only includes measures in the domains of affective and behavioural development, with the exception of one distal factor;
- Few measures of the neighbourhood context are assessed;
- Few health measures have been included, particularly for the 1970 cohort and ALSPAC;
- School context measures have not been included for the 1970 Cohort and ALSPAC;
- In the 1958 and 1970 Cohorts there are more measures of distal factors than of proximal processes. The reverse is the case for ALSPAC;

These and similar issues of difference in measurement across the cohorts and across ages in childhood limit the extent to which inferences can be drawn about the true capacity to predict outcomes using measures in different domains at different stages of childhood.

Nonetheless, with these caveats in mind, it is possible to use these results to assess the extent to which outcomes of adolescence and adulthood can be predicted from the available information at earlier stages in childhood.

2.4 Outcomes

In order to maximise the level of generality of the findings, analysis has been undertaken across the three datasets of a very wide range of outcomes. There are many important features of social exclusion and since particular measures may have differential predictive power for different outcomes, we have tested across a wide range of measures so as to provide as detailed an investigation of these relationship as possible and to limit the likelihood that results are only relevant to very specific single outcomes.

The outcomes were coded into a binary form, with positive outcomes denoting a high cost/harm outcome. For each outcome sample members are defined either as positives (i.e. matching the social exclusion status denoted by the variable) or negative.

Table 5 sets out summary statistics for the set of outcomes or dependent variables for this study.

Table 5 here

The outcome measures for the NCDS (1958 Cohort) are mainly assessed at age 42, some time after adolescence. In addition 7 outcomes have been coded at age 23 for that cohort, namely:

- Outcome 33, Age 23: Unskilled or semi-skilled
- Outcome 34, Age 23: Literacy or numeracy problem
- Outcome 35, Had child before 20

- Outcome 36, Had child before 23
- Outcome 37, Teen parent before 23
- Outcome 38, Depressed at 23
- Outcome 39, Low hourly wage at 23

The outcome measures for the 1970 Cohort are mainly assessed at age 30, although several refer to outcomes of adolescence, such as being a teen parent or experiencing high levels of time “Not in Education or Employment” (NEET) before age 18.

It must be noted that whereas some of the binary outcomes operate with a standard cardinal metric such that comparisons between the cohorts can be made, others are based on the relative position of individuals in the distribution in each dataset and so are not comparable. For example, the outcome of high level criminality (outcome 16) is defined as being found guilty in a court of law more than once. This is a relatively objective measure such that the increase in prevalence between the 1958 cohort (1%) and the 1970 cohort (5%) is interpretable as an increase in this outcome between the cohorts. There are two caveats: i) possible change in the reliability of self-reports over the period, ii) the measure for the 1958 Cohort is taken at age 42, whereas that for the 1970 Cohort is taken at age 30. These caveats make interpretation potentially hazardous but the raw data are certainly in line with reasonable assessment of the historical trend. That is not the case for the “NEET” variables (outcomes 29 and 30), for example. For these measures, the number of months during which each sample member was not in education, training or employment is assessed from self-reports. The binary variable indicates that the sample member experienced lengths of time in this state that were in the highest 10%. Therefore, by definition, the prevalence will be 10% and cannot change between cohorts. This definition is useful for the purposes of the analysis of this report but means that the measure does not provide information about the cross-period change. Note, too that the prevalence for outcome 29 in the BCS is 7%. That is because this proved to be a more amenable cut-off for the distribution of NEET status in those data. The cut-off has been selected to best characterise the data on NEET status and to identify a group for whom a negative outcome is most meaningfully recorded.

The outcome measures for ALSPAC are assessed at age 10 and 11 as data for the adolescent period is not yet available. The sample children are assessed in terms of scores at Key stage 2 (age 11) and of clinic-based measures of anti-social behaviour and depressive symptoms at age 10.

Multiple outcomes

We have also constructed indices of multiple cost, harm or social exclusion. This is discussed in more detail in section 5. For the 1970 Cohort, we have constructed an index of 5 high cost/harm outcomes, assessed at age 30, namely:

1. Mental health referral”
2. Social housing”

3. High levels of NEET before age 30
4. Overcrowding
5. Financial problem

Roughly 15% of the sample had 2 or more of these 5 outcomes (outcome 33), roughly 6% have 3 or more.

We have also constructed a measure indicating how many of the 32 outcomes each individual had experienced. We find that roughly 11% of the sample had experienced 9 or more of these 32 outcomes and we take this as a second high risk group (outcome 34). The correlation between these two indicators is fairly high at 0.61.

For the NCDS (1958 Cohort), we have constructed an index of multiple outcomes (outcome 40) based on use of 6 of the measures assessed at age 23 (all measures except low income which has a high rate of missingness). We use a cut-off of 3 of the 6 outcome measures, with 8.7% of the sample experiencing 3 or more of these 6 outcomes.

We also construct a second multiple outcomes (outcome 41) measure based on the age 42 data, indicating the 12% of the sample with more than 7 of the 32 outcomes assessed at that age.

For ALSPAC we constructed an index of 5 measures, comprising indicators of low performance in each of the Key stage 2 tests, anti-social behaviour and depression. 20% experienced 2 or more of these 5 outcomes. We used this as a measure of multiple risk (outcome 7). Note that outcome 4 is also a multiple outcome measure, namely low performance across the board in school tests at KS 2. This group comprised 12% of the sample.

3. Forecasting risk 1: The signalling power of prior data

In this section we assess the capability to identify those with the high cost/harm outcomes indicated in Table 5, drawing on all the predictor variables shown in Table 2. Then in section 4 we focus on specific measures, assessing the predictive power of each particular measure or of small numbers of measures used in combination.

In this current section we assess the accuracy of prediction when all available measures are used. This means that in this section we are assessing the accuracy of prediction when a great many variables are used, depending on the dataset, between 80 and 180. Therefore, the results in this section probably do not precisely reflect the degree of accuracy in early identification that would be achieved in local-based, administrative systems of risk tracking which used only a small number of measures. However, the precise degree of monitoring and observation that would be possible in such systems is not yet known. Local-based systems that had access to detailed observations and assessments of local practitioners would be drawing on very different forms of data that would be more relevant to the problem of identifying risk than are the measures used here which were not collected for this purpose. Therefore, the levels of identification achievable at local level may be greater than that achieved here, with fewer measures.

What the results in this section indicate is the level of accuracy of early signalling of risk in these survey datasets, an indication of what might be possible in local administrative systems.

3.1 Method

There are a number of ways of assessing the predictive power of measures of risk and protection. It is common in statistical analysis to assess the quality of the prediction in terms of the proportion of variance in an outcome explained by the predicting measures, or in terms of the statistical significance of predictor variables. To focus on the issue of the identification of individuals at risk, we use an alternative approach, namely an assessment of the proportions of true and false positives as defined above.

True and false positives

Using linear regression analysis, the measures of childhood circumstances and child development are used to predict later outcomes. This generates a statistical model that effectively weights each childhood construct in terms of its importance as a signal of the risk that the outcome will occur. These weights are then linked to the observed childhood data for sample members to generate for each sample member a probability that they will experience the high cost/harm outcome. This probability can be thought of as an index of the risk for each individual that they will experience the high cost/harm outcome. This probability can be termed a risk propensity score.

For each individual in the data, the data and analysis provides an observation of whether or not they actually experienced the high cost/harm outcome (actuality) and a prediction based on their earlier childhood risk factors of their

3. Forecasting risk 1

likelihood of experiencing the outcome (prediction). This approach creates a two-way table of probabilities, as shown above in Table 1.

For the analysis presented in this note, we undertake a linear regression for each outcome on each construct taken separately, with gender included in the model. This provides a propensity score, for which we use a threshold that matches the sample probability (p_j) of the outcome to identify the group that has the highest levels of risk of the outcome, such that a risk group is identified which has the same proportion p_j of the sample as is observed for the actual outcome.

We can then calculate for each measure the proportions of true and false positives that would result from using that risk score to identify those at risk. The proportion of true positives and the ratio of true to false positives both provide indications about the predictive power or contribution to the identification of those at risk.

Missing data

In this section we assess the predictive power when all available measures at all ages in childhood are used. This leads to very rich regression models with a great many predictive measures. An important methodological issue becomes how cases are treated when information on a particular measure is missing for particular individuals, but when information is available in relation to other measures. It is standard practice to maintain a high sample size in such instances by using mean imputation or other forms of data substitution. Assuming that missingness is at random, conditional on other data in the model, these methods do not cause bias, although standard errors should be corrected.

For these predictive models, the impact of missingness is rather different. If we drop cases with missing data on any single measure then the number of observations maintained in these very rich regression models becomes very low. However, these are cases for which we have the best data and, therefore, for whom the accuracy of prediction will be highest because we have the most information. These results indicate the potential of reliable and high response data to identify children at risk. However, equivalent levels of identification would require that the data were collected without high levels of missingness and attrition, as might in certain circumstances be easier with administrative data collection than with longitudinal surveys. Assuming that cases with full data are not more predictable than other cases, there is no bias induced to the estimate of the true and false positive proportions by generalising from these small samples. However, a degree of sampling bias may result in any small sample analysis and inference becomes more hazardous.

It is to be expected that the average level of high cost/harm outcomes will be lower in the full information data than in the full sample because it tends to be those most at risk on various indicators who are most likely to leave surveys or to fail to respond to specific questions. This is in fact what we find in these

data. Thus the full information samples include fewer cases of high cost/harm than is the case for the full sample.

Therefore, as well as reporting results when only the full information cases are used, we also report, for the BCS and NCDS, results when mean imputation methods are used. These results are based on far larger samples but include cases for whom less information is available. This makes prediction less accurate.

Taken together, the two different sets of results perhaps provide upper and lower bounds as to the degree of early identification possible. The full information cases provide an upper bound, drawing on that small proportion of the sample for which all information is available. Under systems of administrative data collection, the problem of missing data would be much reduced and so these upper bound results may be realisable. However, if the accuracy of forecast for these individuals is particularly high not because they have full information but because the pathways and processes between childhood information and later outcomes are more predictable for these individuals, then the identification rates achieved will be lower. Assuming that administrative data will have less missingness and so less need for data imputation, than the survey data used here, then the achievable level of accurate early identification will be higher than for the lower bound estimates. Providing both sets of estimates enables an assessment of this range of possibilities. It would be possible to experiment with more efficient forms of data imputation to test the levels of prediction in larger maintained samples at the later ages.

In fact, we find that for some outcomes, the mean probability of the outcomes does not vary between the full information sample and the complete sample and that this has no bearing on the difference between the upper and lower bound estimates. This gives up more confidence in the upper bound estimates. We discuss this in more detail in the results section below.

Another point to note is that when all constructs are used and cases with missing data on any single construct are dropped, the regressions may lose so many observations that there are insufficient degrees of freedom for the method to be applied. This is the case for many of the regressions undertaken using all the constructs coded in the 1958 Cohort up to and including age 16 data. This leads to regressions with over 100 constructs and insufficient observations for the forecast statistics to be developed. For these cases the reader will see missing information in the tables of results. It would be possible to undertake more parsimonious regressions using a smaller number of measures in order to complete these tables. We have not yet done this.

3.2 Results: Identification using all constructs

The results in Table 6 report the levels of true and false positive prediction at each age when all previous information is also used. Thus, predictions are built up at each age drawing on the information at that age as well as information from previous ages.

Table 6 here

The levels of true positive identification are striking when all the available information is used in this way, in particular when only those cases with full information are used in the analysis.

Upper bound estimates

For most outcomes in the 1970 Cohort the upper bound true positive rate by age 16 is over 70% and in many cases over 90%, i.e. more than 90% of those who go on to experience the high cost/harm outcome can be identified at age 16, drawing also on prior information. For example, 94% of those who report low self-efficacy (outcome19) could have been predicted to do so at age 16. The false positive rate is only 0.011, i.e. only 1.1% of those who did not report low self-efficacy at age 30 would have erroneously been predicted to do so.

For the multiple risk outcomes, levels of identification of those at risk are also very high. For the age 23 outcomes in the 1958 cohort, using the upper bound full information estimates, the rate of identification of true positives using data to age 11 to predict multiple deprivation at age 23 is 70.8%, as against a false positive rate of 1.4%. When prediction is made to multiple deprivation as far out as age 42 from age 10, 50.4% of cases can be forecast from data to age 10, with a false positive rate of 4.3%. For the 1970 cohort, those with multiple deprivation in terms of 2 or more of the 5 key outcomes at age 30 (outcome 33) can be predicted in 87.1% of cases using data to age 10, with a false positive rate of 1.1%.

These results indicate that where there is full information, very high levels of early identification of those who will go on to experience high cost/harm outcomes is possible, with low levels of forecast error.

Lower bound estimates

As discussed above the number of observations maintained in these very rich regression models is very low because cases are dropped if they have missing information on any of the measures. For the example of multiple deprivation (more than 1 of 5 outcomes) in the 1970 Cohort the sample size is only 399. As Table 6 indicates, 77 constructs were used to make this prediction. These results indicate the great potential of reliable and high response data to identify children at risk. However, equivalent levels of identification would require that the data were collected without high levels of missingness and attrition, as might in certain circumstances be easier with administrative data collection than with longitudinal surveys.

When the lower bound estimate is used, i.e. the estimate using cases without full data, for the same example of multiple deprivation in the 1970 Cohort (outcome 33), predicted from data to age 10, 3559 observations are maintained and the true positive rate is 43.7%, the false positive rate 8.1%. This is a much lower level of accuracy than when only full information cases are used. In part this is because cases are maintained in the regression

3. Forecasting risk 1

analysis for whom no information is available on some or many of the constructs. This introduces noise and inaccuracy into the signals of early risk, leading to lower levels of accurate forecasting.

This is an important caveat. Yet, it should also be noted that if only the earlier sweeps of the survey are used so that data missingness is much reduced, it is still possible to obtain very high levels of true positive identification, even at early ages. For the example discussed above, 36.3% of those who will experience multiple deprivation at age 30, can be identified as being those most at risk of doing so by age 5, with a sample size of 4525. This indicates both that levels of identification can be very high even early in childhood and that the higher levels of identification are not caused by the use of small samples.

Above, we indicated that the full information sample and complete sample differ for some outcomes in terms of the mean prevalence of the outcome but not for others and that this has no bearing on the difference between the upper and lower bound estimates of forecast accuracy.

For example, the outcome of age 30 mental health problem for the 1970 Cohort, the mean prevalence is 24.3% for the 11227 cases in the full sample (Table 5.) For the sub-sample of 401 cases with full information using all data to age 10, this mean is 22.0% (see Table 6.2), not a substantial difference. The differences in the mean does not suggest that the small sample with full information is substantially different to the complete sample and yet the upper bound estimate of true positive rate (71.6%) is substantially higher than that for the lower bound estimate (42.6%).

For the outcome of multiple adult deprivation (outcome 34) indicating more than 8 of the 32` age 30 outcomes, the mean prevalence is 11.5% for the full sample (Table 5.) For the sub-sample of 401 cases with full information using all data to age 10, this mean is 10.5% (Table 6.2), again not a substantial difference. The upper bound estimate of true positive rate (79.1%) is again substantially higher than that for the lower bound estimate (48.4%).

The same is true for many other outcomes, suggesting that the sub-samples with full information do not lead to higher estimates of forecast accuracy due to differences in composition, i.e. that they comprise lower proportions of high cost, high harm individuals. There are also other outcomes for which the mean prevalence in the full information sub-sample does differ to that in the complete sample. This does not appear to lead to any substantive differences in the range between the upper and lower bound estimates for these outcomes, relative to those with less difference in sample composition.

3.3 Summary

It is worth restating some of these findings; by age 5 it is possible to identify over one third of those who will experience multiple deprivation 25 years later in adulthood. By age 10, it is possible to identify between 44% and 87% of those who will experience multiple deprivation as adults, depending on assumptions about measurement and missing data. The lower bound estimate would be improved on by any data collection/assessment

3. Forecasting risk 1

mechanism that ensured full information was collected. The upper bound may be a hard target to achieve as this is the result when only cases with full information are used in the analysis and this may lead to some upward bias in the estimate. However, the mean levels of prevalence do not vary greatly between the full information and complete samples, suggesting that sample composition does not explain the difference between upper and lower bounds. It is likely that the lower bound is biased downwards by use of cases with little information causing error in forecast accuracy.

Nonetheless, depending on assumptions these upper and lower bounds provide a range of reasonable expectation. We would predict that the true picture is likely to be around 70%, that is roughly 70% of individuals who will experience multiple deprivation at age 30 can be identified at age 10.

This would mean that three tenths are not picked up and so unsurprisingly the situation is not set in stone at age 10 and neither should be policy responses. However, in our view, given the stated objectives of government, it would be both unjust and inefficient to ignore this indication of risk, need and potential cost by failing to respond to these clear early warning signals.

4. Forecasting risk 2: Which measures?

We have shown above that subject to important discussions about the treatment of cases with missing information, it is possible to identify in childhood a large proportion of those who will go on to have high cost/harm outcomes. In this section, we consider the contribution to prediction of each specific measure in order to address the question of which measures are most predictive and would be most useful in a system of risk tracking and early preventative intervention and support.

4.1 Method

In this section, in order to assess the specific measures, two methods are used:

1. True and False positives
2. Classification trees

Method 1: True and false positives

This method is applied as above, using each measure separately, we use linear regression analysis to predict outcomes and then assess the accuracy of prediction using the proportions of true and false positives. We do this in two ways:

- a) A narrow, specific model using a fixed rate of risk;
- b) A more general model using receiver operated characteristics

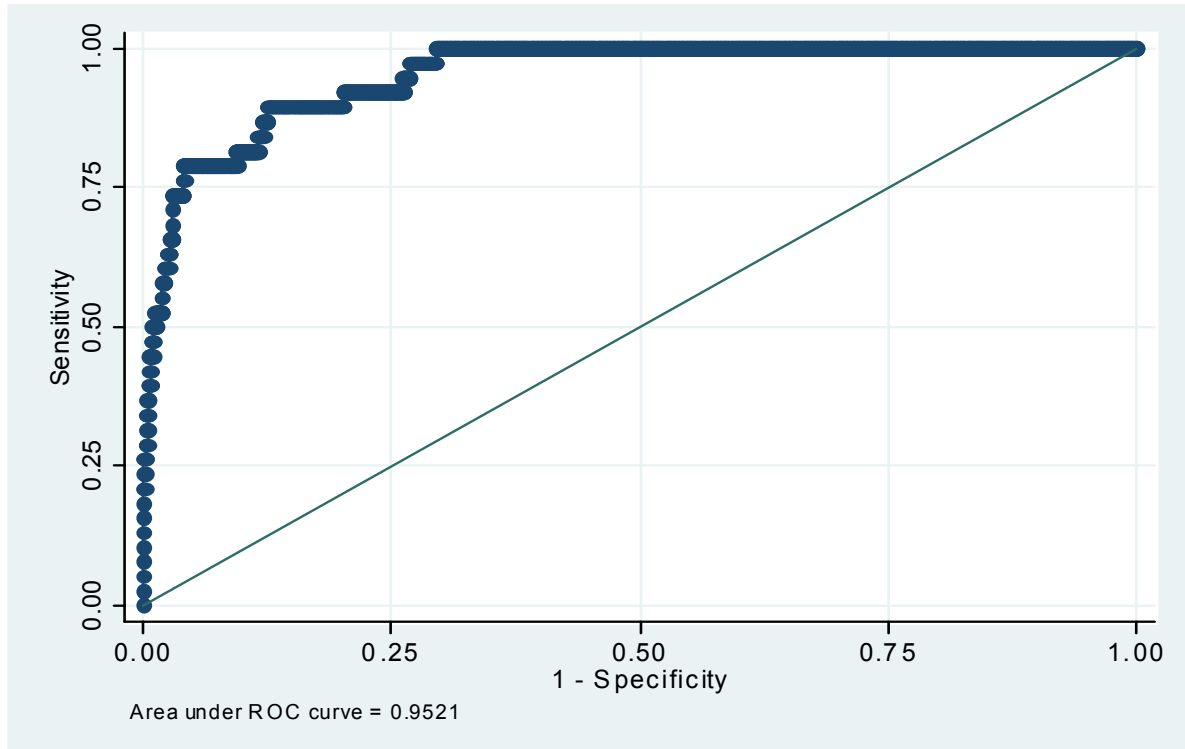
The first approach assumes that there is a fixed proportion of the population that will be classified as at risk, the second allows this proportion to vary. Table 1, above, indicated how true positives are usefully defined as those who i) in truth did experience the outcome under consideration, and ii) were forecast to do so under the decision rule used in the analytical estimation procedure. This decision rule allocates individuals to one or other of two states, D+ (forecast to have the outcome) and D- (forecast not to have the outcome). This is done by fixing the level of forecast probability of the outcome, the point on the individual's propensity score, at which individuals will be classified as D+ or D-, i.e. forecast to have the outcome or not to have it.

In the narrow, specific model we fix this threshold to that of the observed, mean probability of the outcome. Thus, for example, if 5% of the population become teen parents then we take the 5% most at risk (i.e. the 5% with the highest propensity scores) and describe them as D+, the rest (95%) as D-. If we allow this 5% threshold to vary we will get different proportions of the population classified as D+, varying from 0 to 100%. Crucially, with each such threshold, different levels of true and false positives will result.

A very useful device in this context is the Receiver Operated Characteristics (ROC) Curve which graphs the proportion of true and false positives at different levels of the threshold level of risk at which individuals are said to be

“at risk,” or given a positive decision. An example ROC curve is reported in Figure 1, for the case of teenage pregnancy in the 1970 Cohort.

Figure 1: Receiver Operated Characteristics Curve for Teenage Pregnancy, (1970 Cohort, females, from distal information)



In Figure 1 the axes are labelled as “sensitivity” and “1-specificity.” These are other terms for what is described here in terms of true and false positives. Sensitivity denotes the ability to accurately identify those with the outcome of concern (i.e. the true positive rate.) Specificity is the ability to accurately assess those who do not have the outcome, i.e. the true negative rate. Therefore, 1-sensitivity is the false positive rate. Thus the axes of the ROC chart are the true and false positive rates under different thresholds for the decision rule of what level of risk for the outcome will mean that an individual will be forecast to have the outcome.

At high levels of the threshold (bottom left corner of Figure 1), at which few individuals are classified as D+, the true positive rate and false positive rate are both very low. These rates both increase as the threshold becomes more lenient, and more individuals become classified as D+. Assuming that the decision rule (estimation procedure) is better than chance, then as the threshold is reduced and more individuals are classified as D+, the rate at which the true positive proportion increases will be greater than the rate at which the false positive proportion increases. This can be seen in Figure 1 in which the ROC curve is above the 45° line.

Therefore, the area under the ROC curve is a good indicator of the forecast accuracy of the underlying decision rule across the whole distribution of possible levels of specificity.

In order to indicate the predictive power of specific measures we report:

Method 1a) the proportions of true and false positives when the at risk group is set to that of the observed, mean probability of the outcome;

Method 1b) the area under the ROC curve when specific measures are used to predict the adult outcomes.

Because the ROC curves take a very long time to calculate, we have not estimated ROC curves for all the outcomes under consideration, reporting only those for the multiple deprivation outcomes.

A few further methodological comments are in order, in relation to the first approach which is a restricted version of the more general second method.

Comparison across measures

Two technical points that should be noted are firstly that the true positive probability by itself does not offer a full comparison across measures and it is important also to take into account the false positive probability or the ratio of true to false positives. To see this consider a measure such as “looked after child” (LAC) which indicates that a child has been in the care of the Local Authority. The measure will have good discrimination in the sense that the risk of adult high cost/harm outcomes will for a wide range of reasons be relatively high for those who have been so looked after. The ratio of true to false positives will therefore be very high this measure. However, because only a small proportion of the population will be defined as LAC in this way the true positive proportion (i.e. the proportion of those with the outcome who are picked up by this measure) will be very low. In fact we find a negative correlation between the true positive rate (prevalence) and the ratio of true positive to false positives (discriminating accuracy) of -0.34, on average across all outcomes for the BCS, although this does vary by outcome. Measures that identify large proportions of the sample will tend to have high true positive rates and the ratio of true to false positives is then crucial in determining the true contribution of the measure to meaningful identification of those at risk.

The most information will come from measures that combine high precision with high prevalence but such measures are likely to be very rare in practice. It is likely that risk indices will work best when they draw on both i) precise measures with high identification but low prevalence, and, ii) measures with less precision but greater prevalence.

Discrete indicators of risk

Secondly, it should also be noted that it is not always possible to identify the proportions of true and false positive using single constructs that are discrete and a fixed rate of risk, i.e. method 1a. Sometimes there is no group identified

as having a level of risk above the given threshold as the measure groups too many members of the sample into identical risk groups. For example, consider the situation in which i) a measure has two values 0 and 1; ii) 80% of the sample obtain a score of 1 for the measure; iii) the score of 1 is associated with higher risk, and, iv) the probability of the high cost/harm outcome is, say, 20%. Under this scenario, 80% of the sample all have high risk and the attempt by the statistical programme to select those with the 20% highest level of risk will fail. Allowing the threshold to vary overcomes this difficulty to a great extent.

Differences by outcome

It is also important to recognise that the predictive power of different constructs will be different for different outcomes. Different readers may be interested in different specific outcomes. Yet with many outcomes to be considered a full assessment of all measures for all outcomes leads to the production of a very great amount of data that can be hard to interpret. To address this we undertake and present analysis mainly in terms of the multiple outcome indicators that combine outcomes into single indicators of multiple risk.

However, we also provide information in relation to specific outcomes.

Method 2: Classification trees

The second, entirely different, approach to assessing the contribution of constructs to the prediction of outcomes and hence in the identification of those at risk is the use of classification trees. These are the representation of an algorithmic method to determine which binary constructs (akin to right hand side variables in a classical regression model) best distinguish between those with and without particular subsequent outcomes (akin to left hand side or dependent variables).

We constraint all explanatory measures to be binary, although continuous measures can be recoded into a number of different binary measures. (For example, the variable age of the mother may be split into those above and below 30 years and further subdivided into sub-groups between 30 and 45 and above 45.). The algorithm assesses the discriminatory power of each construct and picks as its level 1 construct, the measure that best partials the sample into two groups, with one group relatively likely and the other unlikely to experience the outcome being predicted. In this sense the construct discriminates the likelihood of the outcome. This creates 2 nodes, individuals classified into 2 groups depending on their position in relation to the level 1 construct. The algorithm then considers for each node, which now second level construct best discriminates amongst those in each node into those most and least likely to have the outcome under consideration.

The process continues until the algorithm has resolved the sample into an optimal number of sub-groups, defined in terms of a set of interactions between binary variables, each sub-group having a defined probability of the outcome and where the sub-groups so identified are those that most effectively partial the sample in terms of these probabilities. The constructs

selected by the algorithm as providing such differentiation within the sample can be thought of as having maximal identification capability.

Approach adopted in this note

In the results section below we report for each outcome considered the set of measures that are highest in the classification tree down to three levels. We also report the range of likelihood of the outcome using these three sets of measures. For some of the outcomes with low prevalence among the population the classification tree analysis algorithm failed to produce an optimal number of sub-groups. In this case, the algorithm was re-estimated by age and the variables from the first three levels kept for a final estimation that contained these variables for all ages. Results are reported separately by gender.

4.2 Results 1: True and false positives

1970 Cohort

We begin by reporting results for the 1970 Cohort. Below are listed the ten most predictive constructs for the identification of individuals at risk in terms of the two measures of multiple outcomes in the 1970 Cohort. Prediction is assessed in terms of i) the proportion of true positives, and, ii) the ratio of true to false positives. Table 7 provides the detailed statistics.

Table 7 here

In terms of the proportion of true positives it can be seen that the most predictive measures include both distal and proximal constructs. For the two multiple outcome indicators these are:

More than 1 of 5 key outcomes:

- Mother's qualifications
- Father's qualifications
- Telephone present in home (as an income proxy)
- Father reads to child (assessed at age 5)
- Mother expects child to stay on in school (age 10)

More than 8 of 32 outcomes

- Family income
- Father reads to child (age 5)
- Mother expects child to stay on in school (age 10)
- Telephone present in home (income proxy)
- Mother's Occupational SEG

We have then conducted the exercise of entering only these 5 constructs with the highest level of true positive probability into regression equations to predict these outcomes.

4. Forecasting risk 2

For the first measure, we find that if only the 5 constructs above are used to predict having more than 1 of the 5 key outcomes, the true positive probability is 0.35, compared to a false positive probability of 0.12 (N=6875).

For the second measure, we find that if only the 5 constructs above are used to predict having more than 8 of the 32 outcomes, the true positive probability is 0.35, compared to a false positive probability of 0.14 (N=5113).

These figures are worth restating. Using only 5 constructs or pieces of information, assessed at ages 0, 5 or 10, we can identify by age 10, 35% of those who will experience 2 or more of 5 key high cost/harm outcomes at age 30, or, using a slightly different set of 5 constructs, 35% of those who will experience 9 or more from 30 diverse high cost/harm outcomes.

If instead of only considering the true positive probability for each construct, we choose those 5 measures with the highest ratios of true to false probabilities, the most predictive constructs are as follows:

More than 1 of 5 key outcomes:

- Truants at age 10
- Father has dismissive attitude to child (teacher report, age 10)
- Mother has dismissive attitude to child (teacher report, age 10)
- Housing density
- Soiling difficulties (age 10)

More than 8 of 32 outcomes

- Father has dismissive attitude to child (teacher report, age 10)
- Attends special school (age 10)
- Truants at age 10
- Mother has dismissive attitude to child (teacher report, age 10)
- Child communication skills (teacher rated, age 10)

Identifying high cost/harm individuals from only these 5 constructs we find for the first measure a true positive probability of 0.24, compared to a false positive probability of 0.09 (N=7381).

For the second measure, we find a true positive probability is 0.27, compared to a false positive probability of 0.11 (N=12,330).

1958 Cohort

For the 1958 Cohort we have conducted analysis to assess the predictive power of each of the constructs for identifying those likely to obtain at least 3 of the 6 high cost/harm outcomes assessed at age 23. In terms of the two measures of predictive power, the 5 most important constructs are:

Highest true positive probability

- Attendance in bottom 30% (age 16)
- Single index of teacher's rating child's progress at age 16

4. Forecasting risk 2

- Father reads books or technical journals
- Attendance in bottom 30% (age 15)
- Standardised maths score at age 11

Highest true to false positive probability ratio

- "Looked after" status at or before age 16
- Cannot read well enough for everyday needs at age 16
- "Mental backwardness" at age 11 (teacher report)
- Behavioural difficulties at age 11 (teacher report)
- Receiving special help for behaviour difficulties at age 11 (teacher report)

The specific values for these findings are reported in Table 8.

Table 8 here

These results again indicate the distinct importance of some very specific information that can be readily gained from teachers in data collection exercises that in technical terms could easily be standard practice. Most of the variables indicated as important are wither already assessed within administrative data collection, such as school achievement and attendance.

Again, we then use the two sets of measures above to identify those with 3 or more of the age 23 outcomes for the 1958 Cohort. The results for the measures with the highest true positive rates are 37% true positives and 5% false positives (N=2223). The results for the measures with the highest TP/FP ratios are 17% and 3% (N=5453).

ALSPAC

For the ALSPAC sample we have undertaken analysis on each of the two measures of multiple signals of risk, the first measure identifying the 12% of children who were struggling in Key stage 2 tests in English, maths and science (outcome 4). The second measure (outcome 7) identified the 20% of children who were experiencing 3 or more of 5 outcomes from low performance in any of 3 subjects at Key stage 2, high levels of anti-social behaviour and depression. Detailed results are reported in Table 9.

Table 9 here

The most predictive constructs were as follows:

Outcome 4. Low performance in Key stage 2 – High True Positive Rate

- KS1: Writing points score
- KS1: Reading points score
- KS1: Maths score

- Entry Assessment: Maths
- Entry Assessment: Writing

Outcome 4. Low performance in Key stage 2 – High TP/FP Ratio

- KS1: Maths score
- KS1: Writing points score
- KS1: Reading points score
- KS1: Spelling score
- Entry Assessment: Maths

Outcome 7. 3 or more of 5 high risk outcomes – High True Positive Rate

- KS1: Maths score
- KS1: Reading points score
- KS1: Writing points score
- Family type, 24 months
- Entry Assessment: Language

Outcome 7. 3 or more of 5 high risk outcomes – High TP/FP Ratio

- KS1: Reading points score
- KS1: Maths score
- KS1: Writing points score
- Entry Assessment: Maths
- KS1: Spelling points score

It is clear and striking from these results how predictive are Key stage 1 and entry assessment scores which tend to have not just high true positive rates but also high levels of precision in terms of the ratio of true to false positives.

When the set of measures with the highest level of true positive rates is used jointly for each multiple outcome measure the rates of true and false positives are 52% true positive and 5% false positive for outcome 4 (N=6041); and 60% and 10% for outcome 7 (N=3948). When the measures with the highest ratios of true to false positives are used (precision), the results are similar as the measures tend to be the same measures. True and false probabilities are 37% and 5% respectively for outcome 4 (N=7374), 53% and 10% for outcome 7 (N=4293).

4.3 Results 2: ROC curve areas

The second method for assessing the predictive utility of the constructs used in this study is the assessment of the area under the ROC curve when specific measures are used to predict outcomes. As stated above, we do not report these statistics for all outcomes but only for the multiple deprivation measures for the 1958 and 1970 Cohorts.

Results for 1958 Cohort

Table 10 reports the results for the 1958 Cohort, setting out the 10 most predictive constructs in terms of the area under the ROC curves, for outcomes 40 and 41, multiple deprivation at ages 23 and 42 respectively.

Table 10 here

The predictive variables are fairly similar for both versions of multiple deprivation, emphasising heavily:

- Teacher ratings of children's capability and progress at ages 7,11 and 16
- Performance in reading and maths tests at ages at 7,11 and 16
- Child behaviour and school-leaving plans

Results for 1970 Cohort

Results for the 1970 Cohort are presented in Table 11.

Table 11 here

For both indices of multiple deprivation, we find that there is a core set of measures which are particularly predictive, namely:

- Children's test scores in maths and reading at age10;
- Teacher-rated assessments of general behaviour and communication skills at age10;
- Attitudes, beliefs and behaviours at age 16.

On interesting finding we note is that at age 5, family social context measures are most predictive but that by age 10, the child's own achievements start to become most predictive. This is different for the 1958 Cohort, for whom we find that in the earliest measurement at age 7, the child achievement measures are already more predictive than the family context measures. This may be due to a historical shift between the cohorts, but is also likely to be due to important developmental shifts and the manifestation, for whatever reason, of greater stability in developmental pathways after age 5 and by age 7.

Another interesting point to note is that the precise measures that are most predictive does vary by outcome. There does appear to be a core set of measures that are important for multiple deprivation and indeed for most outcomes but for some specific outcomes of concern there may be one or two measures that become important. For example, the ten most predictive measures for the mental health problem outcome (outcome25) for the 1970 Cohort are as follows:

- Mother-rated conduct, age 16
- Self-esteem, age 16
- Smoker, age 16
- Anti-social behaviour age 16
- Locus of control ,age 16
- Mother's depression, age 5
- Truant, age 16

- Self-esteem, age 10
- Father approves of friends, age 16
- Wants to stay on at school , age 16

There is some degree of commonality in the indicators that are most predictive for this specific outcome and those of multiple deprivation. However, there are also indicators of risk that enter for this specific outcome only, in particular mother's depression in childhood becomes a more important indicator. This highlights the important point that what is a risk factor depends on what outcome is of interest. There are some risk factors that are important for many or most outcomes, others that play a more specific role.

4.4 Results 3: Classification trees

The third method for assessing the predictive utility of the constructs used in this study is the application of classification tree algorithms.

It is important to point out before discussing these results that all explanatory variables have been transformed into discrete variables. Many of the measures have important continuous information which makes selection of cut-offs slightly hazardous and arbitrary when done without detailed sensitivity analysis. Secondly, it can make the interpretation of specific measures difficult as it is not always obvious whether constructs are important as risk or promotive factors as this depends on whether the cut-off point is chosen to indicate a risk group or a group likely to be obtaining positive outcomes. In order to overcome this issue we have homogenised all binary variables to indicate risk and define risk following theoretical and empirical results in the literature. For example, household size is defined as a risk factor if the level is high (more than 6 members) whereas parental education is defined as a risk factor if the level is low (no qualifications).

A limitation is that the classification tree algorithm failed to converge for outcomes with a low prevalence in the population. For the BCS70 these outcomes are teenage parenthood, single parent, high criminality, workless household with children, absent child and addiction to drugs or alcohol. For these outcomes, we performed a two-stage analysis, first undertaking the analysis separately for each age, including all domains, and keeping the variables selected from the first three levels. In the second stage, we then re-estimated the classification tree algorithm including only the information from the selected predictors.

The results are striking in the extent to which they indicate that a small number of binary constructs can differentiate the sample into those with very high likely rates of high cost/harm outcomes and those whose level of risk is much lower.

1970 Cohort

Multiple outcomes

Results in terms of indicators of multiple outcomes for the 1970 Cohort are presented in Table 12.

Table 12 here

Here we have again undertaken analysis in relation to the aggregated measure across the 32 adult and adolescent outcomes. We do not yet have results in relation to the 5 key outcomes discussed above. Additionally, in order to further distinguish between the hardest to reach and those with high but less extreme apparent levels of need, we conducted analysis both in terms of the 10% with 9 or more of the 32 high cost/harm outcomes and the 20% with 6 or more of the 32 outcomes.

Results are presented for females and males separately and also grouped.

The interpretation of Table 12 is as follows. Consider, for example, the results for the first multiple outcomes indicator (outcome 33a in Table 12), for the full sample of both males and females. The variable selected by the algorithm to provide the greatest level of differentiation at the first level is “Child doesn't want to stay on (age 16).” Just using this single measure, the difference in predicted probability between those for whom this is and is not true is given by the range 0.26 : 0.47. in other words, 47% of those indicated by this measure are in the 20% with 6 or more of the 32 high cost/harm outcomes, compared to 26% of those who did want to stay on at age 16.

At lower levels of the classification tree, the difference between these probabilities for those in the highest and lowest risk groups continues to increase. At level 2 for this outcome are two measures “having a teenage mother” and vocabulary scores at age 5.” These each differentiate within one of the two nodes created by the first level binary variable. In other words the algorithm does not require one single variable to do the work of differentiation between all nodes at any given level. Rather, once a set of 2 nodes is created at the first level, each node is differentiated further at level 2 by whichever variable creates the greatest degree of differentiation.

The group created at level 2 by this interaction of variables that has the highest level of risk of having more than 6 of these outcomes has a probability of 67% of doing so, compared to 25% for those with the lowest risk. We do not report the probabilities for all the nodes created out of these two and three level interactions because that would overload the presentation.

At level 3 for this outcome, three further measures are useful. This increases the range in prediction to 17% and 77%. These results make apparent how a small number of measures can strongly differentiate between those with high and low risks of adult or adolescent cost/harm.

Separate outcomes

The results for all 32 separate outcomes in the BCS are reported in Table 13.

Table 13 here

For some outcomes, the algorithm fails to resolve due to the number of constructs entered, as discussed above. For these outcomes, we have rerun the analysis using measures at each age and keeping the first three levels of predictors. Then we re-estimate the classification tree with the sets of level 3 predictors. These results are reported in Table 14.

Table 14 here

We discuss results first for females, next for males.

For females, the most important distal variable that predicts outcomes is parental employment during childhood. This variable predicts teenage motherhood, homelessness, dissatisfaction with life, criminality, not voting in elections, low income, being on benefits, workless household with and without dependent children, and high levels of time not in employment, education, or training upto 30 (NEET30). Other important distal variables are mothers' and fathers' education, measured by lack of qualifications. These variables predict risk of depression, the risk of having mental health problems, having financial problems, not achieving Level 2 qualifications, not achieving Level 4 qualifications. Finally, parental SES is important in predicting overcrowding, obesity, not voting, not achieving Level 4 qualification, obesity and smoking

Measurements of child's school attainment and teacher's rating of child's school progress, cognitive and behavioural ability are very important predictors of risk in terms of the classification tree analysis. For example, maths and reading test scores predict 6 outcomes (depression, not achieving qualifications, racial intolerance, low self-efficacy mental health problems and NEET 30). Teacher ratings of child ability, both cognitive and behavioural, predict depression, psychiatric disturbance, not achieving Level 2 qualifications, racial intolerance, low self-efficacy, low income, and overcrowding, among other outcomes.

For males, the most important distal variable in predicting the risk of outcomes is parental education (measured either as parents not having qualifications or parents not staying on in post compulsory schooling). This variable enters at one of the first three levels for 11 of the 32 outcomes: depression, homelessness, victimisation, life dissatisfaction, low self-efficacy, being on benefits, workless household, absent child in the house, racial intolerance, addiction to drug or alcohol, and living in overcrowded housing. The next most important variable is parental SES, which predicts 7 outcomes: smoking,

depression, life dissatisfaction, workless household, mental health problems, other mental health issues, and NEET 30. Contrary to the predictive power for outcomes for females outcomes, parental unemployment only predicts one outcome for males.

Variables that measure cognitive, affective and behavioural development from birth to 16 enter the classification trees at high levels in predicting many of the outcomes in adulthood. Maths, reading and copying scores predict living in a one adult household, living in social housing, not achieving any qualifications and not achieving Level 2 qualification, living on benefits, living in a workless household with children, depression and living in overcrowded housing. Teachers' reports of children's abilities further predict living in overcrowded housing, victimisation, smoking, not achieving qualifications and not achieving Level 2 qualifications, social housing and low self-efficacy. In combination these two sets of variables predict 11 outcomes.

As mentioned above, some of the outcomes could only be investigated using a restricted set of constructs (teenage motherhood, single father, high criminality for men, workless household with children for male cohort members and addiction to drugs and alcohol for female cohort members). We find that distal variables such as parental education and SES, cognitive ability variables such as test scores and teacher's rating of child's cognitive and behavioural progress are among the most important predictors of the selected outcomes.

1958 Cohort

Separate outcomes

Results in terms of separate outcomes for the 1958 Cohort are presented in Table 15.

Table 15 here

We only focus on age 23 outcomes (unskilled or semi-skilled occupations, literacy or numeracy problems, having a child before the age of 20 and risk of depression). We discuss results first for females, next for males.

For females, the most important distal variable in predicting these outcomes are whether the mother smoked regularly before and during pregnancy, the mother's family size and the father's age. Measurements of cognitive and behaviour predict every single of these outcomes. For example, reading scores predict literacy, unskilled and semi-skilled occupations. Maths scores further predict teenage motherhood.

For males the most important distal variable is father's SES, which predicts two of the four outcomes. However, measures of cognitive ability and also important in predicting these outcomes, and sometimes at different levels. For example, in predicting literacy and numeracy problems reading scores at age

7, reading scores at age 16 are level 2 predictors and teacher's report or reading problems at 7 is a level 3 predictor. Vocabulary scores at age 11 is a level 1 predictor of depression, and teacher's rating of reading ability at age 7 is a level 1 predictor of unskilled and semi-skilled occupations.

4.5 Summary

The classification tree and regression analyses suggests that a small number of carefully chosen constructs can provide quite accurate identification of children at risk of adult and adolescent outcomes associated with high levels of personal, family and community cost and harm. The precise measures that might be collected as part of exercises in preventative activity will depend on the legal, policy and administrative contexts in which such measures are to be collected. Much will depend on:

- the level of local autonomy in what is to be collected,
- the ages at which data is to be collected,
- the sources of data collection,
- the level of dependence on administrative data,
- the extent of new types of data collection,
- the degree of linking between sources of measurement,
- the interventions that the identification is to be designed to support,
- the mechanisms for allocating interventions and support.

Many other factors will also play a role but this list should make clear that it is not possible or sensible to specify in advance a universal set of measures that would be most effective. Two general principles that emerge from this analysis of the longitudinal data may be worth stating:

- i. There is greater pay-off to increasing the ages, sources and domains of measurement than there is to obtaining more information from a single source;
- ii. Children move in and out of risk in terms of their own development and their levels of contextual risk. Therefore, it is important that the policy mechanisms allocating interventions and support to children and families are flexible and able to track and monitor levels of risk, not always intervening at the first sign of risk but equally able to provide early interventions that may reduce the need for more substantive and costly later interventions. This requires a considerable degree of local practitioner skill.

In terms of the specifics of particular measures the findings of this study indicate that Key stage scores and entry assessment data can be very predictive to later school success. This is well known but the extent to which this translates into identification of those at risk (true positives) has not been shown before. Much more could be done around these administratively collected data to build in personalised and targeted interventions that would reduce school drop-out and disengagement, with resulting reductions in high cost/harm outcomes.

4. Forecasting risk 2

Measures of family income and family background also add considerably to identification. These are routinely collected in tax and benefit data.

Teachers and health visitors can also provide very important information about early child development, parenting, family and child attitudes and behaviour that add greatly to identification. The extent to which this can be added to identification processes is not altogether clear. Some information such as developmental milestones and temperament are routinely collected by health visitors but this is not universal and there is currently no universal storage of this data and it is not linked to other health data maintained in the health system, let alone to education data.

Similarly the analysis here suggests that teachers are very well able to assess the development, behaviour and aspirations and expectation of children and the attitudes and behaviours of parents yet there is no universal mechanism for collection, linking or analyses of such perspectives.

However, information on truancy is routinely collected as is information on exclusions and attendance at special schools and most importantly, special educational needs status. These are very predictive constructs that could be linked. The Every Child Matters agenda also creates a context for ensuring that schools monitor levels of pupil behaviour. This would create a new form of administrative data that could with appropriate safeguards also be linked to this identification system. These safeguards would have to include measures to ensure that pupils were not stigmatised and that the developing record of behaviour was used to support improvement and development and not to create vicious cycles of worsening relationships with schools and teachers. Without such safeguards the use of such records may simply make the problems worse for both schools and pupils.

Finally, it may well be that development and reform of the SEN system to enable the collection and maintenance of more informative and precise data flow would enable integration and access of much of the required information on child, family and school context.

The SEN system is intended to determine the allocation of extra funding and intervention to support the learning of a relatively large proportion of the school population who have particular extra need. In this sense, the system rather mirrors the type of mechanism being proposed here in that there are many forms of SEN and many potential interventions to address these needs. A broadening of the general category of SEN together with greater precision around the types of provision and funding associated with SEN or a new form of status and the establishment of systems to enable clearer monitoring and assessment of developing risk may facilitate the type of data gathering and local monitoring being proposed here.

A broadening of the range of interventions that are triggered by SEN status would also enable the formation of a system that i) identifies, ii) tracks and monitors, and, iii) supports and protects children and families so identified as having levels of developmental risk that are likely to lead to subsequent

4. Forecasting risk 2

problems and the further inter-generational and social transmission of the difficulties.

5. Multiple outcomes – The Fergusson model

In this section we report analysis following work by Fergusson *et al.*, (1994)³, in New Zealand. Their analysis explores how the probability of individuals in different risk groups being defined as high cost changes depending on the threshold set for such a definition.

There are two essential elements to these models; i) the level of risk in childhood, and, ii) the degree of cost of outcomes in adulthood. Both of these elements vary between individuals and, depending on the thresholds set for each, individuals will differentially be classified as “at risk” in childhood, and “high cost/harm” in adulthood.

For example, we can count for each individual in the 1970 Cohort, the number of the 32 adult outcomes that they experienced, creating a variable that ranges from 0 to 32. The cut-off point at which individuals are said to be high cost/harm is fairly arbitrary without substantial additional information about the social cost of each outcome, and even then it is a matter of judgement.

Similarly, for each of the 32 outcomes, we can use the methods of previous sections to assess the level of risk of each individual for each outcome. Again, the level of risk at which individuals are said to be “at risk” can be a fairly arbitrary judgement; should one take the 5% most at risk, the 10% most at risk and so on.

In the analysis in this section, we allow each of these thresholds to vary and show how as greater levels of adult deprivation are assessed the level of risk of those at most risk in childhood grows substantially relative to that of those at less risk in childhood. In other words, although most children will grow up to experience one or a few high cost/harm outcomes as defined here, very few will experience many such outcomes and those who are the most vulnerable situations in childhood are particularly at risk of multiple deprivation in adulthood.

We conduct this analysis for the 1970 Cohort.

Fergusson *et al.*, set two levels of childhood risk: the 5% most at risk and the 50% least at risk. They consider two thresholds for defining high harm in terms of the number of outcomes. We have reproduced this analysis but allowed the threshold of high cost to vary (Figure 2) and the levels of risk to vary (Figure 3).

As the threshold increases the probability of the outcome for the high risk group relative to that for the low risk group, increases exponentially.

This is shown in Figure 2 for the 1970 Cohort.

³ Fergusson, D., Horwood, J., and Linskey, M., (1994). “The childhoods of multiple problem adolescents: A 15 year longitudinal study.” *Journal of Child Psychology and Psychiatry*, Vol 35, No. 6, pp1123-1140.

Figure 2 here

The horizontal axis shows the minimum number of the 32 outcomes experienced by each individual. The vertical axis plots the proportion of individuals experiencing at least this number of the 32 high cost/harm outcomes. The three lines plot these probabilities for three groups: i) for the 50% at least risk (a large low risk group), ii) the mean level of risk, and, iii) for the 5% most at risk. For these estimates risk is defined on the basis of distal and proximal features of the family up to age 10. No information about the child is used, only data on the family context.

The three lines all start at 1 on the vertical axis because by definition the probability of experiencing at least 0 of the 32 outcomes is 1. As we move along the horizontal axis, the number of outcomes experienced goes down on average and for the high and low risk groups. For example, the probability of experiencing at least 5 of the 32 outcomes is 27% for the lowest risk group, 37% on average and 86% for those defined as at highest risk based on their childhood family circumstances. The ratio of risk between the low risk group and the high risk group is 4.2 ($0.86/0.27$).

At higher levels of adult deprivation, this ratio increases. Thus, the probabilities of experiencing 10 or more of the 32 outcomes are 1% for the low risk group, 12% on average and 51% for the high risk group. The ratio at this point is 46.1. In fact, this ratio increases to a maximum of 313 for the probabilities of experiencing 12 or more outcomes, as none of the low risk group experienced higher levels of multiple adult deprivation and so the ratio tends to infinity.

In part, this is just a property of small numbers. The absolute difference between risk groups is larger for the lower threshold but in relative terms, grows exponentially as a higher threshold is set for the definition of high cost.

However, this also highlights the very strong relationship between high childhood risk and multiple adult deprivation. Although none of the 50% at lowest risk experienced 12 or more of these outcomes, 49% of those in the highest risk group did, a very stark difference.

Another slice through this is shown in Figure 3, which holds constant the level of adult deprivation and allows the level of childhood risk to vary.

Figure 3 here

In Figure 3, the horizontal axis reports the level of risk for the individual based on their childhood family circumstances, up to age 10. Children are grouped into 20 different levels of risk. Those at percentile 97.5 for example, in Figure

5. Multiple deprivation

3 are those in the highest 5% of risk. Those at percentile 2.5 are the 5% with the lowest level of risk.

The 6 lines of Figure 3 indicate the probability of multiple adult deprivation for each of these 20 groups for 6 different definitions of multiple deprivation. The lowest line is for the case of multiple adult deprivation defined as experience of 15 or more of the 32 outcomes. For those in the 1970 Cohort, only those in the riskiest childhood circumstances ended up at age 30 with this very high level of multiple adult deprivation, in fact only those in the highest decile of risk (with a slight exception for group 15 at percentile 72.5). However for the 5% most at risk in terms of childhood circumstance, 24% experienced 15 or more of the 32 outcomes.

Other lines in Figure 3 report the probability of multiple adult deprivation, when deprivation is defined in less extreme terms. However, even with adult deprivation defined in terms of less extreme situations, there is still a very strong increase in disadvantage for those with the highest levels of childhood risk. For example, for the 5% most at risk in terms of childhood circumstance, the probability of experiencing 6 or more of the 32 adult outcomes is 83%, compared with 48% for those whose risk is between the 85th and the 90th percentiles and 20% around the mean. This non-linearity indicates the substantial level of extra risk associated with extreme disadvantage and deprivation in childhood.

6. Models of the costs and benefits of intervention

6.1 Overview and summary of section

Thus far we have established that it is possible to make good predictions about which children are likely to experience high cost, high harm outcomes as adults. In this section we address the question of whether and under what assumptions, these levels of forecasting accuracy suggest that policy response to the signals provided would be cost effective.

It will never be possible to know with 100% accuracy which children will experience adult deprivation and there will always be false positives (those wrongly identified) and false negatives (missed cases of need). The question we address in this section is what reasonable and empirical levels of accuracy imply in terms of the costs and benefits of intervention.

We set out a framework for assessing the cost-effectiveness of interventions and for evaluating the relative cost-benefit of different levels of targeting of interventions. The framework can be applied to evaluate the cost-effectiveness at different levels of targeting of any given interventions for which there is good evidence on its effectiveness. The purpose of the analysis is to assess the value of targeting interventions using the identification potential set out in previous sections and to demonstrate the conditions under which this level of identification will add to the cost-effectiveness of early intervention.

An example intervention

In this section we set out the relative costs and benefits of an example intervention under 3 different allocation rules:

1. Assignment to all of an eligible population
2. Random assignment to a fixed proportion of the eligible population
3. Non-random allocation to a targeted group with a special case, fixed rate of intervention.

This analysis makes transparent the assumptions under which early intervention will be cost-effective and indicates that under most reasonable assumptions, early preventative action is feasible. Future work should extend the analysis to the case of an intervention with evidence-informed, optimised targeting (ROC analysis).

In order to clarify the cost-effectiveness of early intervention given the empirical capability to identify those who may benefit, we consider the outcome of teenage parenthood, as an example. This is a difficult outcome to address as for many on the political left and right it raises difficult moral and policy questions. It is treated here as an outcome in adolescence/early adulthood that may cause personal and social difficulty. We do not at all suggest that all cases of teenage parenthood are necessarily costly, negative or cause for public reaction. However, many young women who do have children during their teenage years suffer educational and economic and social hardship as a result. The intervention considered here involves teenagers spending time looking after toddlers and so gaining deeper

6. Cost-benefit analysis

appreciation of the realities of parenthood. This is often very different to what they had expected. The evidence suggests that this substantially reduces the likelihood of early parenthood through impacts on the choices and behaviours of these young people.

We develop a model of the costs and benefits of such an intervention, working from the following evidence-informed assumptions (internal DfES costings):

Observed parameters for intervention cost-effectiveness

1. Risk in the eligible population = 5%;
2. Social cost of the outcome = £60K
3. Cost of intervention = £1K
4. Effectiveness = 80%

The policy has been found to have an effectiveness rate of 80%. This is a very high rate and may well not be achievable in a wider programme. The robustness of the evaluation evidence may be questionable. However, we show below that the effectiveness of the intervention would only have to be 4% to demonstrate cost-effectiveness.

Given these parameters and the cost benefit model set out below, together with the observed (lower bound) capability to identify those at risk, we find that the parameters required for a favourable benefit/cost ratio are:

Requirements for cost-effectiveness

1. Social cost of the outcome > £3.3K
2. Cost of intervention < £18K
3. Effectiveness > 4%

These values can be compared to those observed, suggesting that under very reasonable assumptions, intervention can be expected to be highly cost-effective.

6.2 Calibration of cost-effectiveness

Notation

As some degree of mathematical modelling is required, we must first set out some notation:

$C(T+)$ = The social cost of the outcome (T =truth)

$P(T+)$ = The unconditional probability of the outcome; $P(T+) + P(T-) = 1$

$C(D+)$ = The cost of the intervention (D =outcome decision or prediction)

f = the effectiveness of the intervention; $0 \leq f \leq 1$. We may also be interested in conditional values of f , i.e. for different target groups and one may wish to target groups with high expected values of f for the intervention. It is likely that f is a function of readiness to participate and also that estimates of f from non-

6. Cost-benefit analysis

experimental evaluations may over-estimate f , particularly if they ignore the issue of choice of participation. Heterogeneity in f is an important issue that can be built into more sophisticated models. In more sophisticated models one should also allow for f to vary with the level of targeting of the intervention, being lower for more universal interventions due to the problem of scalability and roll-out.

$P(D+)$ = the level of targeting of the intervention, where $P(D+)=1$ denotes a universal intervention. If 5% are targeted, $P(D+)=0.05$.

An important element of the method and notation is the notions of true and false positives, based on the relationship of the predicted and actual outcomes. These were indicated in Table 1 above, where the terms are defined in terms of the accuracy of the predicted outcome relative to truth, i.e. as conditional expectations of the form $P(D | T)$. We reproduce Table 1 here.

Table 1: A tabular description of forecast accuracy for binary outcome.

		Truth		
		Positive	Negative	
Decision (Predicted outcome)	Positive	True positive a	False positive b	a + b
	Negative	False negative c	True negative d	c + d
		a + c	b + d	

As Table 1 highlights, the range of possibilities are as follows:

- i) True negatives - accurate identification of those not requiring extra support;
- ii) False positives - deadweight intervention;
- iii) False negatives - missed intervention needs resulting in failure to intervene;
- iv) True positives - accurately targeted interventions.

Two particular probabilities summarise the range of possibilities:

$P(TP) = P(D+ | T+) =$ the true positive probability $= a / (a+c)$.

$P(FP) = P(D+ | T-) =$ the false positive probability $= b / (b+d)$

These two probabilities summarise the range of possibilities as the other two probabilities are the complements of these, (e.g. $P(FN) = 1-P(TP)$ ⁴ and $P(TN) = 1-(FP)$)

Note, too, that the marginal distributions can also be expressed in terms of the prior notation $P(T)$ and $P(D)$. Thus, $(a+c)/(a+b+c+d) = P(T+)$ and $(a+b)/(a+b+c+d) = P(D+)$ Of course, $P(T+) + P(T-) = P(D+) + P(D-) = 1$

The relationship between joint, conditional and marginal probabilities is given by Bayes Theorem;

⁴ $P(FN) = P(D- | T+) = P(D-, T+) / P(T+) = c/(a+c) = 1 - \{a/(a+c)\} = 1-P(TP)$

6. Cost-benefit analysis

$$P(T+, D+) = P(D+ | T+) * P(T+) = P(TP) * P(T+)$$

As well as these probabilities, we also have important costs and benefits associated with each.

$C(T-, D+)$ = The cost of a positive decision and hence intervention for those who would not experience the outcome (deadweight +stigmatisation costs)

$C(T+, D-)$ = The cost of a negative decision and hence failure to intervene for those who would experience the outcome

$B(T+, D+)$ = The net benefit of a positive decision and intervention for those who would experience the outcome (net of intervention and stigmatisation costs)

$B(T-, D-)$ = The net benefit of a negative decision and absence of intervention for those who would not experience the outcome.

We can operationalise a general model by stating that an intervention will be cost-effective when its benefits exceed its costs. It may be useful to specify this either as

$$C/B < 1$$

or

$$B - C > 0$$

In our first general statement of these expectations we ignore $C(T+, D-)$, the costs of missed at risk cases and we also set $B(T-, D-) = 0$. We have,

$$\begin{aligned} B(T+, D+) &= P(D+, T+) * f * C(T+) - P(D+, T+) * C(D+) \\ &= P(TP) * P(T+) * f * C(T+) - P(TP) * P(T+) * C(D+) \\ &= P(TP) * P(T+) \{f * C(T+) - C(D+)\} \end{aligned}$$

Similarly,

$$\begin{aligned} C(T-, D+) &= P(D-, T+) * C(D+) \\ &= P(FP) * P(T-) * C(D+) \end{aligned}$$

Thus,

$$B - C = P(TP) * P(T+) \{f * C(T+) - C(D+)\} - P(FP) * P(T-) * C(D+) \quad (1)$$

Moreover,

$$\frac{C}{B} = \frac{P(FP) * P(T-) * C(D+)}{P(TP) * P(T+) \{f * C(T+) - C(D+)\}}$$

In terms of P(FP) and P(TP), the requirement for cost-effectiveness is given by:

$$\frac{P(TP)}{P(FP)} > \frac{P(T-)}{P(T+)} * \frac{C(D+)}{\{f * C(T+) - C(D+)\}} \quad (2)$$

Given this set up of the basic framework we now describe the costs of benefits of a particular intervention under 3 scenarios about the allocation of treatment. We consider the example of the outcome teenage pregnancy. Estimates from the PMSU of social cost are £60K, estimates of the effectiveness of a prior intervention is given as 80% success rate at a cost per intervention of £1K:

<p>Example: Teenage pregnancy</p> <p>Risk in the eligible population = 5%; P(T+)=0.05 Social cost of the outcome = C(T+) = £60K Cost of intervention = C(D+) = £1K Effectiveness = f = 0.8</p>
--

Model 1: Cost-effectiveness under universal assignment

In this straight-forward first case, the universality of the programme means that allocation bears no relationship to the individual's probability of the outcome, i.e.:

$$P(FP) = P(TP) = P(D+) = 1$$

Thus, equation (1) is simplified to:

$$\begin{aligned} B - C &= P(T+) \{f * C(T+) - C(D+)\} - P(T-) * C(D+) \\ &= P(T+) \{f * C(T+)\} - C(D+) \{P(T+) + (1-P(T+))\} \\ &= P(T+) * f * C(T+) - C(D+) \end{aligned} \quad (3)$$

In this model B > C if:

$$f * C(T+) * P(T+) > C(D+)$$

Thus in this model, a programme is more likely to be beneficial,

6. Cost-benefit analysis

- i) the more effective it is;
- ii) the greater the social cost of the costly outcome, and;
- iii) the greater the prevalence of the costly outcome.
- iv) the lower the intervention costs

This simple model:

- i) assumes no costs of mis-assigned intervention other than deadweight, i.e. no stigmatisation costs;
- ii) ignores the possibility of using prior information to increase effectiveness by targeting the intervention and;
- iii) does not recognise the social costs of non-intervention where it may be appropriate, i.e. $C(T+, D-)$

Nonetheless, the basic formula is useful. In terms of our example:

$$\begin{aligned} B-C &= f * C(T+) * P(T+) - C(D+) \\ &= (0.8 * £60K * 0.05) - £1K \\ &= £1.4K \end{aligned}$$

This model can be calibrated in the dimensions of any of the variables. Holding the costs and risks fixed, we can show, for example, that the required level of intervention effectiveness for the programme to provide a positive net benefit is given by

$B - C > 0$ if ;

$$f > \frac{C(D+)}{C(T+) * P(T+)}$$

$$> 1/3$$

Or assuming that the intervention costs and efficiency are given, we can consider the level of social cost at which the intervention becomes beneficial:

$$B - C > 0 \text{ if } C(T+) > £25K$$

Equivalently, *ceteris paribus* in terms of the cost of the intervention:

$$B - C > 0 \text{ if } C(D+) < £2.4K$$

Model 2: Random assignment to a fixed proportion

In this second case, we maintain random allocation of the programme but reduce the rate of intervention. Thus,

$$P(TP) = P(FP) = P(D+)$$

Equation (1) can be simplified to:

6. Cost-benefit analysis

$$\begin{aligned} B - C &= P(D+) * P(T+) \{f * C(T+) - C(D+)\} - P(D+) * P(T-) * C(D+) \\ &= P(D+)\{f * P(T+) * C(T+) - C(D+)\} \end{aligned} \quad (4)$$

Note that under non-universal but random assignment, the level of rationing of the intervention influences the magnitude of the overall net benefit but not its sign. As in model 1, cost-effectiveness requires:

$$f * C(T+) * P(T+) > C(D+)$$

but the level of average benefit is factored down by the amount $P(D+)$. If we restricted the intervention to a fixed proportion of the eligible population where the proportion was given by the average population risk ($P(D+) = P(T+) = 0.05$), then the average net benefit would be $0.05 * \text{£}1.4\text{K} = \text{£}70$.

Model 3: Non-random, targeted allocation to a fixed proportion

We now relax the assumption that $P(D+)$ is independent of $P(T+)$ but maintain the allocation threshold such that the programme is given only to those most at risk and where it is rationed such that, as in model 2, the proportion of those who receive the intervention is the same as require it, although of course, because of targeting failures, not all those who receive it will require it and not all of those who require it will receive it, unless $P(FP) = 0$ and $P(TP)=1$.

This model introduces programme targeting into the model and emphasises the issue of the decision rule or identification of those at risk while retaining an important simplifying assumption

Previously, in model 2, the decision rule was given by the level of targeting $P(D+)$, which was allocated randomly across those with negative or positive predicted outcomes. Thus in model 2, $P(FP) = P(TP) = P(D+)$. Now, we allow these probabilities to diverge, increasing efficiency by reducing the probability of deadweight intervention and increasing the likelihood of accurately targeted intervention. The net intervention benefit now depends on $P(TP)$ and $P(FP)$ as well as on the other parameters of the model in equations 1 and 2.

Increasing the accuracy of prediction will increase the likelihood of positive net benefit.

In model 3, we fix the threshold for intervention in such a way as to set $P(D+) = P(T+)$. This is a simplifying restriction because it also removes the necessity of introducing $P(FP)$ into the model because under this restriction, $P(TP)$ is sufficient to summarise the identification possibilities. It is equivalent to the restriction $b=c$ in Table 1.

$$\text{Since } P(T+, D+) + P(T+, D-) = P(T+)$$

$$\text{and } P(T+, D+) + P(T-, D+) = P(D+)$$

6. Cost-benefit analysis

$P(D+) = P(T+)$ implies the equality of the off-diagonal probabilities:

$$P(T+, D-) = P(T-, D+)$$

This means that $P(FP)$ can be derived from $P(TP)$ since,

$$\begin{aligned} P(T+, D-) &= P(T+) - P(T+, D+) \\ &= P(T+) - P(D+ | T+) * P(T+) \\ &= P(T+) [1 - P(TP)] \end{aligned}$$

and

$$\begin{aligned} P(T-, D+) &= P(D+ | T-) P(T-) \\ &= P(T-) * P(FP) \end{aligned}$$

Therefore, by the equality of the off-diagonals,

$$P(FP) = \frac{P(T+)}{P(T-)} * [1 - P(TP)] \quad (5)$$

In this special case, and using equation 5, equation 1 can be simplified to:

$$\begin{aligned} B - C &= P(TP) * P(T+) \{f * C(T+) - C(D+)\} - P(FP) * P(T-) * C(D+) \\ &= P(TP) * P(T+) \{f * C(T+) - C(D+)\} - [1 - P(TP)] * P(T+) * C(D+) \\ &= P(T+) \{P(TP) f C(T+) - C(D+)\} \end{aligned} \quad (6)$$

Therefore, to support intervention we require:

$$P(TP) * f * C(T+) > C(D+)$$

Using the parameters of our example, cost effectiveness requires

$$P(TP) > 0.02$$

This is a very lax threshold because of the very high balance of savings to costs of this intervention for this outcome. We only have to accurately identify with our decision rule, 2% of those genuinely in need to make the intervention worthwhile in these terms.

A central question for this project is the empirical assessment of the likely ratio of false to true positives, drawing on longitudinal survey data. Without going into detail here, estimates suggest that for the 1970 cohort, drawing on data from age 0 to age 10 for female cohort members only, we obtain a true positive rate of 0.38 and a false positive rate of 0.014. These results are

6. Cost-benefit analysis

based on a fixed threshold model with a decision rule that predicts the outcome of teenage motherhood for those females in the sample with the highest probabilities of that outcome based on a logistic regression analysis and where the cut-off for a positive prediction matches the observed probability of the outcome, i.e. 5%. In other words if we take the 5% of females that are suggested by the regression analysis to be most at risk and compare these probability-based decision outcomes to the observed outcome in the data we find that 38% of those who became teenage mothers were predicted to become so ($P(TP) = P(D+ | T+) = 0.38$) and that 1.4% of those who did not become teenage mothers were erroneously predicted to become teenage mothers.

Thus, empirical analysis suggests that a feasible level of $P(TP)$ is 0.38, although in reality we could probably do much better than that. In any case, this is much higher than the 0.02 that the parameters of example 1 suggest is the minimum if this intervention is to provide positive net benefit.

The calculated average net benefit for an intervention targeted on 5% of the eligible population selected on the basis of the empirical decision rule would be as given by equation 6:

$$\begin{aligned} B - C &= P(T+) \{ (P(TP) * f * C(T+) - C(D+)) \} \\ &= \text{£}862 \end{aligned}$$

This is much higher than the £70 we achieved by random selection of 5% of the eligible population in model 2. Model 3 is more effective because it targets a higher proportion of those in need, i.e. has a higher hit rate and its relative efficacy increases with $P(TP)$. For the parameters of this example, the difference in expected value is given by:

$$V^3 - V^2 = \text{£}2.4K * \{P(TP) - 0.05\}$$

It should also be noted that we have not yet included in the estimation the continued social cost of those who needed the intervention but didn't get it. Here, note that the cost of this must be included in the decision about which decision rule is preferable and is given by:

$$\begin{aligned} C(D-, T+) &= P(D-, T+) * C(T+) \\ &= P(D- | T+) * P(T+) * C(T+) \end{aligned}$$

In models 2 and 3, this is given by

$$C_2(D-, T+) = 0.95 * 0.05 * C(T+)$$

$$C_3(D-, T+) = 0.62 * 0.05 * C(T+)$$

Thus, model 2 has an average net benefit that is lower than that of model 3 by £792, and also leaves an average per person net remaining social cost of £2.85K compared to £1.86K for model 3.

6. Cost-benefit analysis

To indicate the increased effectiveness of the targeted intervention of model 3, we can also calibrate the other parameters as before using the formula of equation 6:

$$B > C \text{ if } (P(TP) * f * C(T+)) > C(D+)$$

Intervention effectiveness requires *ceteris paribus*:

$f > 0.044$ (whereas in reality, $f = 0.8$)

$C(T+) > \text{£}3.29$ (reality = $\text{£}60\text{K}$)

$C(D+) < \text{£}18.24\text{K}$ (reality = $\text{£}1\text{K}$)

These empirically based thresholds are much less stringent than those resulting from the random assignment in models 1 or 2. This is summarised in Table 16.

Table 16: Criteria for net positive benefit of teen pregnancy intervention

	Model 1,2: Random assignment	Model 3: Targeted by prior info, fixed rate targeting	Estimated truth
Minimum effectiveness, f	0.33	0.044	0.8
Minimum cost of social outcome, C(T+)	£25K	£3.28K	£60K
Maximum cost of treatment, C(D+)	£2.4K	£18.24K	£1K

This, it can be seen that by using prior information on risk of the outcome we can significantly improve the cost effectiveness of intervention.

Intervention with evidence-informed, optimised targeting (ROC analysis)

We have not thus far modelled or estimated the continued cost of the social outcome for those who required but did not receive the intervention or for whom the intervention did not work. We have also chosen an arbitrary level of targeting that matched the average population level of risk of the outcome but did not take into account the risk for those below the threshold.

These issues can be taken into account in models which vary the threshold of the decision rule in terms of the level of targeting while still taking prior information into account. Since such models take into account the social cost of those who do not receive an intervention that was required, $C(D-, T+)$, this more general model enables an assessment of the optimal level of targeting as well as indicating the narrow net benefit of the intervention.

These models are based on analysis of Receiver Operated Characteristics (ROC) curves, which were described in more detail in section 4.3.

It has been shown⁵ that in cost-benefit terms, the optimal level of targeting of an intervention is that level at which the slope of the ROC curve, S^* , is given by:

$$S^* = \frac{P(T-)}{P(T+)} * \frac{B(T-, D-) + C(T-, D+)}{B(T+, D+) + C(T+, D-)}$$

Thus assessment of the costs and benefits of intervention and failure to intervene, together with knowledge of the prevalence of the outcome and of the identification possibilities summarised by the ROC curve, can inform decision-making about the optimal level of targeting.

6.3 Summary

In this section we have set out a standard framework for assessment of the cost-effectiveness of intervention and shown how the empirically based, risk forecast accuracy analysis of previous sections can inform decision-making about the targeting and effectiveness of interventions.

It is not the case that 50% accuracy in identification of those who will experience adult outcomes is either necessary or sufficient to support a policy of targeted preventative action to reduce adult social exclusion and to protect vulnerable children. The precise level of accuracy required depends on the levels of true and false positives within a context of assessment of the social cost of failure to intervene and of the effectiveness and cost of possible interventions.

In this section we have shown that if effective policies are available than can reduce the likelihood of high cost/harm outcomes without excessive stigmatisation or other side-effects, then it is likely to be the case that it is possible to identify enough of those who may be at risk with sufficient accuracy to make such intervention cost-effective.

We have not shown in this report that any specific intervention is appropriate in relation to any specific outcome or set of risks or undertaken any analysis of stigmatisation or other side-effects. These issues would have to be addressed in any real world application of this framework. However, we have indicated that our ability to identify those at risk should not be a barrier to preventative action.

⁵ Swets, J. A., Dawes, R. M. and Monahan, J. (2000). "Psychological Science Can Improve Diagnostic Decisions." Psychological Sciences in the Public Interest (a supplement to Psychological Science), 1(1)"

Table 2: Sources and ages of data collection in the NCDS, BCS and ALSPAC

	Source	Age 0	Age 1	Age 2	Age 3	Age 4	Age 5	Age 6	Age 7	Age 8	Age 9	Age 10	Age 11	Age 12	Age 13	Age 14	Age 15	Age 16
1958 cohort	Teacher								√				√					√
	Parent	√							√				√					√
	Clinic/Medical												√					√
	Child												√					√
1970 cohort	Teacher											√						√
	Parent	√					√					√						
	Clinic/Medical											√						
	Child											√						√
ALSPAC	Teacher								√									
	Parent	√	√	√	√			√		√	√	√	√	√				
	Clinic/Medical		√	√		√	√		√	√								
	Child									√	√	√						

Table 3.1: Summary statistics for predictor variables, 1958 Cohort (NCDS)

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
<i>Measures at age 0</i>						
<u>Distal family factors</u>						
fageleft_1	father left schooling at minimum leaving age	12991	0.78	0.41	0	1
fageleft_2	father stays on 1 year post minimum schooling	12991	0.03	0.18	0	1
fageleft_3	father stay on 2 years post minimum schooling	12991	0.08	0.26	0	1
fageleft_4	father stays on 3 or more years post minimum schooling	12991	0.11	0.32	0	1
fses_1	father ses 1&2	16459	0.17	0.38	0	1
fses_2	father ses 3	16459	0.61	0.49	0	1
fses_3	father ses 4	16459	0.12	0.33	0	1
fses_4	father ses 5	16459	0.10	0.30	0	1
maage20	mother age in last birthday <= 20	17402	0.10	0.30	0	1
mabthord	birth order of the mother	17231	2.88	2.16	1	10
mafsize	family size of the mother	17237	4.43	2.58	1	10
mageleft_1	mother left schooling at minimum leaving age	14350	0.84	0.36	0	1
mageleft_2	mother stays on 1 year post minimum schooling	14350	0.04	0.20	0	1
mageleft_3	mother stays on 2 years post minimum schooling	14350	0.05	0.22	0	1
mageleft_4	mother stays on 3 or more years post minimum schooling	14350	0.06	0.24	0	1
paage	husband's age in years, 1958	16736	30.57	6.45	16	78
pproom0	persons per room in 1958	16921	1.55	1.00	1	6
ppses_1	mother's father ses 1	15814	0.03	0.16	0	1
ppses_2	mother's father ses 2	15814	0.14	0.35	0	1
ppses_3	mother's father ses 3	15814	0.47	0.50	0	1
ppses_4	mother's father ses 4	15814	0.14	0.34	0	1
ppses_5	mother's father ses 5	15814	0.14	0.34	0	1
ppses_6	mother's father ses nopa	15814	0.09	0.28	0	1
ppskilled	mother's father skilled worker	14288	0.38	0.49	0	1
r1	risk factor: father sc 4 or 5, sick or unemployed at birth	16513	0.22	0.41	0	1
ffses_1	father's ses 1&2	11830	0.16	0.37	0	1
ffses_2	father's ses 3	11830	0.50	0.50	0	1
ffses_3	father's ses 4	11830	0.25	0.43	0	1
ffses_4	father's ses 5	11830	0.09	0.29	0	1
<u>Proximal family processes</u>						
smokeb4	mother smoked in 12 months prior to pregnancy	17351	0.41	0.49	0	1
smokeduring	mother smoked during pregnancy	17193	0.14	0.35	0	1
<u>Neighbourhood context</u>						
<u>School context</u>						

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
<u>Cognitive development</u>						
<u>Affective and behavioural development</u>						
<u>Health</u>						
qnweight_1	birthweight in 1st quintil	17343	0.20	0.40	0	1
qnweight_2	birthweight in 2nd quintil	17343	0.20	0.40	0	1
qnweight_3	birthweight in 3rd quintil	17343	0.21	0.41	0	1
qnweight_4	birthweight in 4th quintil	17343	0.20	0.40	0	1
qnweight_5	birthweight in 5th quintil	17343	0.19	0.39	0	1
Measures at age 7						
<u>Distal family factors</u>						
care7	leaving in care at age 7	14669	0.02	0.15	0	1
fsize_1	family size 1 or 2 sibs	13151	0.43	0.50	0	1
fsize_2	family size 3 or 4 sibs	13151	0.42	0.49	0	1
fsize_3	family size 5 to 12 sibs	13151	0.15	0.36	0	1
fsize_4	family size 13 sibs	13151	0.00	0.02	0	1
fsize_5	family size 14 sibs	13151	0.00	0.01	0	1
hardship7	financial hardship at age 7	14589	0.08	0.26	0	1
moves	number of times family moves since child's birth	14408	1.21	1.50	0	22
nchildren7	number children under 21 at age 7	14590	3.09	1.62	1	14
noinlav	no indoor lavatory or shared it	14497	0.16	0.37	0	1
nopa7	no father at age 7	14703	0.03	0.17	0	1
noteng	mother quality of spoken english	14606	0.01	0.11	0	1
pproom7	persons per room at age7	16921	1.55	1.00	1	6
noma7	no mother at age 7	14705	0.03	0.16	0	1
pses_1	Mother's SES 1 or 2 at birth	17214	0.04	0.21	0	1
pses_2	Mother's SES 3 at birth	17214	0.21	0.41	0	1
pses_3	Mother's SES 4 at birth	17214	0.10	0.30	0	1
pses_4	Mother's SES 5 at birth	17214	0.03	0.16	0	1
pses_5	Mother's SES 'no job' at birth	17214	0.61	0.49	0	1
region7_1	living in northwest	18558	0.12	0.33	0	1
region7_2	living north	18558	0.07	0.25	0	1
region7_3	living in yorkshire	18558	0.08	0.27	0	1
region7_4	living in northmidlands	18558	0.07	0.26	0	1
region7_5	living in east	18558	0.07	0.25	0	1
region7_6	living in london & southeast	18558	0.19	0.39	0	1
region7_7	living south	18558	0.05	0.22	0	1
region7_8	living southwest	18558	0.05	0.22	0	1
region7_9	living midlands	18558	0.09	0.28	0	1
region7_10	living in wales	18558	0.05	0.22	0	1

Table 3.1: p2 of 10

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
region7_11	living in scotland	18558	0.11	0.31	0	1
region7_12	not in p??	18558	0.06	0.24	0	1
bthord	birth order	14270	2.28	1.44	1	7
ndiffs1	housing difficulty at age7	14432	0.12	0.33	0	1
<u>Proximal family processes</u>						
cmstayon7	parents stay on expectation CM at age 7	13178	0.91	0.28	0	1
fread7	father read to children at age 7 weekly	13966	0.36	0.48	0	1
int7_a	Teacher rated parental interest at age 7	14009	2.20	0.73	1	3
int7_1	Teacher rated parental interest at age 7 (no interest)	14009	0.16	0.36	0	1
int7_2	Teacher rated parental interest at age 7 (medium interest)	14009	0.43	0.49	0	1
int7_3	Teacher rated parental interest at age 7 (high interest)	14009	0.39	0.49	0	1
int7_4	Teacher rated parental interest at age 7 (over concerned)	14009	0.03	0.17	0	1
mabook7	Mother reads books or technical journals	14352	0.56	0.50	0	1
manews7	Mother reads newspapers or magazines	14497	0.72	0.45	0	1
mread7	Mother read to children at age 7 weekly	14442	0.49	0.50	0	1
pabook7	Father reads books or technical journals	13971	0.71	0.45	0	1
panews7	Father reads newspapers or magazines	14060	0.85	0.36	0	1
outwithma	Mother goes out with child for walks, picnics, outings, visits etc	13673	0.93	0.25	0	1
outwithpa	Father goes out with child for walks, picnics, outings, visits etc	13173	0.90	0.30	0	1
sepweek7	Child separate from mother more than a week at age 7	14249	0.50	0.50	0	1
<u>Neighbourhood context</u>						
<u>School context</u>						
startsch	age started FT school	14570	5.40	0.83	2	8
presch	attended any preschool or nursery	13434	0.22	0.41	0	1
classz7	class size at age 7	14931	35.26	7.56	2	72
indep7	Child in independent school at age 7	14979	0.03	0.18	0	1
specs7	Child in special school at age 7	14056	0.04	0.20	0	1
spedu7	formally as special education treatment	14046	0.01	0.11	0	1
pgroup7a	Proportion in CM classroom with fathers' SES 1	11860	23.73	24.51	0	100
pgroup7b	Proportion in CM classroom with fathers' SES 5	11860	19.35	19.98	0	100
pgroup7c	Proportion in CM classroom whose parents discussed with teacher	13254	0.52	0.52	0	47
settles7_1	CM settles at school at 7 (no applicable)	13788	0.77	0.42	0	1
settles7_2	CM settles at school at 7 (within a month)	13788	0.17	0.37	0	1
settles7_3	CM settles at school at 7 (between 1 to 3 months)	13788	0.06	0.24	0	1
<u>Cognitive development</u>						
spchtest7	misspronounced word test at age 7	13896	1.51	2.55	0	36
t71m	age 7 standardised maths score	14898	0.00	1.00	-2.05	1.96
t72r	age 7 standardised reading score	14931	0.00	1.00	-3.27	0.93
t73trating	Teacher's rating child's progress at age 7	15042	0.00	1.00	-1.98	2.21

Table 3.1: p3 of 10

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
t74draw	age 7 standardised draw man score	14648	0.00	1.00	-3.37	4.12
i1A1T	poor or very poor oral ability says teacher at 7	15017	0.22	0.41	0	1
i1A6T	very poor oral ability says teacher at 7	15017	0.04	0.20	0	1
i1A1T_m	missing for i1A1T	18558	0.19	0.39	0	1
i1A6T_m	missing for i1A6T	18558	0.19	0.39	0	1
n1AT1	oral ability teacher rating at 7	15017	2.11	0.96	0	4
i1A2T	limited knowledge or largely ignorant of world says teacher at 7	15008	0.28	0.45	0	1
i1A7T	largely ignorant of world says teacher at 7	15008	0.04	0.21	0	1
i1A2T_m	missing for i1A2T	18558	0.19	0.39	0	1
i1A7T_m	missing for i1A7T	18558	0.19	0.39	0	1
n1AT2	world awareness teacher rating at 7	15008	1.93	0.85	0	4
i1A3T	poor or non reader says teacher at 7	14993	0.26	0.44	0	1
i1A8T	non reader says teacher at 7	14993	0.03	0.18	0	1
i1A3T_m	missing for i1A3T	18558	0.19	0.39	0	1
i1A8T_m	missing for i1A8T	18558	0.19	0.39	0	1
n1AT3	reading teacher rating at 7	14993	2.08	0.92	0	4
i1A4T	little or no creativity says teacher at 7	15001	0.33	0.47	0	1
i1A9T	no creativity says teacher at 7	15001	0.04	0.19	0	1
i1A4T_m	missing for i1A4T	18558	0.19	0.39	0	1
i1A9T_m	missing for i1A9T	18558	0.19	0.39	0	1
n1AT4	creativity teacher rating at 7	15001	1.85	0.83	0	4
i1A5T	slow or no understanding in numbers says teacher at 7	15013	0.35	0.48	0	1
i1A10T	no understanding in numbers says teacher at 7	15013	0.04	0.19	0	1
bward7	mental backwardness at age 7	14848	0.13	0.34	0	1
talks2	CM talked by age 2	14457	0.94	0.24	0	1
walks18m	CM walked by 18 months	14534	0.96	0.20	0	1
Affective and behavioural development						
bsag7tot	Total of all syndromes at age 7	14932	8.82	8.87	0	64
dfclt7	CM behavioural difficulties at age 7	14946	0.10	0.30	0	1
disob7_1	CM is disobedient at home at age 7 (never)	14567	0.41	0.49	0	1
disob7_2	CM is disobedient at home at age 7 (little)	14567	0.55	0.50	0	1
disob7_3	CM is disobedient at home at age 7 (frequently)	14567	0.04	0.20	0	1
happysch7_1	CM is happy in current school at age 7. Rep. Mother	14504	0.93	0.25	0	1
happysch7_2	CM is not very happy in current school at age 7. Rep. Mother	14504	0.06	0.24	0	1
happysch7_3	CM is unhappy in current school at age 7. Rep. Mother	14504	0.00	0.06	0	1
n1PRm7	average Rutter parental scale at age 7	18558	0.40	0.33	0	1.05
tearful7_1	CM is tearful at age 7 (never)	14568	0.58	0.49	0	1
tearful7_2	CM is tearful at age 7 (little)	14568	0.38	0.49	0	1
tearful7_3	CM is tearful at age 7 (frequently)	14568	0.04	0.20	0	1
wets3	CM wets during day age 3	14516	0.04	0.20	0	1
wets5	CM soils during day age 4	14538	0.11	0.31	0	1
wets4	CM wets at night by age 5	14528	0.01	0.11	0	1

Table 3.1: p4 of 10

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
attend7	School attendance at age 7	14820	0.91	0.10	0	1
bullied7_1	CM is bullied by children at age 7 (never)	14383	0.65	0.48	0	1
bullied7_2	CM is bullied by children at age 7 (little)	14383	0.30	0.46	0	1
bullied7_3	CM is bullied by children at age 7 (frequently)	14383	0.05	0.23	0	1
dnwsch7	CM reluctant to go to school at age 7	14524	0.10	0.31	0	1
fight7_1	CM fights other children (never)	14492	0.41	0.49	0	1
fight7_2	CM fights other children (little)	14492	0.53	0.50	0	1
fight7_3	CM fights other children (frequently)	14492	0.06	0.23	0	1
i1S24	stays at home to help parent at age 7 BSAG	18558	0.31	0.46	0	1
i1S27	unforthcomingness over 50% BSAG score at age 7	18558	0.33	0.47	0	1
i1S30	writing off adults and as'standards over 80% BSAG score at age 7	18558	0.29	0.46	0	1
memo7_1	emotional maladjustment from med summ at age 7 (no problem)	13954	0.95	0.21	0	1
memo7_2	emotional maladjustment from med summ at age 7 (problem, no handicap)	13954	0.03	0.18	0	1
memo7_3	emotional maladjustment from med summ at age 7 (slight handicap)	13954	0.01	0.10	0	1
memo7_4	emotional maladjustment from med summ at age 7 (moderate handicap)	13954	0.00	0.05	0	1
memo7_5	emotional maladjustment from med summ at age 7 (severe handicap)	13954	0.00	0.02	0	1
Health						
height7	height at age 7 in inches	13782	48.17	2.40	36	61
wtk1	child weight in pounds at age 7	13648	23.88	3.73	10.88	44.90
h1s1	at least one health condition present at age 7	14302	0.31	0.46	0	1
h6	in danger of becoming overweight or overweight at 7 (over 85th pctile, by gender)	14054	0.97	0.17	0	1
mtot7	total score on 14 health conditions at age 7	13973	0.54	1.15	0	27
n1TS17_1	respiratory problems and colds bsag7 (none)	18558	0.85	0.35	0	1
n1TS17_2	respiratory problems and colds bsag7 (poor res)	18558	0.05	0.22	0	1
n1TS17_3	respiratory problems and colds bsag7 (many col)	18558	0.10	0.29	0	1
n1TS19_1	physical defects bsag7 (none)	18558	0.93	0.26	0	1
n1TS19_2	physical defects bsag7 (bad eyes)	18558	0.03	0.17	0	1
n1TS19_3	physical defects bsag7 (squint)	18558	0.01	0.09	0	1
n1TS19_4	physical defects bsag7 (bulging)	18558	0.00	0.05	0	1
n1TS19_5	physical defects bsag7 (poor hea)	18558	0.01	0.12	0	1
n1TS19_6	physical defects bsag7 (no co-or)	18558	0.01	0.11	0	1
n1TS19_7	physical defects bsag7 (contorte)	18558	0.00	0.04	0	1
n1TS19_8	physical defects bsag7 (abnormal)	18558	0.00	0.05	0	1
bmi1	Body Mass Index at 7	14054	51.14	11.98	0	99
Measures at age 11						
Distal family factors						
care11	in care at age 11	13388	0.02	0.13	0	1
fsm11	Receives free school meal at age 11	13955	1.90	0.30	1	2
hardship11	Financial hardship at age 11	13710	0.11	0.31	0	1
hhsz11	household size at age 11	13806	5.16	1.68	2	18
noma11	no mother present at age 11	18558	0.01	0.07	0	1

Table 3.1: p5 of 10

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
nopa11	no father present at age11	13860	0.05	0.22	0	1
nchildren11	number of children under 21 CM age 11	13790	3.07	1.58	1	9
proom11	persons in CM room at age 11	13586	0.82	0.94	0	6
<u>Proximal family processes</u>						
activity1_1	CM reads books (never)	13606	0.12	0.32	0	1
activity1_2	CM reads books (most days)	13606	0.44	0.50	0	1
activity1_3	CM reads books (sometimes)	13606	0.44	0.50	0	1
activity2_1	CM reads newspapers (never)	13545	0.09	0.29	0	1
activity2_2	CM reads newspapers (most days)	13545	0.39	0.49	0	1
activity2_3	CM reads newspapers (sometimes)	13545	0.52	0.50	0	1
cmlll11	parents further training for CM after education at age 11	11181	0.89	0.31	0	1
cmstayon11	parents stay on expectation CM at age 11	11181	0.93	0.25	0	1
int11_a	Teacher rated parental interest at age 11	13824	1.90	0.92	1	3
int11_1	Teacher rated parental interest at age 11 (no interest)	13824	0.11	0.31	0	1
int11_2	Teacher rated parental interest at age 11 (medium interest)	13824	0.14	0.35	0	1
int11_3	Teacher rated parental interest at age 11 (high interest)	13824	0.38	0.49	0	1
int11_4	Teacher rated parental interest at age 11 (over concerned)	13824	0.37	0.48	0	1
so11_child~1	Age 11: child wants to stay on in education - 1=No	18558	0.16	0.36	0	1
<u>Neighbourhood context</u>						
amenH1	Age 11: The neighbourhood has park	13737	0.89	0.31	0	1
amenU1		13737	0.82	0.38	0	1
amenH2	Age 11: The neighbourhood has playground	13685	0.62	0.49	0	1
amenU2		13685	0.50	0.50	0	1
amenH3	Age 11: The neighbourhood has swimming pool	13726	0.80	0.40	0	1
amenU3		13726	0.73	0.44	0	1
amenH4	Age 11: The neighbourhood has indoor play centre	13579	0.88	0.32	0	1
amenU4		13579	0.48	0.50	0	1
amenH5	Age 11: The neighbourhood has cinema	13728	0.75	0.43	0	1
amenU5		13728	0.51	0.50	0	1
amenH6	Age 11: The neighbourhood has library	13734	0.95	0.22	0	1
<u>School context</u>						
classz11	Class size at age 11	13834	34.31	7.61	1	90
pgrp11c	2s % of 11 yr olds suitable for 5 GCSE or O-Levels	13559	26.40	17.60	0	99
<u>Cognitive development</u>						
highab11	CM in high ability class at age 11	13914	0.13	0.34	0	1
lowab11	CM in low ability class at age 11	13914	0.10	0.30	0	1
spedu11	formally as special education treatment at age 11	18558	0.02	0.15	0	1
stream11	CM in streamed class at age 11	13935	0.32	0.47	0	1
t111m	standardised maths score at age 11	14129	0.00	1.00	-1.61	2.26
t112r	standardised reading score at age 11	14133	0.00	1.00	-2.54	3.02
t113trating	Single index of teacher's rating child's progress at age 11	14128	0.00	0.95	-2.54	2.50
t114copy	standardised copying score at age 11	14104	0.00	1.00	-5.58	2.45
bward11	CM mental backwardness at age 11	14142	0.10	0.30	0	1

Table 3.1: p6 of 10

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
i2G1	referral to agency at 11	13763	0.11	0.31	0	1
<u>Affective and behavioural development</u>						
bsag11tot	Bristol Soc.Adjustment at age 11	14160	8.50	8.99	0	70
bullied11_1	CM is bullied by children at age 11 (never)	13436	0.75	0.43	0	1
bullied11_2	CM is bullied by children at age 11 (little)	13436	0.21	0.41	0	1
bullied11_3	CM is bullied by children at age 11 (frequently)	13436	0.04	0.20	0	1
dfc11	CM behavioural difficulties at age 11	14137	0.02	0.13	0	1
disob11_1	CM is disobedient at age 11 (never)	13770	0.49	0.50	0	1
disob11_2	CM is disobedient at age 11 (little)	13770	0.48	0.50	0	1
disob11_3	CM is disobedient at age 11 (frequently)	13770	0.03	0.17	0	1
fight11_1	CM fights other children at age 11 (never)	13612	0.49	0.50	0	1
fight11_2	CM fights other children at age 11 (little)	13612	0.46	0.50	0	1
fight11_3	CM fights other children at age 11 (frequently)	13612	0.04	0.20	0	1
tearful11_1	CM is tearful at age 11 (never)	13752	0.59	0.49	0	1
tearful11_2	CM is tearful at age 11 (little)	13752	0.37	0.48	0	1
tearful11_3	CM is tearful at age 11 (frequently)	13752	0.04	0.19	0	1
n1PRm11	avg Rutter parental scale at 11	18558	0.43	0.35	0	1
dnwsch11	CM reluctant to go to school at age 11	13732	0.07	0.26	0	1
att11	percentile position on attendance at age 11	11789	48.91	27.64	1	90
i2B7	receiving special help for behaviour difficulties teacher at age 11	14137	0.01	0.12	0	1
i2B9	has had bad dreams or night terrors says parent age 11	13644	0.10	0.30	0	1
i2S5	not good attendance at age 11 BSAG	13732	0.08	0.28	0	1
n2S3	depression bsag at age 11	14162	1.05	1.55	0	10
n2S4	anxiety for acceptance by adults bsag at age 11	14162	0.56	1.21	0	12
n2S5	hostility towards adults bsag at age 11	14162	0.90	1.95	0	19
n2S6	writing off adults and adult standards bsag at age 11	14162	1.02	1.70	0	16
n2S7	anxiety for acceptance by children bsag at age 11	14162	0.33	0.80	0	8
n2S8	hostility towards children bsag at age 11	14162	0.29	0.81	0	10
n2S9	restlessness bsag at age 11	14162	0.23	0.57	0	4
<u>Health</u>						
htm2	child height at age 11 in centimeter w/ stadiometer	12733	1.44	0.07	1.02	1.70
wtk2	child weight in pounds at age 11	12606	36.57	7.32	15.42	84.80
i2S24	depression over 80% BSAG score at age 11	14162	0.15	0.36	0	1
i227_m	missing for poor physical co-ordination, e.g., running says teacher at age 11	18558	0.26	0.44	0	1
i227_1	poor physical co-ordination, e.g., running says teacher at age 11 (not at all)	13653	0.85	0.36	0	1
i227_2	poor physical co-ordination, e.g., running says teacher at age 11 (somewhat)	13653	0.13	0.33	0	1
i227_3	poor physical co-ordination, e.g., running says teacher at age 11 (certainly)	13653	0.03	0.16	0	1
mot111	laterality health measure at age 11	12790	0.91	0.66	0	1.75
h5s2	whether hearing difficulty constitutes a problem at age 11	11984	0.01	0.11	0	1
h7s2	whether vision defect constitutes problem at age 11	12307	0.02	0.15	0	1
bmi2		12548	17.47	2.60	8.90	37.76

Measures at age 16

Distal family factors

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
care16	in care at age 16	11104	0.02	0.12	0	1
noma16	no mother present at age 16	11691	0.01	0.12	0	1
nopa16	no father present at age 16	11688	0.07	0.26	0	1
hardship16	Financial hardship at age 16	11601	0.10	0.30	0	1
fsm16	Receives free school meal at age 16	11509	1.90	0.30	1	2
hsize16	household size at age 16	11628	4.89	1.64	2	17
nchildren16	number of children under 20 at age 16	11628	1.74	1.52	0	14
proom16	persons in CM room at age 16	11630	0.53	0.78	0	7
masmokes	mother smokes more than 5 cigarettes per day	11276	0.39	0.49	0	1
<u>Proximal family processes</u>						
int16_a	Teacher rated parental interest 16	11830	1.78	0.96	0	3
int16_1	Teacher rated parental interest 16 (not applicable)	11830	0.05	0.21	0	1
int16_2	Teacher rated parental interest 16 (no interest)	11830	0.13	0.33	0	1
int16_3	Teacher rated parental interest 16 (medium interest)	11830	0.16	0.37	0	1
int16_4	Teacher rated parental interest 16 (high interest)	11830	0.33	0.47	0	1
int16_5	Teacher rated parental interest 16 (over concerned)	11830	0.33	0.47	0	1
cmlll16	CM further training after education at age 16	12026	0.64	0.48	0	1
cmstayon16	parents stay on expectation CM at age 16	10900	0.61	0.49	0	1
so16_chi~a_1	Likely to leave school at age 16	11131	0.68	0.47	0	1
so16_chi~a_2	Likely to leave school at age 17	11131	0.08	0.27	0	1
so16_child~3	Likely to leave school at age 18	11131	0.25	0.43	0	1
<u>Neighbourhood context</u>						
<u>School context</u>						
ratio	pupils per teacher at school at age 16	12452	17.91	8.32	1.62	885
classz16m	3s no. children in maths class	12092	25.11	6.22	0	56
classz16e	3s no. children in english class	12146	25.62	5.87	0	65
pgroup	Proportion of CM whose parents are in manual, studied GCE&CSE and stayed on	9979	0.49	0.23	0.06	1
S16comp	CM in comprehensive school at age 16	12749	0.58	0.49	0	1
S16gram	CM in grammar school at age 16	12749	0.11	0.31	0	1
S16priv	CM in private secondary at age 16	12749	0.06	0.24	0	1
schatt11_16	number of schools attended age 11 to 16	11613	1.25	0.53	1	6
so16_teach~1	Child will benefit it for staying in education. Resp teacher	10904	0.49	0.50	0	1
so16_teach~2	Child won't benefit it for staying in education. Resp teacher	10904	0.51	0.50	0	1
pgp16a	Percentage of fathers in non-manual jobs for childs in school at age < 16	10956	4.07	2.36	1	9
pgp16b	Percentage studying for GCE O level	12026	25.81	33.65	0	100
pgp16c	Percentage studying for CSE only	12026	28.28	26.89	0	100
pgp16d	Percentage remained at school after they could have left	11817	59.90	27.78	0	100
pgp16e	Percentage pupils expelled	12111	0.00	0.00	0	0.16
<u>Cognitive development</u>						
high_1	High ability group in streamed english/math class at age 16 (= NO)	8276	0.47	0.50	0	1
high_2	High ability group in streamed english/math class at age 16 (= only 1)	8276	0.14	0.34	0	1
high_3	High ability group in streamed english/math class at age 16 (= in both)	8276	0.39	0.49	0	1

Table 3.1: p8 of 10

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
t161m	standardised maths score at age 16	11921	0.00	1.00	-1.82	2.61
t162r	standardised reading score at age 16	11987	0.00	1.00	-3.57	1.37
t16trating	Single index of teacher's rating child's progress at age 16	5180	0.00	1.00	-1.84	1.50
t16childse-p	Total score for CM self-perception of ability at age 16	18558	0.00	0.83	-1.57	2.29
bcskills16	CM does not have basic skills by age 16	12573	0.03	0.18	0	1
englow	Low ability group in streamed english class at age 16	8686	0.46	0.50	0	1
mathlow	Low ability group in streamed math class at age 16	10228	0.50	0.50	0	1
i3a1T	Cannot read well enough for everyday needs at age 16	12412	0.02	0.13	0	1
i3a2T	cannot do calculations for shopping at age 16	12373	0.03	0.16	0	1
<u>Affective and behavioural development</u>						
n1PRm16	Average Rutter parental scale at age 16	18558	0.45	0.43	0	1
fight16_1	CM fights other children at age 16 (never)	11625	0.88	0.33	0	1
fight16_2	CM fights other children at age 16 (little)	11625	0.11	0.31	0	1
fight16_3	CM fights other children at age 16 (frequently)	11625	0.02	0.13	0	1
external16	Single index of teacher's rating external behavioural problems at age 16	12541	1.21	0.33	1	3
internal16	Single index of teacher's rating internal behavioural problems at age 16	12508	1.23	0.30	1	3
jim	Scores for factor 1 at age 16	18558	0.00	0.90	-2.78	1.78
spneed16	special need after FT-education at age 16	11452	1.97	0.18	1	2
i3m13	factor score on views at age 16 in most negative 20%	11519	0.20	0.40	0	1
ch_absent	child absent from school at age16	11625	0.11	0.31	0	1
rut26_1	3s bullies other children (does not apply)	12444	0.92	0.27	0	1
rut26_2	3s bullies other children (applies)	12444	0.06	0.24	0	1
rut26_3	3s bullies other children (certainly)	12444	0.02	0.14	0	1
rut10_1	3s often appears miserable,unha (does not apply)	12459	0.85	0.35	0	1
rut10_2	3s often appears miserable,unha (applies)	12459	0.13	0.33	0	1
rut10_3	3s often appears miserable,unha (certainly)	12459	0.02	0.14	0	1
rut15_1	3s is often disobedient (does not apply)	12455	0.82	0.38	0	1
rut15_2	3s is often disobedient (applies)	12455	0.14	0.34	0	1
rut15_3	3s is often disobedient (certainly)	12455	0.04	0.20	0	1
rut16_1	3s cannot settle more than a fe (does not apply)	12452	0.79	0.41	0	1
rut16_2	3s cannot settle more than a fe (applies)	12452	0.17	0.37	0	1
rut16_3	3s cannot settle more than a fe (certainly)	12452	0.04	0.20	0	1
rut17_1	3s fearful of new situations & (does not apply)	12436	0.77	0.42	0	1
rut17_2	3s fearful of new situations & (applies)	12436	0.20	0.40	0	1
rut17_3	3s fearful of new situations & (certainly)	12436	0.03	0.16	0	1
i3b1	has been in trouble with police says teacher at age 16	11368	0.09	0.29	0	1
i3b11	attendance in bottom 30% in 1973	12510	0.30	0.46	0	1
i3b15	attendance in bottom 30% in 1972 and 3	11311	0.19	0.39	0	1
<u>Health</u>						
bmi3	body mass index at 16	11063	20.61	2.89	11.56	43.95
htm3	child height at age 16 in feet/100	11149	1.66	0.09	1.21	1.98
med	any medical condition at age 16	18558	0.26	0.44	0	1
med3	3m mental retardation	11375	0.02	0.15	0	1

Table 3.1: p9 of 10

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
obese16	overweight at age 16	11428	0.11	0.31	0	1
mtot16	total score on 14 health conditions at age 16	11174	0.48	1.23	0	29

Table 3.2: Summary statistics for predictor variables, 1970 Cohort (BCS)

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
<i>Measures at age 0</i>						
<u>Distal family factors</u>						
pasoc01	Father's SES 1	17100	0.05	0.21	0	1
pasoc02	Father's SES 2	17100	0.11	0.31	0	1
pasoc03	Father's SES 3 non-manual	17100	0.11	0.32	0	1
pasoc04	Father's SES 3 manual	17100	0.44	0.50	0	1
pasoc05	Father's SES 4	17100	0.14	0.35	0	1
pasoc06	Father's SES 5	17100	0.06	0.25	0	1
pasoc07	Father's SES other	17100	0.03	0.17	0	1
pasoc08	Father's SES missing	17100	0.05	0.21	0	1
masoc01	Mother's SES 1 or 2	15689	0.09	0.29	0	1
masoc02	Mother's SES 3 non-manual	15689	0.30	0.46	0	1
masoc03	Mother's SES 3 manual	15689	0.05	0.23	0	1
masoc04	Mother's SES 4	15689	0.21	0.41	0	1
masoc05	Mother's SES 5	15689	0.01	0.12	0	1
masoc06	Mother's SES other	15689	0.01	0.08	0	1
masoc07	Mother homemaker / OR missing??	15689	0.33	0.47	0	1
f1nold	Number of older siblings	18615	1.17	1.35	0	17
fethwi	West-indian	18615	0.01	0.11	0	1
fethukwi	UK/West Indian	18615	0.00	0.04	0	1
englang	English as first language	13265	0.97	0.18	0	1
maageg1	Mother age 14-18	17094	0.10	0.30	0	1
maageg2	Mother age 19-24	17094	0.43	0.50	0	1
maageg3	Mother age 25-34	17094	0.23	0.42	0	1
maageg4	Mother age 35+	17094	0.24	0.43	0	1
nopa0	No father present	18615	0.05	0.21	0	1
paune0	Father unemployed	17197	0.04	0.18	0	1
cmregb1	Birth region 1	17197	0.09	0.28	0	1
cmregb2	Birth region 2	17197	0.13	0.33	0	1
cmregb3	Birth region 3	17197	0.06	0.24	0	1
cmregb4	Birth region 4	17197	0.10	0.30	0	1
cmregb5	Birth region 5	17197	0.03	0.17	0	1
cmregb6	Birth region 6	17197	0.06	0.24	0	1
cmregb7	Birth region 7	17197	0.04	0.19	0	1
cmregb8	Birth region 8	17197	0.02	0.12	0	1
cmregb9	Birth region 9	17197	0.17	0.37	0	1
cmregb10	Birth region 10	17197	0.00	0.00	0	0
cmregb11	Birth region 11	17197	0.12	0.33	0	1
cmregb12	Birth region 12	17197	0.09	0.29	0	1

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
cmregb13	Birth region 13	17197	0.04	0.19	0	1
cmregb14	Birth region 14	17197	0.00	0.03	0	1
mumso0	Mother stayed on in school post-16	17050	0.35	0.48	0	1
<u>Proximal family processes</u>						
<u>Neighbourhood context</u>						
<u>School context</u>						
<u>Cognitive development</u>						
<u>Affective and behavioural development</u>						
<u>Health</u>						
weight0q_1	Birthweight quintile 1	17140	0.21	0.41	0	1
weight0q_2	Birthweight quintile 2	17140	0.20	0.40	0	1
weight0q_3	Birthweight quintile 3	17140	0.21	0.41	0	1
weight0q_4	Birthweight quintile 4	17140	0.19	0.39	0	1
weight0q_5	Birthweight quintile 5	17140	0.19	0.40	0	1
<i>Measures at 22 months (sub-sample)</i>						
<u>Distal family factors</u>						
pasoc221	Father's SES 1	2291	0.06	0.23	0	1
pasoc222	Father's SES 2	2291	0.16	0.37	0	1
pasoc223	Father's SES 3 non-manual	2291	0.10	0.30	0	1
pasoc224	Father's SES 3 manual	2291	0.46	0.50	0	1
pasoc225	Father's SES 4	2291	0.16	0.37	0	1
pasoc226	Father's SES 5	2291	0.05	0.22	0	1
paune22	Father unemployed	2457	0.07	0.25	0	1
<u>Proximal family processes</u>						
<u>Neighbourhood context</u>						
<u>School context</u>						
<u>Cognitive development</u>						
cubes22	Cubes	2436	38.25	22.36	0	100
lang222	Language	2436	49.95	43.23	0	100
pers2221	Personal development	2436	69.66	36.61	0	100
pencil22	Copying designs	2436	37.90	21.94	0	100

Table 3.2: p2 of 10

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
<u>Affective and behavioural development</u>						
<u>Health</u>						
height22q_1	Height quintile 1	2277	0.23	0.42	0	1
height22q_2	Height quintile 2	2277	0.25	0.43	0	1
height22q_3	Height quintile 3	2277	0.15	0.35	0	1
height22q_4	Height quintile 4	2277	0.21	0.41	0	1
height22q_5	Height quintile 5	2277	0.17	0.37	0	1
<i>Measures at 42 months (sub-sample)</i>						
<u>Distal family factors</u>						
pasoc421	Father's SES 1	2087	0.06	0.23	0	1
pasoc422	Father's SES 2	2087	0.16	0.37	0	1
pasoc423	Father's SES 3 non-manual	2087	0.12	0.33	0	1
pasoc424	Father's SES 3 manual	2087	0.45	0.50	0	1
pasoc425	Father's SES 4	2087	0.15	0.36	0	1
pasoc426	Father's SES 5	2087	0.04	0.19	0	1
pasoc427	Father's SES other	2087	0.01	0.11	0	1
paune42	Father unemployed	2315	0.03	0.16	0	1
<u>Proximal family processes</u>						
<u>Neighbourhood context</u>						
<u>School context</u>						
<u>Cognitive development</u>						
count42	Counting	2297	35.13	33.66	0	100
speak42	Speaking	2297	50.03	29.75	0	100
copy421	Copying Designs	2297	41.81	26.56	0	100
copy422		2297	16.80	30.86	0	100
<u>Affective and behavioural development</u>						
t424ext	Externalising behaviour age 42 months	18779	0.00	0.35	-1.06	2.11
t424ext_m	Externalising behaviour age 42 months_missing	18779	0.88	0.33	0	1
<u>Health</u>						
height42q_1	Height quintile 1	2200	0.27	0.44	0	1
height42q_2	Height quintile 2	2200	0.21	0.41	0	1
height42q_3	Height quintile 3	2200	0.13	0.34	0	1
height42q_4	Height quintile 4	2200	0.20	0.40	0	1

Table 3.2: p3 of 10

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
height42q_5	Height quintile 5	2200	0.19	0.39	0	1

Measures at age 5

Distal family factors

dadso5		12962	0.42	0.59	0	2
dadso51	Father education, age 16	12962	0.63	0.48	0	1
dadso52	Father education, age 17	12962	0.32	0.47	0	1
dadso53	Father education, age 18	12962	0.05	0.22	0	1
mqual51	Mother, qualifications level 1	12606	0.56	0.50	0	1
mqual52	Mother, qualifications level 2	12606	0.14	0.35	0	1
mqual53	Mother, qualifications level 3	12606	0.18	0.38	0	1
mqual54	Mother, qualifications level 4	12606	0.04	0.18	0	1
mqual55	Mother, qualifications level 5	12606	0.03	0.16	0	1
mqual56	Mother, qualifications level 6	12606	0.02	0.15	0	1
mqual58	Mother, qualifications level (no mother)	12606	0.01	0.10	0	1
mqual59	Mother, qualifications level unknown	12606	0.00	0.06	0	1
pqual51	Father, qualifications level 1	12420	0.46	0.50	0	1
pqual52	Father, qualifications level 2	12420	0.10	0.30	0	1
pqual53	Father, qualifications level 3	12420	0.16	0.36	0	1
pqual54	Father, qualifications level 4	12420	0.07	0.25	0	1
pqual55	Father, qualifications level 5	12420	0.00	0.05	0	1
pqual56	Father, qualifications level 6	12420	0.01	0.09	0	1
pqual57	Father, qualifications level 6	12420	0.13	0.34	0	1
pqual58	Father, qualifications level (no father)	12420	0.02	0.14	0	1
pqual59	Father, qualifications level unknown	12420	0.05	0.22	0	1
pasoc51	Father's SES 1	12940	0.05	0.22	0	1
pasoc52	Father's SES 2	12940	0.07	0.25	0	1
pasoc53	Father's SES 3 non-manual	12940	0.19	0.39	0	1
pasoc54	Father's SES 3 manual	12940	0.08	0.28	0	1
pasoc55	Father's SES 4	12940	0.44	0.50	0	1
pasoc56	Father's SES 5	12940	0.13	0.33	0	1
pasoc57	Father's SES other	12940	0.05	0.21	0	1
pasoc58	Father's SES missing????	12940	0.00	0.04	0	1
masoc51	Mother's SES 1 or 2	13101	0.00	0.05	0	1
masoc52	Mother's SES 3 non-manual	13101	0.06	0.24	0	1
masoc53	Mother's SES 3 manual	13101	0.11	0.31	0	1
masoc54	Mother's SES 4	13101	0.03	0.18	0	1
masoc55	Mother's SES 5	13101	0.12	0.33	0	1
masoc56	Mother's SES other	13101	0.04	0.20	0	1
masoc57	Mother homemaker ???????	13101	0.01	0.08	0	1
masoc58	Mother's SES missing??????	13101	0.62	0.48	0	1

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
paageg1	Father age 14-18	12947	0.03	0.16	0	1
paageg2	Father age 19-24	12947	0.29	0.45	0	1
paageg3	father age 25-34	12947	0.25	0.43	0	1
paageg4	father age 35+	12947	0.38	0.49	0	1
paageg5	Father age missing????	12947	0.05	0.22	0	1
paune5	Father unemployed	13136	0.07	0.25	0	1
nyoung5	Number of younger siblings	13063	0.52	0.64	0	5
pproom	People per room	12874	0.90	0.33	0.07	6
noma5	No mother	13136	0.00	0.06	0	1
nopa5	No father	13136	0.05	0.22	0	1
care5	Experience of care by 5	16449	0.02	0.13	0	1
<u>Proximal family processes</u>						
mamal5_c	Mother depressed	9055	4.33	3.63	0	23
pareads	father reads to child	12780	0.47	0.50	0	1
mareadpw	How often mother reads to CM per week	12281	3.68	2.90	0	7
rpw	How often read to per week	12557	4.29	2.59	0	7
antitv	Mother anti-TV attitude	9132	-0.01	0.52	-1.14	1.75
unauth	Mother low-authoritarian attitude	9132	0.02	0.85	-2.34	2.69
<u>Neighbourhood context</u>						
poorned		12712	0.08	0.27	0	1
goodned		12712	0.23	0.42	0	1
<u>School context</u>						
<u>Cognitive development</u>						
t51copy	Copying score age 5	13136	58.60	25.20	0	100
t52soc	Anti-social behaviour score	13136	51.97	12.63	0	100
t53reads	Reading score	12229	40.79	36.35	0	100
t54hfd1	Human Figure Drawing age 5	13136	45.76	13.74	0	100
t55hfd2	Human Figure Drawing age 5	13136	44.78	14.70	0	100
t56profl	Profile Drawing age 5	13136	40.98	26.09	0	100
t57vocab	Vocabulary score age 5	13136	57.40	26.52	0	100
<u>Affective and behavioural development</u>						
ext5	Externalising behavs age 5	14379	0.00	0.88	-1.36	5.27
int5	Internalising behavs age 5	14374	0.00	0.88	-1.28	4.43
wets5	Soiling probs age 5	9132	0.03	0.16	0	1
<u>Health</u>						

Table 3.2: p5 of 10

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
height5q_1	Height quintile 1	12822	0.22	0.42	0	1
height5q_2	Height quintile 2	12822	0.18	0.39	0	1
height5q_3	Height quintile 3	12822	0.26	0.44	0	1
height5q_4	Height quintile 4	12822	0.16	0.36	0	1
height5q_5	Height quintile 5	12822	0.18	0.39	0	1

Measures at age 10

<u>Distal family factors</u>						
pasoc101	Father's SES 1	13187	0.06	0.24	0	1
pasoc102	Father's SES 2	13187	0.21	0.41	0	1
pasoc103	Father's SES 3 non-manual	13187	0.08	0.27	0	1
pasoc104	Father's SES 3 manual	13187	0.43	0.49	0	1
pasoc105	Father's SES 4	13187	0.11	0.32	0	1
pasoc106	Father's SES 5	13187	0.04	0.19	0	1
pasoc107	Father's SES other	13187	0.07	0.25	0	1
masoc101	Mother's SES 1 or 2	9538	0.01	0.10	0	1
masoc102	Mother's SES 3 non-manual	9538	0.18	0.39	0	1
masoc103	Mother's SES 3 manual	9538	0.33	0.47	0	1
masoc104	Mother's SES 4	9538	0.09	0.28	0	1
masoc105	Mother's SES 5	9538	0.29	0.46	0	1
masoc106	Mother's SES other	9538	0.09	0.29	0	1
paune10	Mother homemaker ??????	13872	0.10	0.31	0	1
f2nyoung	Number of younger siblings	18615	0.52	0.54	0	5
noma10	No mother	18453	0.01	0.07	0	1
nopa10	No father	18453	0.05	0.23	0	1
tel	Own a telephone????	13033	0.57	0.49	0	1
finc_1	Income < £35	8766	0.02	0.12	0	1
finc_2	Income £35-£49	8766	0.04	0.21	0	1
finc_3	Income £50-£99	8766	0.29	0.45	0	1
finc_4	Income £100-£149	8766	0.35	0.48	0	1
finc_5	Income £150-£199	8766	0.17	0.38	0	1
finc_6	Income £200-£249	8766	0.07	0.25	0	1
mqual101	Mother, qualifications level 1	6081	0.31	0.46	0	1
mqual102	Mother, qualifications level 2	6081	0.36	0.48	0	1
mqual103	Mother, qualifications level 3	6081	0.07	0.25	0	1
mqual104	Mother, qualifications level 4	6081	0.05	0.21	0	1
mqual105	Mother, qualifications level 5	6081	0.06	0.23	0	1
mqual106	Mother, qualifications level (no mother)	6081	0.06	0.23	0	1
mqual107	Mother, qualifications level unknown	6081	0.09	0.29	0	1
pqual101	Father, qualifications level 1	8533	0.28	0.45	0	1
pqual102	Father, qualifications level 2	8533	0.21	0.41	0	1

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
pqual103	Father, qualifications level 3	8533	0.12	0.32	0	1
pqual104	Father, qualifications level 4	8533	0.00	0.05	0	1
pqual105	Father, qualifications level 5	8533	0.01	0.11	0	1
pqual106	Father, qualifications level 6	8533	0.16	0.37	0	1
pqual107	Father, qualifications level (no father)	8533	0.10	0.30	0	1
pqual108	Father, qualifications level unknown	8533	0.12	0.32	0	1
care10	In care by age 10	16449	0.01	0.11	0	1
reg1	Region 1	18615	0.32	0.46	0	1
reg2	Region 2	18615	0.04	0.21	0	1
reg3	Region 3	18615	0.07	0.25	0	1
reg4	Region 4	18615	0.09	0.28	0	1
reg5	Region 5	18615	0.05	0.22	0	1
reg6	Region 6	18615	0.07	0.26	0	1
reg7	Region 7	18615	0.02	0.14	0	1
reg9	Region 8	18615	0.05	0.22	0	1
reg10	Region 9	18615	0.04	0.19	0	1
reg11	Region 10	18615	0.07	0.26	0	1
reg12	Region 11	18615	0.00	0.05	0	1
<u>Proximal family processes</u>						
so_	Mother wants child to stay on past 16	12743	0.58	0.49	0	1
so_2	Mother wants child to stay on past 17	12743	0.42	0.49	0	1
spprobm1	Mother "over-concerned" with child, teacher-rated	18615	0.05	0.21	0	1
spprobm2	Mother "hostile to child," teacher-rated	18615	0.00	0.05	0	1
spprobm3	Mother "dismissive to child," teacher-rated	18615	0.01	0.12	0	1
spprobp1	Father "over-concerned" with child, teacher-rated	18615	0.03	0.16	0	1
spprobp2	Father "hostile to child," teacher-rated	18615	0.00	0.04	0	1
spprobp3	Father "dismissive to child," teacher-rated	18615	0.01	0.09	0	1
<u>Neighbourhood context</u>						
<u>School context</u>						
s2indep	Independent school	12758	0.02	0.15	0	1
remed101		12345	3.48	3.46	0	35
remed102		11359	1.93	3.28	0	35
remed103		10226	0.15	0.91	0	28
spg1PrH	Peer group: High level of high SES	18615	14.47	13.97	0	100
spg3PrL	Peer group: Low level of high SES	18615	33.07	18.90	0	100
spg4AcL	Peer group: Low level of high academic achievement	18615	7.17	6.34	0	100
sset1rH	High set for reading	18311	0.12	0.32	0	1
sset2rM	Medium set for reading	18311	0.15	0.36	0	1

Table 3.2: p7 of 10

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
sset3rL	Low set for reading	18311	0.09	0.28	0	1
sset5mM	Medium set for maths	18311	0.19	0.39	0	1
sset6mL	Low set for maths	18311	0.10	0.31	0	1
shrs1	Curriculum time on subject 1	12616	1.00	0.60	0	6.5
shrs2	Curriculum time on subject 2	12615	1.59	0.72	0	6
shrs3	Curriculum time on subject 3	12616	1.95	1.41	0	10.5
shrs5	Curriculum time on subject 4	12616	0.96	0.77	0	10.5
shrs6	Curriculum time on subject 5	12616	1.56	0.83	0	10
shrs7	Curriculum time on subject 6	12616	4.71	1.13	0	12.5
shrs8	Curriculum time on subject 7	12616	0.98	0.59	0	5
shrs9	Curriculum time on subject 8	12615	1.96	1.23	0	15.95
shrs10	Curriculum time on subject 9	12616	0.30	0.76	0	12
shrs11	Curriculum time on subject 10	12616	1.89	0.70	0	10.5
shrs12	Curriculum time on subject 11	12616	1.03	0.53	0	11.5
shrs13	Curriculum time on subject 12	12616	0.39	0.41	0	6
shrs14	Curriculum time on subject 13	12616	1.92	0.71	0	8
shrs15	Curriculum time on subject 14	12616	2.20	2.29	0	24.25
shrs16	Curriculum time on subject 15	12616	0.41	1.15	0	21
shrs17	Curriculum time on subject 16	12617	0.47	1.20	0	27.5
<u>Cognitive development</u>						
t101m	Maths test age 10	11720	0.00	1.00	-3.43	2.22
t102r	Reading test age 10	12791	0.00	1.00	-2.91	1.92
t112bas	British Ability Scales	12498	0.00	1.00	-3.77	3.42
special	Special school	18779	0.01	0.07	0	1
<u>Affective and behavioural development</u>						
t103cont	Locus of control age 10	12445	0.00	1.00	-2.64	2.72
t104este	Self esteem age 10	12520	0.00	1.00	-2.93	1.87
t105asoc	Anti-social behaviour	12758	0.00	1.00	-1.51	4.39
t106peer	Peer relations age 10	12758	0.00	1.00	-3.45	2.02
t107att	(In)attention behavs age 10	12758	0.00	1.00	-2.93	1.78
t108ext	Externalising behavs age 10	12758	0.00	1.00	-3.07	2.11
t111com	Communication skills	12758	0.00	1.00	-3.21	2.27
t113anx	Anxiety	12758	0.00	1.00	-2.04	3.75
t114clu	Clumsy behaviour	12758	0.00	1.00	-1.45	5.44
t115hyp	Hyperactivity	12758	0.00	1.00	-1.55	4.28
ts101ext	Externalising behavs age 10	13543	21.68	13.65	0	100
ts102int	Internalising behavs age 10	13536	28.08	16.91	0	100
wets10	Soiling probs b/w 5 & 10	14423	0.01	0.07	0	1
truant10	Truant at age 10	12758	0.01	0.08	0	1

Table 3.2: p8 of 10

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
<u>Health</u>						
height10q_1	Height quintile 1	6667	0.22	0.41	0	1
height10q_2	Height quintile 2	6667	0.19	0.39	0	1
height10q_3	Height quintile 3	6667	0.19	0.39	0	1
height10q_4	Height quintile 4	6667	0.21	0.40	0	1
height10q_5	Height quintile 5	6667	0.19	0.39	0	1
Measures at age 16						
<u>Distal family factors</u>						
care16	Experience of being in care by age 16???	16449	0.01	0.12	0	1
<u>Proximal family processes</u>						
<u>Neighbourhood context</u>						
<u>School context</u>						
<u>Cognitive development</u>						
<u>Affective and behavioural development</u>						
fqcont1_sp	Frequency of going to sports centre	6355	1.19	1.32	0	4
fqcont1_cc	Frequency of going to community centre	6359	0.23	0.78	0	4
fqcont2	Frequency of going to youth club	6391	0.74	1.31	0	4
fqcont3	Frequency of going to uniformed youth club	6386	0.37	1.05	0	4
fqcont4	Frequency of going to church activities	6417	0.61	1.16	0	4
vol_1	non-paid activity:babysitting elsewhere	6417	0.46	0.50	0	1
vol_2	non-paid activity:helping old people	6417	0.32	0.47	0	1
vol_3	non-paid activity:nature conservation	6417	0.08	0.27	0	1
vol_4	non-paid activity:helping single parents	6417	0.07	0.26	0	1
vol_5	other voluntary community activities	6417	0.12	0.33	0	1
sact_1	participation in school concerts / plays	6417	0.28	0.45	0	1
sact_2	participation in excursions / day trips	6417	0.51	0.50	0	1
sact_3	participation in school holidays	6417	0.28	0.45	0	1
sact_4	participation in field courses	6417	0.29	0.45	0	1
sact_5	participation in clubs / societies	6417	0.21	0.41	0	1
sact_6	participation in charitable activities	6417	0.23	0.42	0	1
sact_7	participation in other school activity	6417	0.05	0.22	0	1
truant	Has truanted	13503	0.44	0.35	0	1.70
smoker16	Smokes at 16	13503	0.24	0.30	-0.18	1.40
drink_lwk	Has had alcoholic drink in last wk	13503	1.31	0.89	-0.09	7

Table 3.2: p9 of 10

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
sloc16	Locus of control age 16	13503	9.68	2.36	-0.06	16.02
so16	Wants to stay on age 16	13503	1.18	0.65	-0.89	2.39
selfe16	Self efficacy age 16	13503	14.72	2.38	0	20
asoc1_gen	CM engaged in gen. antisoc acts	13503	11.29	7.35	-0.06	104
asoc2_crim	CM engaged in criminal antisoc acts	13503	2.04	3.38	-1.35	52
asoc3_pal	Close friends engaged in antisoc acts	13503	0.43	0.90	-0.38	13
asoc4_cas	Casual friends engaged in antisoc acts	13503	0.68	0.98	-0.39	13
asoc5_acq	Acquaintances engaged in antiso acts	13503	1.48	1.38	-0.08	13
papprove_a~s	Parents approve of activities	18725	2.65	0.34	1	3
papprove_fr	Parents approve of friends	18725	2.67	0.33	1	3
cond161_ma~e	Mother-rated (MR) behaviour: general conduct disorder	13503	-0.01	0.75	-2.03	4.05
cond162_ma~e	MR behaviour: clumsy naughty	13503	-0.01	0.71	-2.19	6.88
cond163_ma~e	MR behaviour: clumsy frustrated	13503	0.01	0.73	-8.13	2.54
cond164_ma~e	MR behaviour: internalising	13503	-0.01	0.74	-4.08	2.24

Health

Table 3.3: Summary statistics for predictor variables, ALSPAC

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
<i>Measures at age 0</i>						
<u>Distal family factors</u>						
birthorder	Birthorder	13059	1.85	0.97	1	14
breastfeed_1	Breastfeed==never	11146	0.25	0.43	0	1
breastfeed_2	Breastfeed==<1 month	11146	0.16	0.37	0	1
breastfeed_3	Breastfeed==1-<3 months	11146	0.16	0.37	0	1
breastfeed_4	Breastfeed==3-<6 months	11146	0.13	0.33	0	1
breastfeed_5	Breastfeed==6 or more	11146	0.30	0.46	0	1
mageatb	Mother's age at birth of study child	14082	28.00	4.96	15	44
mageatb2	Mother's age at birth of study child (squared)	14082	808.56	282.43	225	1936
mchhome_1	Mother's own childhood home type==council housing	9664	0.22	0.42	0	1
mchhome_2	Mother's own childhood home type==being bought	9664	0.40	0.49	0	1
mchhome_3	Mother's own childhood home type==owned	9664	0.29	0.45	0	1
mchhome_4	Mother's own childhood home type==other rented	9664	0.03	0.17	0	1
mchhome_5	Mother's own childhood home type==other	9664	0.05	0.22	0	1
medu_1	Mother's education==cse	11723	0.15	0.36	0	1
medu_2	Mother's education==vocational	11723	0.10	0.31	0	1
medu_3	Mother's education==o level	11723	0.37	0.48	0	1
medu_4	Mother's education==a level	11723	0.24	0.43	0	1
medu_5	Mother's education==degree	11723	0.14	0.34	0	1
methnic	Mother's ethnic group	12407	1.03	0.16	1	2
mfedu_1	Mother's father's education==cse	6799	0.27	0.44	0	1
mfedu_2	Mother's father's education==vocational	6799	0.25	0.44	0	1
mfedu_3	Mother's father's education==o level	6799	0.15	0.36	0	1
mfedu_4	Mother's father's education==a level	6799	0.18	0.38	0	1
mfedu_5	Mother's father's education==degree	6799	0.15	0.36	0	1
mfindiff	Mother has experienced financial difficulties	12168	2.91	3.54	0	15
mfindiff2	Mother has experienced financial difficulties (squared)	12168	21.00	40.27	0	225
mfjobstatu~1	Mother's father's job status==foreman	12578	0.06	0.24	0	1
mfjobstatu~2	Mother's father's job status==manager	12578	0.17	0.38	0	1
mfjobstatu~3	Mother's father's job status==supervisor	12578	0.07	0.25	0	1
mfjobstatu~4	Mother's father's job status==leading hand	12578	0.06	0.23	0	1
mfjobstatu~5	Mother's father's job status==self employed	12578	0.14	0.35	0	1
mfjobstatu~6	Mother's father's job status==none of these	12578	0.34	0.47	0	1
mfjobstatu~7	Mother's father's job status==status missing	12578	0.16	0.37	0	1
mfpresent	Mother's father present during own childhood	12609	0.78	0.41	0	1
mhomeless	Mother experienced homelessness during pregnancy	11939	0.02	0.15	0	1
mmedu_1	Mother's mother's education==cse	7214	0.33	0.47	0	1

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
mmedu_2	Mother's mother's education==vocational	7214	0.22	0.42	0	1
mmedu_3	Mother's mother's education==o level	7214	0.18	0.39	0	1
mmedu_4	Mother's mother's education==a level	7214	0.20	0.40	0	1
mmedu_5	Mother's mother's education==degree	7214	0.06	0.24	0	1
mmpresent	Mother's mother present during own childhood	12609	0.90	0.30	0	1
mmteen	Mother's mother was a teen mum	11145	0.09	0.28	0	1
msibs	Mother's total no. of siblings	12559	2.21	1.59	0	15
msibs2	Mother's total no. of siblings (squared)	12559	7.42	12.71	0	225
msmoke_1	Mother smoked during pregnancy==none	12241	0.80	0.40	0	1
msmoke_2	Mother smoked during pregnancy==1-9	12241	0.08	0.27	0	1
msmoke_3	Mother smoked during pregnancy==10-19	12241	0.09	0.29	0	1
msmoke_4	Mother smoked during pregnancy==20+	12241	0.03	0.18	0	1
msoccat_d1	Mother's social class category==I	15209	0.04	0.19	0	1
msoccat_d2	Mother's social class category==II	15209	0.21	0.41	0	1
msoccat_d3	Mother's social class category==III non-manual	15209	0.28	0.45	0	1
msoccat_d4	Mother's social class category==III manual	15209	0.05	0.22	0	1
msoccat_d5	Mother's social class category==IV	15209	0.07	0.25	0	1
msoccat_d6	Mother's social class category==V	15209	0.01	0.12	0	1
msoccat_d7	Mother's social class category==missing	15209	0.16	0.37	0	1
msocnet12a	Mother's social network score at 12 weeks	12243	23.27	3.92	1	29
msocsupp12a	Mother's social support score	11578	19.59	5.09	0	30
pat32wk	Mother had a partner at 32wks	12135	0.98	0.15	0	1
pedu_1	Father's education==cse	10783	0.18	0.38	0	1
pedu_2	Father's education==vocational	10783	0.09	0.29	0	1
pedu_3	Father's education==o level	10783	0.24	0.43	0	1
pedu_4	Father's education==a level	10783	0.29	0.45	0	1
pedu_5	Father's education==degree	10783	0.202541	0.4019119	0	1
pethnic	Partner's ethnic group==non-white	12152	1.04	0.19	1	2
psoccat_d1	Father's social class category==I	15209	0.08	0.27	0	1
psoccat_d2	Father's social class category==II	15209	0.25	0.43	0	1
psoccat_d3	Father's social class category==III non-manual	15209	0.08	0.27	0	1
psoccat_d4	Father's social class category==III manual	15209	0.23	0.42	0	1
psoccat_d5	Father's social class category==IV	15209	0.07	0.26	0	1
psoccat_d6	Father's social class category==V	15209	0.02	0.14	0	1
psoccat_d7	Father's social class category==missing	15209	0.08	0.28	0	1

Proximal family processes

Mother characteristics

mschexp1	Mother's attitude to school : Good marks important	12406	0.69	0.46	0	1
----------	--	-------	------	------	---	---

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
mschexp2	Mother's attitude to school : Uselss to try, other kids better	12358	0.09	0.29	0	1
mschexp3	Mother's attitude to school : Mum ever expelled/suspended	12450	0.04	0.19	0	1
mschexp4	Mother's attitude to school : Mum often truant	9142	0.09	0.28	0	1
mschexp5	Mother's attitude to school : Mum liked school	9590	2.41	1.00	1	5
mschexp6	Mother's attitude to school : School valuable	9716	2.11	0.91	1	5
mloc12a	Mother's Locus of Control score at 12 wks	10694	4.37	2.17	0	12

Pre-school factors

Neighbourhood context

School context

Cognitive development

Affective and behavioural development

Health

bweightq_1	Child's birth weight==bottom quintile	15209	0.18	0.39	0	1
bweightq_2	Child's birth weight==2nd quintile	15209	0.19	0.40	0	1
bweightq_3	Child's birth weight==3rd quintile	15209	0.18	0.38	0	1
bweightq_4	Child's birth weight==4th quintile	15209	0.17	0.38	0	1
bweightq_5	Child's birth weight==top quintile	15209	0.18	0.38	0	1

asures between 6 months and 24 months

Distal family factors

Proximal family processes

ho6_book_1	Number of children's books in the home at 6 mths==none	11434	0.1893476	0.3918016	0	1
ho6_book_2	Number of children's books in the home at 6 mths==1-2	11434	0.27	0.44	0	1
ho6_book_3	Number of children's books in the home at 6 mths==3-9	11434	0.35	0.48	0	1
ho6_book_4	Number of children's books in the home at 6 mths==10+	11434	0.19	0.39	0	1
ho6_toys	Toy score at 6 mths	11478	3.65	1.10	0	5
pp6_actout	Activity score at 6 mths	11453	14.38	2.71	0	20
pp6_actout2	Activity score at 6 mths (squared)	11453	214.01	74.70	0	400
pp6_mcint	Mother-child interaction score at 6 mths	11485	10.38	1.72	0	12
pp6_pcint	Partner-child interaction score at 6 mths	11289	9.45	3.84	0	16
pp6_teach_1	Mother tries to teach child at 6 mths==yes freq not stated	11329	0.00	0.05	0	1
pp6_teach_2	Mother tries to teach child at 6 mths==no child too young	11329	0.27	0.44	0	1

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
pp6_teach_3	Mother tries to teach child at 6 mths==no don't have time	11329	0.01	0.08	0	1
pp6_teach_4	Mother tries to teach child at 6 mths==yes occassionally	11329	0.35	0.48	0	1
pp6_teach_5	Mother tries to teach child at 6 mths==yes often	11329	0.37	0.48	0	1
ho18_book_1	Number of children's books in the home at 18 mths==none		0.01	0.10	0	1
ho18_book_2	Number of children's books in the home at 18 mths==1-2		0.04	0.20	0	1
ho18_book_3	Number of children's books in the home at 18 mths==3-9		0.2826087	0.450288	0	1
ho18_book_4	Number of children's books in the home at 18 mths==10+		0.66	0.47	0	1
ho18_toys_1	Toy score at 18 mths==0.0		0.00	0.01	0	1
ho18_toys_2	Toy score at 18 mths==1.0		0.00	0.02	0	1
ho18_toys_3	Toy score at 18 mths==2.0		0.01	0.09	0	1
ho18_toys_4	Toy score at 18 mths==3.0		0.64	0.48	0	1
ho18_toys_5	Toy score at 18 mths==4.0		0.23	0.42	0	1
ho18_toys_6	Toy score at 18 mths==5.0		0.13	0.34	0	1
pp18_actout	Activity score at 18 mths		18.06176	3.223117	0	32
pp18_actout2	Activity score at 18 mths (squared)		336.6146	115.4452	0	1024
pp18_mcint	Mother-child interaction score at 18 mths		32.08653	4.321333	0	36
pp18_mcint2	Mother-child interaction score at 18 mths (squared)		1048.218	225.9409	0	1296
pp18_pcint	Partner-child interaction score at 18 mths		22.73	8.04	0	36
pp18_pcint2	Partner-child interaction score at 18 mths (squared)		581.31	309.18	0	1296
pp18_teach_1	Mother tries to teach child at 18 mths==no child too young		0.03	0.17	0	1
pp18_teach_2	Mother tries to teach child at 18 mths==no don't have time		0.00	0.06	0	1
pp18_teach_3	Mother tries to teach child at 18 mths==yes occassionally		0.37	0.48	0	1
pp18_teach_4	Mother tries to teach child at 18 mths==yes often		0.60	0.49	0	1
pp18_teach~g	Teaching score at 18 mths: No. things try to teach child		7.99	1.59	0	10
pp18_teach~g2	Teaching score at 18 mths: No. things try to teach child (squared)	11097	66.41	23.96	0	100
<u>Mother characteristics</u>						
<u>Pre-school factors</u>						
cc24_hrs	Childcare hours at 24 hrs	10319	11.28	16.89	0	160
cc24_type_1	Childcare type at 24 mths==none or relative/friend	10319	0.77	0.42	0	1
cc24_type_2	Childcare type at 24 mths==nanny/babysitter	10319	0.04	0.21	0	1
cc24_type_3	Childcare type at 24 mths==child-minder, day nursery/other	10319	0.19	0.39	0	1
<u>Neighbourhood context</u>						
<u>School context</u>						
<u>Cognitive development</u>						
cd6_comm	Communication score at 6 mths	10826	16.36	2.87	0	24

Table 3.3: p4 of 11

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
cd6_comm2	Communication score at 6 mths (squared)	10826	275.99	98.86	0	576
cd18_comm	Communication score at 18 mths	11102	15.94	4.79	0	28
cd18_comm2	Communication score at 18 mths (squared)	11102	276.88	156.90	0	784
<u>Affective and behavioural development</u>						
cd6_citact	Carey Infant Temperament Scale: Activity score	10503	40.12615	6.340193	13	60
cd6_citact2	Carey Infant Temperament Scale: Activity score (squared)	10503	1650.302	509.0616	169	3600
cd6_citada	Carey Infant Temperament Scale: Adaptability score	10508	13.99867	5.707122	0	42
cd6_citada2	Carey Infant Temperament Scale: Adaptability score (squared)	10508	228.5308	170.4181	0	1764
cd6_citapp	Carey Infant Temperament Scale: Approach score	10495	15.01363	6.3574	0	43
cd6_citapp2	Carey Infant Temperament Scale: Approach score (squared)	10495	265.8216	204.6066	0	1849
cd6_citdis	Carey Infant Temperament Scale: Distractability score	10506	14.02351	5.605372	0	36
cd6_citdis2	Carey Infant Temperament Scale: Distractability score (squared)	10506	228.0761	162.8162	0	1296
cd6_citint	Carey Infant Temperament Scale: Intensity score	10494	25.03202	5.587893	4	49
cd6_citint2	Carey Infant Temperament Scale: Intensity score (squared)	10494	657.8235	286.8131	16	2401
cd6_citmoo	Carey Infant Temperament Scale: Mood score	10496	15.81164	5.908438	0	44
cd6_citmoo2	Carey Infant Temperament Scale: Mood score (squared)	10496	284.9143	192.4852	0	1936
cd6_citper	Carey Infant Temperament Scale: Persistence score	10504	13.57178	5.107787	0	34
cd6_citper2	Carey Infant Temperament Scale: Persistence score (squared)	10504	210.2803	145.4396	0	1156
cd6_citrhy	Carey Infant Temperament Scale: Rhythmicity score	10500	15.96629	6.927516	0	46
cd6_citrhy2	Carey Infant Temperament Scale: Rhythmicity score (squared)	10500	302.9082	250.4323	0	2116
cd6_citthr	Carey Infant Temperament Scale: Threshold score	10490	27.50667	6.05676	3	50
cd6_citthr2	Carey Infant Temperament Scale: Threshold score (squared)	10490	793.2979	341.5554	9	2500
cd6_fmotor	Fine motor score at 6 mths	10471	21.09818	6.361875	0	33
cd6_fmotor2	Fine motor score at 6 mths (squared)	10471	485.6026	272.1465	0	1089
cd6_gmotor	Gross motor score at 6 mths	11078	16.32208	5.800315	0	39
cd6_gmotor2	Gross motor score at 6 mths (squared)	11078	300.0509	240.0972	0	1521
cd6_socdev	Social development score at 6 mths	10652	17.20597	4.840718	0	30
cd6_socdev2	Social development score at 6 mths (squared)	10652	319.4758	171.1499	0	900
cd18_socdev	Social development score at 18 mths	11087	19.14738	3.888408	0	28
cd18_socdev2	Social development score at 18 mths (squared)	11087	381.7405	146.0572	0	784
cd18_fmotor	Fine motor score at 18 mths	11090	26.68575	3.059099	0	32
cd18_fmotor2	Fine motor score at 18 mths (squared)	11090	721.4867	151.7445	0	1024
cd18_gmotor	Gross motor score at 18 mths	11087	19.44719	2.876727	0	22
cd18_gmotor2	Gross motor score at 18 mths (squared)	11087	386.468	94.27997	0	484
<u>Health</u>						

Measures between 30 months and 47 months

Distal family factors

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
hhtenure33_1	Housing tenure at 33m: owned/mortgaged	9708	0.78	0.41	0	1
hhtenure33_2	Housing tenure at 33m: council rented	9708	0.12	0.32	0	1
hhtenure33_3	Housing tenure at 33m: HA/council bought	9708	0.03	0.18	0	1
hhtenure33_4	Housing tenure at 33m: private rental/other	9708	0.07	0.25	0	1
income33_d1	Family income per week at 33 mths==less than £100	15209	0.05	0.22	0	1
income33_d2	Family income per week at 33 mths==£100-£199	15209	0.10	0.30	0	1
income33_d3	Family income per week at 33 mths==£200-£299	15209	0.17	0.37	0	1
income33_d4	Family income per week at 33 mths==£300-£399	15209	0.12	0.33	0	1
income33_d5	Family income per week at 33 mths==more than £400	15209	0.14	0.35	0	1
income33_d6	Family income per week at 33 mths==missing	15209	0.06	0.24	0	1
mempcat_1	Mother's employment category==full-time by 18m	10966	0.10	0.29	0	1
mempcat_2	Mother's employment category==part-time by 18m	10966	0.31	0.46	0	1
mempcat_3	Mother's employment category==by 18m (hours not known)	10966	0.10	0.31	0	1
mempcat_4	Mother's employment category==19-33m	10966	0.08	0.28	0	1
mempcat_5	Mother's employment category==not by 33m	10966	0.31	0.46	0	1
mempcat_6	Mother's employment category==not by 21m	10966	0.09	0.29	0	1
memploy33	Mother employed at 33 mths	9539	0.50	0.50	0	1
pat33m	Mother had partner at 33 mths	9717	0.94	0.24	0	1
famtype47	Family type at 47 mths==two adults	9422	0.10	0.30	0	1
income47_d1	Family income per week at 47 mths==less than £100	15209	0.04	0.21	0	1
income47_d2	Family income per week at 47 mths==£100-£199	15209	0.09	0.29	0	1
income47_d3	Family income per week at 47 mths==£200-£299	15209	0.15	0.36	0	1
income47_d4	Family income per week at 47 mths==£300-£399	15209	0.13	0.33	0	1
income47_d5	Family income per week at 47 mths==more than £400	15209	0.16	0.37	0	1
income47_d6	Family income per week at 47 mths==missing	15209	0.06	0.24	0	1
memploy47	Mother employed at 47 mths	8539	0.55	0.50	0	1
pat47m	Mother had partner at 47 mths	9570	0.94	0.24	0	1
<u>Proximal family processes</u>						
ho30_book	Few or no children's books in the home at 30 mths	10307	0.13	0.34	0	1
ho30_toys	Low toys score at 30 mths	10272	0.05	0.21	0	1
pp30_actout	Activities outside the home at 30 mths	10320	18.44	3.14	0	32
pp30_rwarmth	Reunion warmth score at 30 mths	10218	5.64	0.68	0	6
pp30_teach	Do you try to teach your child at 30m	10278	3.62	0.53	1	4
pp30_teach~g	Teaching score at 30 mths: No. things try to teach child	10302	6.32	1.08	0	7
mbondauth33	Intimate Bond measure at 33 mths: authority	8588	41.48	5.94	12	48
mbondwarm33	Intimate Bond measure at 33 mths: warmth	8754	21.86	8.45	12	48
mbondanop33	Intimate Bond at 33 mths: No partner	15209	0.04	0.20	0	1
mbondwnop33	Intimate Bond at 33 mths: No partner	15209	0.04	0.20	0	1
pp38_mcint	Mother-child interaction score at 38 mths	10166	24.42	3.42	0	27

Table 3.3: p6 of 11

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
pp38_pcint	Partner-child interaction score at 38 mths	10118	19.42	6.80	0	27
ho42_book	Number of children's books in the home at 42 mths	10026	3.93	0.28	1	4
ho42_toys	Toy score at 42 mths	10030	5.20	0.62	1	6
pp42_actout	Activities outside the home at 42 mths	10024	18.65	2.93	1	29
pp42_disvar	Discipline variation at 42 mths	9453	10.20	3.71	0	20
pp42_mcint	Mother-child interaction score at 42 mths	10090	28.43	5.38	0	36
pp42_ocint	Other-child interaction score at 42 mths	9902	8.71	9.99	0	36
pp42_pcint	Partner-child interaction score at 42 mths	10041	20.01	8.24	0	36
pp42_rwarmth	Activities outside the home at 42 mths	9913	5.68	0.64	0	6
pp42_teach	Do you try to teach your child at 42 mths	10014	3.57	0.54	1	4
pp42_teach~g	Teaching score at 42 mths: No. things try to teach child	9888	6.68	0.71	1	7
<u>Mother characteristics</u>						
mcceianx33	Mother's CCEI score at 33 mths: anxiety	9573	4.71	3.58	0	16
mcceianx332	Mother's CCEI score at 33 mths: anxiety (squared)	9573	35.05	49.22	0	256
mcceidep33	Mother's CCEI score at 33 mths: depression	9545	4.21	3.22	0	16
mcceidep332	Mother's CCEI score at 33 mths: depression (squared)	9545	28.0989	39.53554	0	256
mcceisom33	Mother's CCEI score at 33 mths: somatic	9611	2.873686	2.079198	0	14
mcceisom332	Mother's CCEI score at 33 mths: somatic (squared)	9611	12.58	20.56	0	196
mepds33	Mother's EPDS score at 33 mths	9595	6.27	5.04	0	30
mselfe33	Mother's self esteem at 33 mths	9478	29.34881	6.357992	0	40
mselfe332	Mother's self esteem at 33 mths (squared)	9478	901.77	358.94	0	1600
<u>Pre-school factors</u>						
<u>Neighbourhood context</u>						
<u>School context</u>						
<u>Cognitive development</u>						
cd38_is	dv: intelligibility score	10016	5.61	0.88	0	6
cd38_mcpast	dv: past tense score	9981	33.40	10.51	0	42
cd38_mcplu	dv: plurals score	10027	10.14	2.26	0	12
cd38_mcvocab	dv: vocabulary score	10137	228.17	32.67	0	246
cd38_mcwcomb	dv: word combination score	9959	21.98	5.10	0	26
<u>Affective and behavioural development</u>						
cd30_fmotor	Fine motor score at 30 mths	10292	26.36932	3.755723	0	32
cd30_fmotor2	Fine motor score at 30 mths (squared)	10292	709.4449	183.8169	0	1024
cd30_fpsai	Pre-school Activities Inventory at 30 mths: Female play	9839	30.60687	7.757672	11	54

Table 3.3: p7 of 11

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
cd30_fpsai2	Pre-school Activities Inventory at 30 mths: Female play (squared)	9839	996.9559	489.4213	121	2916
cd30_mpsai	Pre-school Activities Inventory at 30 mths: Male play	9958	36.61709	7.016509	12	60
cd30_mpsai2	Pre-school Activities Inventory at 30 mths: Male play (squared)	9958	1390.038	525.3332	144	3600
cd30_gmotor	Gross motor score at 30 mths	10278	19.51304	2.516342	0	22
cd30_gmotor2	Gross motor score at 30 mths (squared)	10278	387.09	87.23116	0	484
cd38_easact	EAS Temperament: Activity score	9927	21.61207	3.127441	5	25
cd38_easem	EAS Temperament: Emotionality score	9822	12.44869	4.235382	5	25
cd38_easshy	EAS Temperament: Shyness score	9918	12.45937	4.086998	5	25
cd38_eassoc	EAS Temperament: Sociability score	9706	18.17711	3.087936	5	25
cd42_fmotor	Fine motor score at 42 mths	10025	29.71691	4.060061	0	34
cd42_fpsai	Pre-school Activities Inventory at 42 mths: Female play	10038	35.83622	9.256223	12	60
cd42_gmotor	Gross motor score at 42 mths	10032	26.35437	3.58587	0	30
cd42_mpsai	Pre-school Activities Inventory at 42 mths: Male play	10038	37.5784	7.751764	12	60
cd42_rrcond	Revised Rutter Behaviour Scale: Conduct difficulties score	10024	3.634976	2.368256	0	16
cd42_rrrem	Revised Rutter Behaviour Scale: Emotional difficulties score	10024	2.555567	1.747136	0	12
cd42_rrhyp	Revised Rutter Behaviour Scale: Hyperactivity score	10024	2.638168	1.814357	0	8
cd42_rrpro	Revised Rutter Behaviour Scale: Prosocial score	10024	15.32572	3.614844	0	22
cd42_socdev	Revised Rutter Behaviour Scale: Social development score	10020	22.0984	3.14117	0	26

Health

Measures at 60 months and 84 months

Distal family factors

Proximal family processes

Mother characteristics

Pre-school factors

Neighbourhood context

School context

Cognitive development

entlang	Entry Assessment: Language	9836	3.54	1.10	0	5
entmath	Entry Assessment: Maths	9837	3.25	1.10	0	5
entread	Entry Assessment: Reading	9839	3.05	0.89	0	5
entwrit	Entry Assessment: Writing	9839	2.88	0.88	0	5

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
k1maths	KS1: Maths score	11643	3.23	1.37	0	5
k1read	KS1: Reading points score	11647	3.27	1.51	0	5
k1spell	KS1: Spelling score	9871	1.08	0.65	0	2
k1write	KS1: Writing points score	11647	2.62	1.25	0	5
<u>Affective and behavioural development</u>						
entlmotor	Entry Assessment: Large Motor	3399	3.302442	1.02055	0	5
entprob	Entry Assessment: Problem solving	3001	3.393869	1.00473	0	5
entsmotor	Entry Assessment: Small Motor	3804	3.410095	1.158291	0	5
entsoc	Entry Assessment: Social Skills	5209	3.393934	1.077066	0	5
<u>Health</u>						
Measures between 90 months and 114 months						
<u>Distal family factors</u>						
<u>Proximal family processes</u>						
<u>Mother characteristics</u>						
<u>Pre-school factors</u>						
<u>Neighbourhood context</u>						
<u>School context</u>						
<u>Cognitive development</u>						
f7_letter	Shapes game score: word - focus at 7	587	34.20	5.71	0	40
f7_phoneme	Phoneme task score - focus at 7	8125	20.00	9.56	0	40
f7_read	Reading score - focus at 7	8145	28.11	9.44	0	52
f7_spella	Spelling score (no. correct) - focus at 7	8024	7.73	4.44	0	15
f7_spellb	Spelling score - focus at 7	8024	25.81	12.70	0	45
f8_conatta	Attention: Opp worlds task - normative 1 - focus at 8	7214	18.53	1.46	1	19.00
f8_conattb	Attention: Opp worlds task - normative 2 - focus at 8	7208	18.24	1.70	1	19
f8_divatt	Attention: Dual task - focus at 8	5557	0.88	0.26	0	1
f8_listcomp	Speech & Language: Wold comprehension - focus at 8	7383	7.46	1.96	2	15
f8_selatt	Attention: Sky search - focus at 8	7186	5.21	1.92	0	46.58333
f8_stmem	Short Term Memory - focus at 8	7367	7.23	2.51	0	12

Table 3.3: p9 of 11

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
f8_wiscperf	WISC: Performance IQ - focus at 8	6815	99.74	17.02	46	151
f8_wiscTOT	WISC: Total IQ - focus at 8	6751	104.24	16.40	48	151
f8_wiscverb	MISC: Verbal IQ - focus at 8	7300	106.99	16.76	46	155
f9_readnw	Nonword reading score - focus at 9	7589	5.20	2.50	0	10
f9_readrw	Reading score - focus at 9	7603	7.50	2.49	0	10
f9_sentence	Sentence decision, overall - focus at 9	600	38.71	0.96	20	39
f9_spell	Spelling score - focus at 9	7586	10.20	3.49	0	15
f9_storyac~r	Accuracy: standardised score - focus at 9	6888	103.69	13.67	69	131
f9_storycomp	Comprehension - standardised - focus at 9	6888	99.92	11.89	69	131
f9_storyrate	Words read per minute - focus at 9	6869	105.14	12.56	69	131
<u>Affective and behavioural development</u>						
f7_balance	Heel to toe, co-ordination test 1 - focus at 7	7313	13.7994	2.0802	0	15
f7_pegA	Peg game: coordination test 3 - focus at 7	7264	22.5424	3.984151	14	72
f7_pegb	Peg game: coordination test 4 - focus at 7	7120	25.73511	5.582874	14	149
f7_string	String game: coordination test 5 - focus at 7	6666	23.31923	8.200787	9	118
f7_throw	Bean bag game: coordination test 6 - focus at 7	7322	5.822999	2.075498	0	10
f8_antisoc	Antisocial activities score - focus at 8	7125	0.3601404	0.8530324	0	11
f8_friends	Friends score - focus at 8	7136	3.44745	2.437625	0	15
f8_locscore	Locus of control mean - focus at 8	7150	49.51809	17.57396	0	100
f8_loctype	Locus of control summary score - focus at 8	7150	1.998741	0.7063634	1	3
f8_psaI	Gender - cai score - focus at 8	7062	49.85095	15.52165	3.45	81.35
f8_scholcomp	Self esteem: scholastic - focus at 8	6952	16.97756	3.6758	6	24
f8_selfworth	Self esteem: global self worth - focus at 8	6941	19.22605	3.426552	6	24
f9_selfmath	Self perception maths score - focus at 8	7478	23.74418	5.456034	8	32
f9_selfread	Self perception reading score - focus at 8	7479	24.5129	4.751578	8	32
<u>Health</u>						
f7_bmiq_1	Focus at 7: bmi: Bottom quintile	8210	0.20	0.40	0	1
f7_bmiq_2	Focus at 7: bmi: 2nd quintile	8210	0.20	0.40	0	1
f7_bmiq_3	Focus at 7: bmi: 3rd quintile	8210	0.20	0.40	0	1
f7_bmiq_4	Focus at 7: bmi: 4th quintile	8210	0.20	0.40	0	1
f7_bmiq_5	Focus at 7: bmi: Top quintile	8210	0.20	0.40	0	1
f7_heightq_1	Focus at 7: Height: Bottom quintile	8224	0.21	0.40	0	1
f7_heightq_2	Focus at 7: Height: 2nd quintile	8224	0.20	0.40	0	1
f7_heightq_3	Focus at 7: Height: 3rd quintile	8224	0.20	0.40	0	1
f7_heightq_4	Focus at 7: Height: 4th quintile	8224	0.20	0.40	0	1
f7_heightq_5	Focus at 7: Height: Top quintile	8224	0.20	0.40	0	1
f7_weightq_1	Focus at 7: Weight: Bottom quintile	8211	0.20	0.40	0	1
f7_weightq_2	Focus at 7: Weight: 2nd quintile	8211	0.22	0.41	0	1

Table 3.3: p10 of 11

Variable name	Variable	Obs	Mean	Std. Dev.	Min	Max
f7_weightq_3	Focus at 7: Weight: 3rd quintile	8211	0.18	0.39	0	1
f7_weightq_4	Focus at 7: Weight: 4th quintile	8211	0.20	0.40	0	1
f7_weightq_5	Focus at 7: Weight: Top quintile	8211	0.19	0.40	0	1

Table 3.3: p11 of 11

Table 4: Number of constructs, each dataset, by age and domain

	<i>NCDS (1958 Cohort)</i>				<i>BCS (1970 Cohort)</i>						<i>ALSPAC</i>				
	Age 0	Age 7	Age 11	Age 16	Age 0	22 months	42 months	Age 5	Age 10	Age 16	Age 0	6-24m	30-47m	60, 84m	90, 114m
Distal family factors	12	14	8	9	10	2	2	12	12	1	23	0	9	0	0
Proximal family processes	2	11	6	6	0	0	0	6	7	0	0	13	22	0	0
Mother characteristics											2	0	5	0	0
Pre-school factors											0	2	0	0	0
Neighbourhood context	0	0	1	0	0	0	0	2	0	0					
School context	0	10	2	15	0	0	0	0	5	0					
Cognitive development	0	9	10	10	0	4	3	7	4	0	0	2	5	8	12
Affective and behavioural development	0	15	19	16	0	0	1	3	14	12	0	15	17	4	4
Health	1	8	8	6	1	1	1	1	1	0	1	1	2	0	0

Table 5: Summary stats for binary outcomes

<i>NCDS (1958Cohort)</i>			<i>BCS (1970 Cohort)</i>			<i>ALSPAC</i>		
outcome	N	M	outcome	N	M	outcome	N	M
Assessed at age 42			Assessed at age 30			Assessed at age 10/11		
1 Smoker	11375	0.26	1 Smoker	11170	0.37	1 Bottom quartile, KS2 English	12190	0.24
2 Obese	11419	0.18	2 Obese	11262	0.14	2 Bottom quartile, KS2 Maths	12207	0.24
3 Depression	11281	0.18	3 Depression	11082	0.17	3 Bottom quartile, KS2 Science	12335	0.24
4 Psychiatric disturbance	11270	0.21	4 Psychiatric disturbance	11071	0.20	4 Bottom quartile, KS2 E, M & S	12008	0.12
5 Single, separated, divorced	11383	0.29	5 Single, separated, divorced	11121	0.33	5 High anti-social score, age 10	7433	0.16
6 Teen parent	11419	0.10	6 Teen parent	18779	0.03	6 High depression score, age 10	7324	0.10
7 Single parent	11418	0.03	7 Single parent	11226	0.06	Multiple outcomes		
8 Social housing	11350	0.12	8 Social housing	11115	0.15	7 >2 of 5 outcomes	6625	0.20
9 Homeless	6044	0.06	9 Homeless	10219	0.07			
10 Victim	11419	0.03	10 Victim	11262	0.07			
11 Dis-satisfied	11269	0.16	11 Dis-satisfied	11070	0.15			
12 No Qualifications	11233	0.10	12 No Qualifications	11188	0.28			
13 Not Level 2 Qualifications	11233	0.22	13 Not Level 2 Qualifications	11188	0.36			
14 Not Level 4 Qualifications	11233	0.69	14 Not Level 4 Qualifications	11188	0.80			
15 Low level criminality	11273	0.07	15 Low level criminality	11073	0.23			
16 High level criminality	11273	0.01	16 High level criminality	11073	0.05			
17 Racial intolerance	11279	0.06	17 Racial intolerance	11081	0.05			
18 Not voting	11351	0.23	18 Not voting	11183	0.38			
19 Low self-efficacy	11281	0.27	19 Low self-efficacy	11117	0.25			
20 Low hourly wage	7959	0.19	20 Low hourly wage	8067	0.19			
21 On benefits	11388	0.08	21 On benefits	11191	0.10			
22 Workless household	11382	0.08	22 Workless household	11226	0.10			
23 Workless household w children	11378	0.05	23 Workless household w children	11226	0.05			
24 Absent child	11419	0.01	24 Absent child	18779	0.02			
25 Mental health problem	11388	0.25	25 Mental health problem	11227	0.24			
26 Depression	11388	0.22	26 Depression	11227	0.21			
27 Other mental health issue	11388	0.10	27 Other mental health issue	11227	0.09			
28 Addition to drug/alcohol	11388	0.02	28 Addition to drug/alcohol	11227	0.02			
29 NEET18	14675	0.10	29 NEET18	11262	0.07			
30 NEET30	14675	0.10	30 NEET30	11262	0.10			
31 Overcrowding	11390	0.07	31 Overcrowding	18744	0.04			
32 Financial problem	11388	0.08	32 Financial problem	11227	0.07			
Assessed at age 23			Multiple outcomes					
33 Age 23: Unskilled or semi-skilled	10124	0.21	33 > 1 of 5 key outcomes	11112	0.15			
34 Age 23: Literacy or numeracy prob	12537	0.13	34 > 8 of 32 outcomes	18779	0.11			
35 Had child before 20	12535	0.08						
36 Had child before 23	12535	0.26						
37 Teen parent before 23	12535	0.04						
38 Depressed at 23	12537	0.08						
39 Low hourly wage at 23	6750	0.14						
Multiple outcomes								
40 >2 of 6 age 23 outcomes	10123	0.09						
41 >7 of 32 age 42 outcomes	14675	0.12						

Table 6.1: Identification of those with high cost/harm outcomes, up to age of measure (i.e using previous measures), 1958 Cohort

		<i>Dropping cases with missing data, upper bound</i>					<i>Imputing missing data, lower bound</i>					
outcome	age	No. constructs	N	TP	FP	ratio	age	No. constructs	N	TP	FP	ratio
1 Smoker	0	14	5449	0.342	0.221	1.551	0	14	11373	0.364	0.222	1.643
	7	66	2422	0.465	0.175	2.660	7	63	9485	0.424	0.198	2.141
	11	108	1259	0.527	0.132	3.991	11	107	7417	0.460	0.186	2.476
	16	161	106				16	155	5578	0.510	0.165	3.088
2 Obese	0	12	6719	0.246	0.161	1.530	0	14	11417	0.232	0.167	1.388
	7	58	3468	0.404	0.118	3.420	7	56	9501	0.373	0.136	2.749
	11	103	1511	0.534	0.079	6.767	11	98	7493	0.469	0.113	4.153
	16	153	120				16	143	5634	0.502	0.106	4.717
3 Depression	0	13	6054	0.242	0.146	1.656	0	14	11279	0.266	0.157	1.694
	7	78	2288	0.344	0.114	3.006	7	79	9412	0.346	0.139	2.481
	11	131	1044	0.547	0.076	7.187	11	132	7364	0.379	0.129	2.929
	16	187	93				16	189	5543	0.423	0.118	3.578
4 Psychiatric disturbance	0	14	4924	0.268	0.179	1.493	0	14	11268	0.274	0.191	1.437
	7	76	1721	0.359	0.160	2.246	7	78	9407	0.316	0.179	1.764
	11	128	831	0.571	0.109	5.262	11	130	7362	0.354	0.168	2.108
	16	184	81				16	188	5542	0.386	0.157	2.459
5 Single, separated, divorced	0	4	10164	0.310	0.269	1.155	0	4	11381	0.320	0.290	1.104
	7	28	5760	0.388	0.250	1.552	7	28	10129	0.372	0.258	1.446
	11	57	3616	0.415	0.227	1.826	11	52	7982	0.389	0.245	1.587
	16	91	413	0.636	0.115	5.551	16	77	6078	0.415	0.231	1.799
6 Teen parent	0	12	6115	0.262	0.080	3.283	0	12	11417	0.254	0.087	2.906
	7	70	2385	0.402	0.057	6.997	7	71	9501	0.349	0.074	4.695
	11	122	1106	0.593	0.032	18.406	11	122	7429	0.398	0.066	6.015
	16	171	119				16	173	5586	0.434	0.061	7.102
7 Single parent	0	4	8729	0.062	0.025	2.508	0	7	11416	0.063	0.025	2.488
	7	39	4178	0.144	0.023	6.171	7	38	10152	0.075	0.025	3.022
	11	71	2488	0.211	0.023	9.118	11	72	8684	0.133	0.024	5.566
	16	96	951	0.448	0.017	25.832	16	102	8684	0.142	0.024	5.985
8 Social housing	0	13	6081	0.258	0.089	2.912	0	14	11348	0.264	0.100	2.639

Dropping cases with missing data, upper bound

Imputing missing data, lower bound

outcome	<i>Dropping cases with missing data, upper bound</i>						<i>Imputing missing data, lower bound</i>					
	age	No. constructs	N	TP	FP	ratio	age	No. constructs	N	TP	FP	ratio
	7	71	2364	0.421	0.060	7.045	7	73	9464	0.385	0.080	4.816
	11	123	1142	0.624	0.030	20.596	11	123	7402	0.423	0.073	5.802
	16	175	109				16	176	5570	0.478	0.065	7.354
9 Homeless	0	11	3516	0.099	0.058	1.714	0	11	6043	0.092	0.059	1.542
	7	52	1696	0.284	0.046	6.208	7	53	5015	0.195	0.053	3.697
	11	92	890	0.541	0.034	16.017	11	96	3905	0.262	0.046	5.699
	16	126	201				16	132	2910	0.319	0.041	7.753
10 Victim	0	6	7439	0.075	0.024	3.138	0	5	11417	0.076	0.026	2.907
	7	42	3469	0.163	0.020	8.220	7	40	9501	0.130	0.025	5.259
	11	81	1961	0.308	0.019	16.316	11	80	7588	0.176	0.023	7.672
	16	111	374				16	111	5773	0.215	0.022	9.744
11 Dis-satisfied	0	14	5076	0.211	0.140	1.509	0	15	11267	0.229	0.145	1.578
	7	74	2050	0.305	0.108	2.839	7	76	9405	0.281	0.132	2.136
	11	123	1013	0.558	0.059	9.407	11	125	7359	0.313	0.121	2.580
	16	179	91				16	178	5540	0.364	0.112	3.246
12 No Qualifications	0	11	6629	0.263	0.076	3.483	0	12	11231	0.261	0.085	3.070
	7	70	2565	0.473	0.040	11.730	7	71	9359	0.453	0.061	7.385
	11	122	1170	0.646	0.021	31.043	11	121	7331	0.508	0.053	9.554
	16	169	167				16	171	5518	0.602	0.040	14.923
13 Not Level 2 Qualifications	0	15	4897	0.373	0.170	2.195	0	15	11231	0.387	0.175	2.218
	7	80	1691	0.592	0.084	7.030	7	80	9359	0.564	0.122	4.635
	11	134	766	0.745	0.043	17.465	11	132	7331	0.608	0.105	5.783
	16	193	72				16	190	5518	0.676	0.084	8.076
14 Not Level 4 Qualifications	0	8	5440	0.781	0.498	1.569	0	8	11231	0.775	0.500	1.551
	7	35	3082	0.813	0.402	2.019	7	30	9359	0.804	0.429	1.873
	11	62	1727	0.801	0.347	2.312	11	53	7404	0.819	0.394	2.077
	16	100	146				16	87	5639	0.840	0.342	2.453
15 Low level criminality	0	11	6650	0.169	0.056	3.018	0	13	11271	0.201	0.064	3.160
	7	71	2372	0.348	0.047	7.482	7	72	9407	0.268	0.056	4.773
	11	122	1088	0.468	0.042	11.252	11	121	7360	0.291	0.053	5.462
	16	165	167				16	169	5540	0.347	0.048	7.163

Table 6.1: p2 of 6

Dropping cases with missing data, upper bound

Imputing missing data, lower bound

outcome	age	No. constructs	N	TP	FP	ratio	age	No. constructs	N	TP	FP	ratio
16 High level criminality	0	3	10080	0.043	0.009	4.992	0	3	11271	0.067	0.009	7.596
	7	28	6069	0.143	0.007	20.476	7	27	10037	0.090	0.008	11.040
	11	55	3428	0.240	0.006	42.985	11	55	8596	0.154	0.008	19.855
	16	81	1371	0.500	0.002	227.500	16	82	8495	0.169	0.008	22.207
17 Racial intolerance	0	10	6769	0.088	0.057	1.552	0	10	11277	0.082	0.059	1.390
	7	54	3036	0.206	0.046	4.516	7	53	10042	0.142	0.056	2.544
	11	92	1805	0.313	0.039	8.092	11	91	8524	0.192	0.051	3.775
	16	126	284				16	124	8425	0.211	0.050	4.225
18 Not voting	0	15	4944	0.310	0.203	1.528	0	15	11349	0.295	0.201	1.466
	7	75	1766	0.449	0.163	2.758	7	74	9469	0.369	0.186	1.983
	11	119	857	0.590	0.115	5.130	11	119	7468	0.398	0.175	2.273
	16	175	81				16	175	5618	0.435	0.161	2.706
19 Low self-efficacy	0	14	4924	0.354	0.227	1.556	0	13	11279	0.358	0.240	1.487
	7	77	1706	0.476	0.162	2.942	7	74	9412	0.416	0.211	1.972
	11	124	857	0.611	0.103	5.910	11	118	7435	0.436	0.195	2.234
	16	180	79				16	175	5594	0.471	0.181	2.602
20 Low hourly wage	0	13	4298	0.279	0.163	1.719	0	15	7957	0.289	0.165	1.745
	7	76	1809	0.444	0.108	4.100	7	80	6649	0.402	0.138	2.923
	11	128	836	0.622	0.058	10.730	11	132	5233	0.437	0.125	3.490
	16	184	73				16	188	3961	0.494	0.113	4.383
21 On benefits	0	11	6707	0.183	0.064	2.870	0	12	11386	0.173	0.076	2.292
	7	70	2600	0.343	0.047	7.230	7	68	9494	0.296	0.061	4.817
	11	122	1257	0.576	0.021	27.615	11	116	7424	0.326	0.055	5.923
	16	168	180				16	167	5584	0.364	0.050	7.251
22 Workless household	0	11	6704	0.156	0.063	2.482	0	12	11380	0.157	0.073	2.154
	7	70	2594	0.277	0.044	6.333	7	69	9489	0.278	0.060	4.653
	11	121	1255	0.559	0.022	25.729	11	117	7420	0.293	0.055	5.372
	16	164	180				16	165	5581	0.337	0.049	6.834
23 Workless household w children	0	9	6703	0.138	0.036	3.846	0	9	11376	0.115	0.043	2.689
	7	61	2642	0.213	0.029	7.326	7	60	9476	0.207	0.038	5.520

Table 6.1: p3 of 6

Dropping cases with missing data, upper bound

Imputing missing data, lower bound

outcome	<i>Dropping cases with missing data, upper bound</i>						<i>Imputing missing data, lower bound</i>					
	age	No. constructs	N	TP	FP	ratio	age	No. constructs	N	TP	FP	ratio
	11	112	1277	0.500	0.015	34.472	11	108	7413	0.219	0.036	6.140
	16	153	215				16	148	5577	0.297	0.030	9.753
24 Absent child	0	4	9334	0.058	0.012	4.717	0	4	11417	0.063	0.012	5.181
	7	25	5361	0.102	0.010	10.173	7	27	10153	0.114	0.011	10.474
	11	52	3626	0.190	0.009	20.078	11	53	8156	0.156	0.010	15.548
	16	74	1493	0.412	0.007	60.776	16	77	8060	0.158	0.010	15.720
25 Mental health problem	0	15	4960	0.377	0.203	1.858	0	15	11386	0.354	0.218	1.620
	7	81	1715	0.471	0.166	2.831	7	82	9494	0.392	0.205	1.909
	11	135	776	0.632	0.106	5.946	11	136	7424	0.427	0.192	2.221
	16	197	72				16	198	5584	0.460	0.183	2.519
26 Depression	0	15	4960	0.341	0.173	1.965	0	15	11386	0.321	0.188	1.710
	7	79	1715	0.447	0.141	3.164	7	80	9494	0.357	0.177	2.013
	11	129	826	0.622	0.094	6.641	11	130	7424	0.398	0.163	2.435
	16	187	75				16	187	5584	0.438	0.153	2.861
27 Other mental health issue	0	14	4961	0.175	0.085	2.060	0	13	11386	0.166	0.092	1.811
	7	71	1921	0.330	0.070	4.699	7	72	9494	0.208	0.086	2.415
	11	121	902	0.591	0.044	13.361	11	121	7424	0.277	0.079	3.508
	16	169	121				16	168	5584	0.306	0.075	4.085
28 Addition to drug/alcohol	0	3	9477	0.040	0.016	2.538	0	5	11386	0.049	0.016	3.096
	7	35	4918	0.125	0.012	10.835	7	35	9494	0.107	0.014	7.438
	11	64	2575	0.290	0.009	33.572	11	65	7651	0.156	0.014	11.384
	16	93	1169	0.500	0.007	72.063	16	94	5749	0.184	0.013	14.666
29 NEET18	0	11	8492	0.200	0.079	2.543	0	13	14673	0.216	0.088	2.441
	7	71	3208	0.346	0.054	6.446	7	73	12110	0.288	0.076	3.808
	11	123	1456	0.495	0.035	14.028	11	125	9356	0.339	0.068	5.004
	16	171	195				16	177	6973	0.370	0.063	5.874
30 NEET30	0	11	8492	0.241	0.075	3.191	0	13	14673	0.238	0.083	2.879
	7	72	3208	0.380	0.054	7.003	7	73	12110	0.318	0.075	4.246
	11	124	1456	0.526	0.040	13.080	11	125	9356	0.357	0.070	5.136
	16	174	194				16	177	6973	0.412	0.064	6.464

Table 6.1: p4 of 6

Dropping cases with missing data, upper bound

Imputing missing data, lower bound

outcome		age	No. constructs	N	TP	FP	ratio			age	No. constructs	N	TP	FP	ratio
31	Overcrowding	0	8	6734	0.117	0.056	2.074	0	9	11388	0.117	0.061	1.911		
		7	53	3035	0.244	0.043	5.647	7	51	9485	0.178	0.055	3.251		
		11	103	1363	0.533	0.027	19.627	11	93	7478	0.238	0.051	4.641		
		16	140	217				16	132	5625	0.280	0.048	5.861		
32	Financial problem	0	9	6709	0.123	0.071	1.735	0	12	11386	0.127	0.077	1.658		
		7	65	2745	0.223	0.054	4.125	7	65	9494	0.206	0.068	3.028		
		11	115	1347	0.444	0.036	12.504	11	111	7424	0.240	0.063	3.805		
		16	157	200				16	152	5584	0.308	0.058	5.270		
33	Age 23: Unskilled or semi-skilled	0	14	4550	0.347	0.169	2.050	0	14	10124	0.338	0.174	1.945		
		7	76	1644	0.514	0.105	4.891	7	76	8423	0.470	0.135	3.477		
		11	125	761	0.667	0.061	11.009	11	126	6609	0.482	0.129	3.748		
		16	179	59				16	177	4968	0.520	0.120	4.345		
34	Age 23: Literacy or numeracy problerr	0	12	7369	0.211	0.110	1.911	0	13	12537	0.233	0.118	1.972		
		7	76	2764	0.438	0.064	6.879	7	76	10412	0.435	0.082	5.274		
		11	128	1283	0.500	0.047	10.664	11	127	8148	0.484	0.073	6.633		
		16	177	164				16	181	6141	0.514	0.068	7.614		
35	Had child before 20	0	11	7370	0.208	0.055	3.759	0	12	12535	0.231	0.063	3.657		
		7	67	2872	0.350	0.043	8.209	7	68	10411	0.313	0.054	5.771		
		11	118	1362	0.582	0.022	26.922	11	117	8148	0.357	0.048	7.434		
		16	158	197				16	164	6141	0.410	0.042	9.671		
36	Had child before 23	0	15	5482	0.441	0.184	2.397	0	15	12535	0.446	0.190	2.344		
		7	82	1892	0.515	0.143	3.601	7	82	10411	0.503	0.167	3.015		
		11	136	886	0.615	0.099	6.189	11	136	8148	0.542	0.153	3.550		
		16	198	79				16	198	6141	0.576	0.141	4.076		
37	Teen parent before 23	0	9	7370	0.179	0.030	5.880	0	8	12535	0.145	0.037	3.911		
		7	60	2900	0.385	0.023	16.803	7	56	10411	0.254	0.031	8.253		
		11	111	1357	0.512	0.015	33.702	11	105	8228	0.346	0.027	12.758		
		16	152	219				16	147	6202	0.405	0.024	17.136		
38	Depressed at 23	0	11	7370	0.176	0.059	2.975	0	13	12537	0.182	0.067	2.729		
		7	68	2858	0.345	0.033	10.318	7	70	10412	0.285	0.058	4.944		
		11	119	1331	0.536	0.020	26.271	11	120	8148	0.315	0.054	5.814		

Table 6.1: p5 of 6

Dropping cases with missing data, upper bound

Imputing missing data, lower bound

outcome	<i>Dropping cases with missing data, upper bound</i>						<i>Imputing missing data, lower bound</i>					
	age	No. constructs	N	TP	FP	ratio	age	No. constructs	N	TP	FP	ratio
	16	164	186				16	170	6141	0.370	0.048	7.708
39 Low hourly wage at 23	0	13	3682	0.194	0.135	1.437	0	13	6750	0.216	0.135	1.603
	7	77	1390	0.387	0.081	4.742	7	76	5679	0.344	0.110	3.123
	11	128	644	0.686	0.038	17.891	11	126	4467	0.386	0.099	3.910
	16	185	58				16	183	3373	0.416	0.092	4.541
40 >2 of 6 age 23 outcomes	0	11	5906	0.256	0.060	4.246	0	12	10123	0.232	0.073	3.184
	7	72	2234	0.419	0.038	11.130	7	69	8423	0.375	0.057	6.625
	11	124	1062	0.708	0.014	51.304	11	119	6609	0.411	0.052	7.951
	16	171	137				16	171	4968	0.476	0.045	10.568
41 >7 of 32 age 42 outcomes	0	13	7725	0.212	0.094	2.268	0	14	14673	0.228	0.105	2.164
	7	74	2938	0.304	0.068	4.446	7	75	12110	0.310	0.091	3.415
	11	126	1414	0.504	0.043	11.719	11	127	9356	0.339	0.086	3.956
	16	180	127				16	182	6973	0.402	0.076	5.264

Table 6.1: p6 of 6

Table 6.2: Identification of those with high cost/harm outcomes, up to age of measure (i.e using previous measures), 1970 Cohort

		<i>Dropping cases with missing data, upper bound</i>							<i>Imputing missing data, lower bound</i>						
outcome		age	No. constructs	N	M	TP	FP	ratio	age	No. constructs	N	TP	FP	ratio	
1 Smoker		0	10	8088	0.363996	0.460	0.309	1.5	0	10	9436	0.463	0.314	1.5	
		5	40	4539	0.356026	0.491	0.281	1.7	5	38	5877	0.476	0.287	1.7	
		10	79	757	0.321004	0.642	0.169	3.8	10	76	5327	0.517	0.261	2.0	
		16	89	757	0.321004	0.720	0.132	5.4	16	88	5327	0.579	0.227	2.5	
2 Obese		0	4	9488	0.138175	0.188	0.131	1.4	0	4	9489	0.188	0.131	1.4	
		5	21	4929	0.133699	0.226	0.119	1.9	5	22	5948	0.228	0.120	1.9	
		10	49	1143	0.093613	0.449	0.057	7.9	10	49	4986	0.295	0.111	2.7	
		16	58	605	0.097521	0.678	0.035	19.5	16	55	4986	0.297	0.110	2.7	
3 Depression		0	10	8012	0.168248	0.260	0.149	1.7	0	8	9304	0.251	0.152	1.7	
		5	36	4556	0.166155	0.304	0.139	2.2	5	36	5817	0.299	0.136	2.2	
		10	70	850	0.135294	0.496	0.079	6.3	10	69	3636	0.363	0.118	3.1	
		16	83	452	0.119469	0.852	0.020	42.4	16	81	3636	0.387	0.113	3.4	
4 Psychiatric disturbance		0	5	10134	0.201993	0.255	0.188	1.4	0	6	10135	0.253	0.189	1.3	
		5	30	4885	0.201024	0.313	0.173	1.8	5	30	6298	0.302	0.172	1.8	
		10	59	728	0.201923	0.537	0.117	4.6	10	58	3923	0.367	0.159	2.3	
		16	71	361	0.207756	0.827	0.045	18.2	16	70	3923	0.370	0.158	2.3	
5 Single, separated, divorced		0	8	8077	0.326359	0.372	0.295	1.3	0	8	9381	0.362	0.292	1.2	
		5	33	4978	0.315388	0.421	0.267	1.6	5	31	5939	0.403	0.276	1.5	
		10	73	570	0.363158	0.671	0.187	3.6	10	72	5138	0.451	0.255	1.8	
		16	86	297	0.367003	0.890	0.064	13.9	16	85	5138	0.465	0.249	1.9	
6 Teen parent		0	7	16991	0.032488	0.152	0.028	5.4	0	7	17100	0.155	0.028	5.4	
		5	31	7402	0.037287	0.225	0.030	7.5	5	31	8638	0.224	0.030	7.4	
		10	66	1673	0.037657	0.476	0.020	23.2	10	67	3959	0.381	0.031	12.3	
		16	79	835	0.034731	0.759	0.009	87.3	16	80	3959	0.434	0.028	15.3	
7 Single parent		0	9	8130	0.052276	0.214	0.043	5.0	0	9	9484	0.227	0.046	4.9	
		5	34	5037	0.051618	0.227	0.042	5.4	5	39	5910	0.248	0.041	6.1	
		10	70	953	0.039874	0.632	0.015	41.3	10	75	3599	0.367	0.033	11.2	
		16	82	510	0.035294	0.889	0.004	218.7	16	86	3599	0.362	0.033	10.9	
8 Social housing		0	9	9343	0.145885	0.337	0.113	3.0	0	9	9392	0.336	0.113	3.0	
		5	39	4525	0.129282	0.393	0.090	4.4	5	39	5850	0.405	0.091	4.4	

outcome	<i>Dropping cases with missing data, upper bound</i>							<i>Imputing missing data, lower bound</i>					
	age	No. constructs	N	M	TP	FP	ratio	age	No. constructs	N	TP	FP	ratio
	10	77	627	0.08134	0.804	0.017	46.3	10	75	3560	0.440	0.079	5.6
	16	90	304	0.0625				16	88	3560	0.449	0.078	5.8
9 Homeless	0	8	8615	0.066744	0.162	0.060	2.7	0	9	8660	0.164	0.060	2.7
	5	32	4514	0.065352	0.244	0.053	4.6	5	39	5429	0.222	0.054	4.1
	10	64	905	0.060774	0.600	0.026	23.2	10	71	4558	0.282	0.049	5.7
	16	75	449	0.033408				16	81	4558	0.310	0.048	6.5
10 Victim	0	8	9440	0.071822	0.176	0.064	2.8	0	9	9489	0.173	0.064	2.7
	5	32	4949	0.066882	0.211	0.057	3.7	5	34	5894	0.199	0.059	3.3
	10	69	452	0.050885	0.826	0.009	88.6	10	70	3590	0.282	0.051	5.5
	16	81	233	0.04721				16	82	3590	0.286	0.051	5.6
11 Dis-satisfied	0	5	10145	0.147462	0.219	0.136	1.6	0	5	10198	0.211	0.136	1.5
	5	30	4944	0.143204	0.267	0.123	2.2	5	28	6351	0.268	0.126	2.1
	10	56	2993	0.140996	0.351	0.107	3.3	10	56	5493	0.326	0.115	2.8
	16	65	1460	0.122603	0.419	0.081	5.2	16	66	5493	0.340	0.113	3.0
12 No Qualifications	0	10	8052	0.272479	0.441	0.209	2.1	0	10	9395	0.442	0.213	2.1
	5	37	4571	0.252899	0.478	0.177	2.7	5	34	5869	0.475	0.184	2.6
	10	71	407	0.154791	0.778	0.041	19.1	10	67	3584	0.523	0.160	3.3
	16	82	213	0.13615				16	79	3584	0.544	0.153	3.6
13 Not Level 2 Qualifications	0	11	8052	0.358048	0.533	0.259	2.1	0	11	9395	0.533	0.263	2.0
	5	42	4520	0.334292	0.584	0.209	2.8	5	42	5854	0.582	0.216	2.7
	10	83	832	0.189904	0.633	0.086	7.4	10	84	3575	0.626	0.185	3.4
	16	95	442	0.153846	0.912	0.016	56.8	16	97	3575	0.649	0.173	3.8
14 Not Level 4 Qualifications	0	2	10239	0.800274	0.865	0.557	1.6	0	2	10240	0.865	0.557	1.6
	5	18	5198	0.793382	0.875	0.480	1.8	5	20	6359	0.880	0.484	1.8
	10	39	1837	0.730539	0.871	0.349	2.5	10	40	3871	0.900	0.398	2.3
	16	45	1837	0.730539	0.887	0.307	2.9	16	50	3871	0.910	0.357	2.6
15 Low level criminality	0	10	9249	0.222619	0.473	0.152	3.1	0	11	9298	0.471	0.152	3.1
	5	41	4482	0.218206	0.494	0.141	3.5	5	42	5799	0.496	0.144	3.5
	10	78	398	0.188442	0.853	0.034	25.1	10	80	3536	0.557	0.117	4.7
	16	91	208	0.144231				16	93	3536	0.580	0.111	5.2
16 High level criminality	0	8	10148	0.046512	0.208	0.039	5.4	0	8	10201	0.209	0.039	5.4

Table 6.2: p2 of 5

outcome	<i>Dropping cases with missing data, upper bound</i>							<i>Imputing missing data, lower bound</i>					
	age	No. constructs	N	M	TP	FP	ratio	age	No. constructs	N	TP	FP	ratio
	5	35	5229	0.044559	0.309	0.032	9.6	5	35	6315	0.289	0.033	8.8
	10	68	1292	0.02322	0.600	0.010	63.1	10	69	5297	0.349	0.029	11.9
	16	81	698	0.014327	0.800	0.003	275.2	16	81	5297	0.367	0.029	12.8
17 Racial intolerance	0	7	9310	0.04855	0.128	0.045	2.9	0	6	9359	0.130	0.045	2.9
	5	27	4961	0.049385	0.159	0.044	3.6	5	27	5888	0.159	0.042	3.8
	10	56	2058	0.043732	0.378	0.028	13.3	10	55	5091	0.221	0.039	5.6
	16	68	1035	0.039614	0.585	0.017	34.2	16	65	5091	0.266	0.037	7.2
18 Not voting	0	10	8079	0.37319	0.456	0.324	1.4	0	10	9426	0.455	0.326	1.4
	5	35	4890	0.370757	0.470	0.312	1.5	5	37	5870	0.483	0.306	1.6
	10	75	544	0.314338	0.661	0.155	4.2	10	79	3577	0.515	0.277	1.9
	16	86	286	0.293706	0.893	0.045	20.0	16	91	3577	0.537	0.265	2.0
19 Low self-efficacy	0	10	8015	0.242046	0.351	0.207	1.7	0	9	9309	0.357	0.208	1.7
	5	38	4554	0.238033	0.411	0.184	2.2	5	36	5818	0.393	0.190	2.1
	10	78	677	0.192024	0.677	0.077	8.8	10	76	3808	0.450	0.162	2.8
	16	90	334	0.152695	0.941	0.011	88.8	16	89	3808	0.463	0.158	2.9
20 Low hourly wage	0	10	6802	0.185975	0.373	0.143	2.6	0	11	6839	0.372	0.144	2.6
	5	40	3364	0.177765	0.405	0.129	3.1	5	41	4332	0.412	0.131	3.1
	10	78	309	0.084142	0.923	0.007	130.6	10	80	2682	0.489	0.106	4.6
	16	90	166	0.066265				16	92	2682	0.511	0.102	5.0
21 On benefits	0	9	9405	0.103456	0.249	0.087	2.9	0	10	9454	0.248	0.087	2.9
	5	38	4556	0.089772	0.303	0.069	4.4	5	41	5889	0.324	0.069	4.7
	10	76	401	0.044888				10	80	3588	0.387	0.058	6.7
	16	89	209	0.033493				16	92	3588	0.397	0.057	7.0
22 Workless household	0	9	9435	0.099523	0.211	0.087	2.4	0	10	9484	0.212	0.087	2.4
	5	37	4569	0.084045	0.260	0.068	3.8	5	39	5908	0.267	0.069	3.9
	10	74	664	0.066265	0.659	0.024	27.2	10	78	3599	0.343	0.055	6.2
	16	85	318	0.056604	0.944	0.003	283.3	16	87	3599	0.357	0.054	6.6
23 Workless household w children	0	8	9435	0.053418	0.202	0.045	4.5	0	9	9484	0.201	0.045	4.5
	5	32	5100	0.044902	0.227	0.036	6.2	5	37	5910	0.243	0.037	6.6
	10	65	1246	0.02488	0.484	0.013	36.7	10	72	4953	0.348	0.030	11.4
	16	76	666	0.021021	0.786	0.005	170.8	16	82	4953	0.362	0.030	12.1

Table 6.2: p3 of 5

<i>Dropping cases with missing data, upper bound</i>								<i>Imputing missing data, lower bound</i>					
outcome	age	No. constructs	N	M	TP	FP	ratio	age	constructs	N	TP	FP	ratio
24 Absent child	0	6	15605	0.021147	0.085	0.020	4.3	0	6	15613	0.085	0.020	4.3
	5	30	6903	0.023468	0.173	0.020	8.7	5	30	7971	0.178	0.021	8.3
	10	62	2095	0.024821	0.462	0.014	33.7	10	64	5202	0.291	0.025	11.8
	16	75	1042	0.018234	0.737	0.005	150.8	16	77	5202	0.326	0.023	13.9
25 Mental health problem	0	10	9354	0.243105	0.352	0.208	1.7	0	11	9403	0.346	0.210	1.7
	5	40	4531	0.241227	0.378	0.198	1.9	5	40	5858	0.386	0.193	2.0
	10	79	401	0.219451	0.716	0.080	9.0	10	81	3576	0.426	0.176	2.4
	16	92	209	0.239234				16	94	3576	0.461	0.165	2.8
26 Depression	0	9	9354	0.207612	0.325	0.177	1.8	0	9	9403	0.323	0.177	1.8
	5	39	4531	0.203929	0.352	0.166	2.1	5	39	5858	0.346	0.166	2.1
	10	77	401	0.179551	0.764	0.052	14.8	10	78	3576	0.411	0.147	2.8
	16	89	209	0.186603				16	90	3576	0.436	0.140	3.1
27 Other mental health issue	0	4	9448	0.092295	0.153	0.086	1.8	0	4	9497	0.152	0.087	1.7
	5	23	5114	0.091905	0.172	0.084	2.1	5	24	5959	0.174	0.083	2.1
	10	52	656	0.077744	0.588	0.035	16.9	10	54	5399	0.219	0.077	2.9
	16	64	349	0.083095	0.793	0.019	42.3	16	65	5399	0.244	0.074	3.3
28 Addition to drug/alcohol	0	5	10261	0.023877	0.078	0.022	3.5	0	5	10314	0.078	0.022	3.5
	5	28	5684	0.023047	0.145	0.020	7.2	5	28	6397	0.146	0.020	7.4
	10	62	1323	0.015117	0.550	0.007	79.6	10	65	3885	0.306	0.016	19.7
	16	73	711	0.011252	0.875	0.001	615.1	16	75	3885	0.376	0.014	27.0
29 NEET18	0	7	9440	0.073093	0.168	0.065	2.6	0	8	9489	0.169	0.065	2.6
	5	30	4964	0.060637	0.223	0.050	4.4	5	34	8007	0.219	0.057	3.8
	10	64	736	0.050272	0.649	0.019	34.9	10	68	4011	0.261	0.045	5.8
	16	74	352	0.03125				16	78	4011	0.287	0.043	6.6
30 NEET30	0	10	8133	0.093815	0.284	0.074	3.8	0	10	9489	0.298	0.076	3.9
	5	37	4681	0.088015	0.352	0.063	5.6	5	39	5894	0.370	0.062	5.9
	10	74	684	0.054094	0.757	0.014	54.4	10	79	3590	0.431	0.053	8.1
	16	87	326	0.03681				16	91	3590	0.467	0.050	9.4
31 Overcrowding	0	5	15490	0.041382	0.105	0.039	2.7	0	5	15582	0.104	0.039	2.7
	5	25	6888	0.046458	0.128	0.042	3.0	5	27	8045	0.147	0.043	3.4
	10	54	2721	0.039691	0.250	0.031	8.1	10	58	5164	0.235	0.037	6.4
	16	65	1269	0.038613	0.510	0.020	25.9	16	70	5164	0.303	0.034	9.0

Table 6.2: p4 of 5

outcome	<i>Dropping cases with missing data, upper bound</i>							<i>Imputing missing data, lower bound</i>					
	age	No. constructs	N	M	TP	FP	ratio	age	constructs	N	TP	FP	ratio
32 Financial problem	0	9	9410	0.07152	0.156	0.065	2.4	0	9	9459	0.156	0.065	2.4
	5	33	4946	0.069753	0.191	0.061	3.2	5	35	5890	0.186	0.060	3.1
	10	63	1517	0.063942	0.464	0.037	12.7	10	65	3588	0.297	0.049	6.1
	16	70	774	0.047804	0.622	0.019	32.7	16	72	3588	0.332	0.046	7.2
33 > 1 of 5 key outcomes	0	9	9341	0.14795	0.339	0.115	2.9	0	9	9390	0.340	0.114	3.0
	5	38	4525	0.13105	0.363	0.096	3.8	5	39	5849	0.388	0.095	4.1
	10	77	399	0.077694	0.871	0.011	80.1	10	79	3559	0.437	0.081	5.4
	16	90	209	0.066986				16	92	3559	0.496	0.072	6.9
34 > 8 of 32 outcomes	0	9	15521	0.11565	0.226	0.101	2.2	0	9	15613	0.226	0.101	2.2
	5	37	6275	0.131952	0.315	0.104	3.0	5	38	7948	0.321	0.102	3.2
	10	75	411	0.104623	0.791	0.024	32.3	10	76	3652	0.484	0.098	5.0
	16	88	214	0.084112				16	88	3652	0.515	0.092	5.6

Table 6.2: p5 of 5

Table 6.3: Identification of those with high cost/harm outcomes, up to age of measure (i.e using previous measures), ALSPAC

Dropping cases with missing data, upper bound

outcome	age	No.		TP	FP	ratio
		constructs	N			
1 Bottom quartile, KS2 English	0	22	3273	0.389	0.094	4.1
	24	38	2609	0.427	0.089	4.8
	47	70	2104	0.462	0.078	5.9
	84	82	314	0.941	0.007	131.8
	114	64	228	0.944	0.005	198.3
2 Bottom quartile, KS2 Maths	0	15	6034	0.359	0.129	2.8
	24	18	5889	0.363	0.125	2.9
	47	32	4957	0.388	0.113	3.4
	84	44	796	0.650	0.060	10.8
	114	52	389	0.725	0.031519	23.0
3 Bottom quartile, KS2 Science	0	14	6421	0.379	0.123	3.1
	24	16	6270	0.386	0.119	3.2
	47	31	5174	0.394	0.105	3.8
	84	43	847	0.563	0.071	7.9
	114	51	412	0.684	0.032	21.3
4 Bottom quartile, KS2 E, M & S	0	17	6980	0.230	0.048	4.8
	24	27	5942	0.246	0.045	5.5
	47	58	4594	0.318	0.037	8.6
	84	70	688	0.821	0.008	108.4
	114	59	429	0.818	0.005	171.0
5 High anti-social score, age 10	0	24	2737	0.346	0.111	3.1
	24	55	2132	0.437	0.099	4.4
	47	114	1424	0.538	0.074	7.3
	84	126	208			
	114	157	442	0.875	0.018	48.3
6 High depression score, age 10	0	5	3886	0.182	0.088	2.1
	24	7	3577	0.194	0.086	2.3
	47	23	2984	0.220	0.081	2.7
	84	30	680	0.364	0.056	6.5
	114	102	692	0.481	0.042	11.4
7 >2 of 5 outcomes	0	19	2494	0.346	0.104	3.3
	24	25	2374	0.372	0.098	3.8
	47	48	2126	0.393	0.094	4.2
	84	56	1571	0.500	0.070	7.1
	114	87	925	0.633	0.049	12.9

Table 7: Prediction of multiple outcomes, five most important constructs, by level of True Positives and ratio of True to False Positives, BCS

construct	age	domain	rank (TP)	Rank (ratio)	rank(TP), by age	rank (ratio), by age	TP	FP	ratio
Outcome: More than 1 of 5 Key outcomes									
By true positive rate									
Mother's qualifications	5	D	1	59	1	17	0.466	0.241	1.935
Father's qualifications	5	D	2	53	2	13	0.431	0.217	1.981
Telephone present in home (income proxy)	10	D	3	37	1	19	0.390	0.177	2.201
Father reads to child	5	P	4	77	3	24	0.389	0.237	1.642
Mother expects child to stay on in school	10	P	5	46	2	23	0.360	0.171	2.106
By true positive/false positive ratio									
Truants at age 10	10	A	84	1	35	1	0.016	0.003	4.758
Father has dismissive attitude to child (teacher report)	10	P	82	2	33	2	0.019	0.005	3.732
Mother has dismissive attitude to child (teacher report)	10	P	79	3	32	3	0.033	0.010	3.433
Housing density	5	D	51	4	18	1	0.227	0.077	2.938
Soiling difficulties	10	A	89	5	38	4	0.005	0.002	2.864
Outcome: More than 8 of 32 outcomes									
By true positive rate									
Family income	10	D	1	28	1	16	0.345	0.181	1.910
Father reads to child	5	P	2	81	1	24	0.338	0.238	1.419
Mother expects child to stay on in school	10	P	3	40	2	20	0.321	0.178	1.801
Telephone present in home (income proxy)	10	D	4	47	3	23	0.317	0.185	1.716
Mother's Occupational SEG	10	D	5	50	4	25	0.288	0.170	1.688
By true positive/false positive ratio									
Father has dismissive attitude to child (teacher report)	10	P	81	1	33	1	0.018	0.006	3.214
Attends special school	10	C	85	2	36	2	0.012	0.004	3.010
Truants at age 10	10	A	82	3	34	3	0.015	0.006	2.564
Mother has dismissive attitude to child (teacher report)	10	P	79	4	32	4	0.028	0.012	2.389
Child communication skills (teacher rated)	10	A	9	5	5	5	0.260	0.113	2.301

Table 8: Prediction of multiple outcomes, five most important constructs, by level of True Positives and ratio of True to False Positives, NCDS

construct	age	domain	rank (TP)	Rank (ratio)	rank(TP), by age	rank (ratio), by age	TP	FP	ratio
By true positive rate									
attendance in bottom 30% (age 16)	16	A	1	59	1	22	0.457	0.135	3.399
Single index of teacher's rating child's progress at age 16	16	C	2	1	2	1	0.356	0.059	5.993
Father reads books or technical journals	7	P	3	107	1	32	0.342	0.129	2.646
attendance in bottom 30% (age 15)	16	A	4	32	3	14	0.332	0.082	4.041
Standardised maths score at age 11	11	C	5	15	1	8	0.326	0.071	4.588
By true positive/false positive ratio									
"Looked after" status at or before age 16	16	D	162	2	47	2	0.024	0.004	5.695
Cannot read well enough for everyday needs at age 16	16	C	152	3	44	3	0.045	0.008	5.654
"Mental backwardness" at age 11 (teacher report)	11	C	115	4	36	1	0.129	0.024	5.487
Behavioural difficulties at age 11 (teacher report)	11	A	166	5	50	2	0.019	0.003	5.456
Receiving special help for behaviour difficulties at age 11 (teacher report)	11	A	167	5	51	2	0.018	0.003	5.456

Table 9: Prediction of multiple outcomes, five most important constructs, by level of True Positives and ratio of True to False Positives, ALSPAC

construct	age	domain	Rank rank (TP) (ratio)	rank rank(TP), (ratio), by by age	rank (ratio), by age	TP	FP	ratio	
Outcome 4: Low performance in Key stage 2									
By true positive rate									
KS1: Writing points score	84	C	1	2	1	2	0.511	0.079	6.456
KS1: Reading points score	84	C	2	3	2	3	0.501	0.078	6.440
KS1: Maths score	84	C	3	1	3	1	0.410	0.043	9.495
Entry Assessment: Maths	84	C	4	5	4	5	0.324	0.083	3.885
Entry Assessment: Writing	84	C	5	14	5	8	0.307	0.113	2.713
By true positive/false positive ratio									
KS1: Maths score	84	C	3	1	3	1	0.410	0.043	9.495
KS1: Writing points score	84	C	1	2	1	2	0.511	0.079	6.456
KS1: Reading points score	84	C	2	3	2	3	0.501	0.078	6.440
KS1: Spelling score	84	C	10	4	7	4	0.252	0.061	4.132
Entry Assessment: Maths	84	C	4	5	4	5	0.324	0.083	3.885
Outcome 7: 3 or more of 5 high risk outcomes									
By true positive rate									
KS1: Maths score	84	C	1	2	1	2	0.586	0.128	4.576
KS1: Reading points score	84	C	2	1	2	1	0.583	0.121	4.814
KS1: Writing points score	84	C	3	3	3	3	0.507	0.122	4.147
Family type, 24 months	24	X	4	96	1	21	0.481	0.341	1.408
Entry Assessment: Language	84	C	5	19	4	8	0.473	0.245	1.926
By true positive/false positive ratio									
KS1: Reading points score	84	C	2	1	2	1	0.583	0.121	4.814
KS1: Maths score	84	C	1	2	1	2	0.586	0.128	4.576
KS1: Writing points score	84	C	3	3	3	3	0.507	0.122	4.147
Entry Assessment: Maths	84	C	8	4	5	4	0.397423	0.097	4.094
KS1: Spelling points score	84	C	13	5	6	5	0.3419	0.096	3.568

Table 10: Ten most predictive constructs for multiple deprivation outcomes, 1958 Cohort

Age	Domain	Construct	ROC area
Outcome 40: >2 of 6 age 23 outcomes			
16	C	Teacher's rating of child's progress	0.821
16	C	Reading score	0.807
11	C	Maths score	0.797
16	C	Maths score	0.796
11	C	Teacher's rating of child's progress	0.789
11	C	Reading score	0.789
7	C	Teacher's rating of child's capability	0.782
16	A	Teacher-rated externalising behaviour	0.777
16	S	Child will benefit from staying in education, teacher-rating	0.775
7	C	Reading score	0.758
Outcome 41: >7 of 32 age 42 outcomes			
16	C	Teacher's rating of child's progress	0.728
16	C	Maths score	0.708
16	C	Reading score	0.700
16	S	Child will benefit from staying in education, teacher-rating	0.695
11	C	Maths score	0.691
11	C	Teacher's rating of child's progress	0.688
16	A	Teacher-rated externalising behaviour	0.683
7	C	Teacher's rating of child's capability	0.682
11	C	Reading score	0.680
16	P	Likely to leave school at age 16	0.664

Table 11: Ten most predictive constructs for multiple deprivation outcomes, 1970 Cohort

Age	Domain	Construct	ROC area
Outcome 33: > 1 of 5 key outcomes			
10	C	Reading score	0.684
10	C	Maths score	0.684
10	A	Communication score	0.683
10	A	Attention score	0.679
10	C	British Ability Scale	0.674
16	A	Mother-rated conduct scale	0.673
16	A	Self-esteem	0.669
16	A	Locus of control	0.668
16	A	Wants to stay on in school	0.667
0	D	Father's SEG	0.665
Outcome 34: > 8 of 32 outcomes			
16	A	Wants to stay on in school	0.677
16	A	Self-esteem	0.661
16	A	Locus of control	0.660
16	A	Mother-rated conduct scale	0.660
10	A	Communication score	0.656
10	A	Attention score	0.655
10	C	Reading score	0.653
16	A	Smoker	0.650
10	C	Maths score	0.645
10	C	British Ability Scale	0.637

Table 12: First 3 levels for Classification Tree Analysis, BCS

	Level	Variables	Predicted probabilities	
			Minimum	Maximum
Outcome 33a: 6 or more of 32 outcomes (20%)				
Full Sample	1	Child stay on in education (lowest 25%)	0.2553	0.4741
	2	Mother's age (youngest 20%) Vocabulary scores at age 5 (lowest 25%)	0.2502	0.6746
	3	Play truant at age 16 (lowest 25%) Anti-social behaviour, age 10	0.1711	0.7713
Females	1	British Ability Scales, age 10	0.1994	0.3442
	2	Play truant at age 16 (lowest 25%) Mother did not stay on in education	0.1207	0.6424
	3	Mother's age (youngest 20%) Child stay on in education (lowest 25%)	0.1095	0.7695
Males	1	Child stay on in education (lowest 25%)	0.2360	0.4658
	2	Peer group age 10, % children from low SES No older siblings, at birth	0.1839	0.5007
	3	Profile drawing score age 5 (lowest 25%) Father does not have qualifications Frequency of going to church activities (lowest 25%)	0.0976	0.4905
Outcome 33b: 9 or more of 32 outcomes (10%)				
Full Sample	1	British Ability Scales, age 10	0.1063	0.2016
	2	Play truant at age 16 (lowest 25%) Number in class requiring remedial help in reading	0.0733	0.4200
	3	Reading scores at age 5 (lowest 25%) Height age 10, lowest 20%	0.0605	0.4676
Females	1	Child stay on in education (lowest 25%)	0.0994	0.2328
	2	Mother does not have qualifications Play truant at age 16 (lowest 25%)	0.0468	0.3767
	3	Child wets in bed at age 5	0.0431	0.4817
Males	1	Reading scores at age 5 (lowest 25%)	0.1327	0.1910
	2	Height age 10, lowest 20% Peer group age 10, % children low academic achievement	0.1324	0.2713
	3	Participation in other school activity, age 16 Behaviour: general conduct disorder at age 16	0.0973	0.3118

Table 13: First 3 levels for Regression Tree Analysis, all 32 outcomes, BCS

Outcome	Female				Male			
	Level	Variables	Predicted probabilities		Level	Variables	Predicted probabilities	
			Minimum	Maximum			Minimum	Maximum
1 Smoker	1	Child smokes at 16 (lowest 25%)	0.2494	0.5973	1	Child smokes at 16 (lowest 25%)	0.3374	0.6049
	2	Depressed mother, age 5 Mother SES 4 & 5, age 5	0.2309	0.6194	2	Mother SES 4 & 5, at birth No participation in charitable activities	0.2210	0.6674
	3	Human figure drawing (age 5) Participation in school holidays Acquaintances engaged in antisocial acts	0.1605	0.7114	3	High levels of anti-social behaviour, age 10 Wets in bed age 5	0.1944	0.7462
2 Obese	1	Neighbourhood poor education	0.1425	0.1987	1	Mother-rated (MR) behaviour: general conduct disorder_m-75	0.1306	0.1470
	2	Mother SES 4 & 5, at birth	0.1335	0.2345	2	Child doesn't want to stay on (age 16)	0.1262	0.1589
	3	Mother SES 4 & 5, age 5	0.1290	0.2696	3	Locus of control age 10 (lowest 20%)	0.1217	0.1701
3 Depression	1	Teacher's rating of hyperactivity at age 5 (lowest 25%)	0.1829	0.2515	1	Internalising behaviours age 10 (top 20%)	0.1280	0.1923
	2	Reading scores at age 10 (lowest 25%)	0.1652	0.3316	2	Father's SES 4 & 5, age 10 Father's SES 4 & 5, age 5	0.1125	0.2174
	3	Low school reading set	0.1638	0.3918	3	MR behaviour: internalising at 16 (top 20%)-75 Mother does not have qualifications	0.0617	0.2398
4 Psychiatric disturbance	1	British Ability Scale, bottom 25%, age 10	0.2260	0.2336	1	Internalising behaviours age 10 (top 20%)	0.1680	0.1923
	2	Externalising behaviours age 10 (top 20%) Has had alcoholic drink in last wk, age 16	0.2147	0.2780	2	School peer group age 10, low average attainment Young siblings at age 5	0.1621	0.1989
	3	Teacher's rating of vocabulary at age 5 (lowest 25%) Teacher's rating of peers at age 10 (lowest 25%)	0.2046	0.3067	3	Frequency of going to community centre (lowest 25%) Does not have telephone at home, age 5	0.1363	0.2129
5 Single, separated, divorced	1	Number in class requiring remedial help in reading	0.2931	0.3007	1	Mother-rated (MR) behaviour: general conduct disorder_m-75	0.3592	0.4049
	2	Age 5 height, lowest 20% Teacher's rating youth club use at age 11 (lowest 25%)	0.2588	0.3734	2	Human figure drawing, age 5 (lowest 25%) Acquaintances engaged in antisocial acts-75	0.3302	0.4570
	3	No older brothers, age 0 Youth club attendance (age 16)	0.2230	0.4128	3	Copying score age 5 (lowest 25%) No participation in charitable activities CM region London & South East	0.3158	0.5434
6 Teen parent	1	Father's unemployed, age 10	0.0615	0.1134	1	none		
7 Single parent	1	No father at home, age 10	0.1016	0.1475		none		
	2	Youth club attendance (age 16)	0.0685	0.1541				
	3	Parents' expect child to stay on in education, age 10	0.0610	0.1946				
8 Social housing	1	Parents' approve activities	0.0874	0.2238	1	Teacher's rating of attendance at age 10 (lowest 25%)	0.0916	0.1853
	2	Participation in day trips	0.0750	0.2314	2	Reading score at age 10 (lowest 25%) Acquaintances engaged in antisocial acts_p75	0.0796	0.2383
	3	Number in class requiring remedial help in reading	0.0707	0.2400				
9 Homeless	1	Father's income, lowest 20%, age 10	0.0722	0.1270	1	Father's no qualifications, age 5	0.0495	0.0761
	2	Father's unemployed, age 10	0.0681	0.1756				
	3	Uninformed youth club attendance (age 16)	0.0496	0.2002				
10 Victim	1	Casual friends engaged in anti-social acts	0.0342	0.0572	1	Unauthoritarian parenting age 5 (top 20%)	0.0998	0.1059
	2	Mother rated behaviour: clumsy frustrated	0.0269	0.0623	2	Father did not stay on in education Mother did not stay on in education	0.0845	0.1184
					3	Teacher's rating of attendance at age 10 (lowest 25%) Participation in school holidays	0.0590	0.1461
11 Dis-satisfied	1	Father's unemployed, at birth	0.1471	0.2215	1	Low school reading set	0.1395	0.2080
	2	Vocabulary score at age 5 (lowest 25%)	0.1274	0.2647	2	Father's no qualifications, age 5	0.1303	0.2300
	3	British Ability Scale, bottom 25%, age 10	0.1135	0.2841	3	Father's SES 4 & 5, age 5	0.1177	0.2359
12 No Qualifications	1	Maths scores at age 10 (lowest 25%)	0.2029	0.4458	1	Child doesn't want to stay on (age 16)	0.2278	0.4681
	2	Locus of control (age 16)	0.1213	0.4726	2	Reading score at age 10 (lowest 25%) Teacher's rating of anxiety at age 10 (top 20%)	0.1832	0.5637
	3	Parents' expect child to stay on in education, age 10 Age 10 height, lowest 20%	0.0849	0.5745	3	No father present age 10 Age 5 height, lowest 20%	0.1754	0.6022

Outcome	Female				Male			
	Level	Variables	Predicted probabilities		Level	Variables	Predicted probabilities	
			Minimum	Maximum			Minimum	Maximum
13 Not Level 2 Qualifications	1	British Ability Scale, bottom 25%, age 10	0.2576	0.5791	1	Child doesn't want to stay on (age 16)	0.2982	0.6102
	2	Self efficacy (age 16)	0.1994	0.7070	2	Reading score at age 10 (lowest 25%)	0.2500	0.7430
	3	Mother depressed, age 5 Father's no qualifications, age 5 High set for reading (age 10)	0.1173	0.7491	3	Teacher's rating of peers at age 10 (lowest 25%) No father present age 10 Age 5 height, lowest 20%	0.2494	0.7515
14 Not Level 4 Qualifications	1	Child doesn't want to stay on (age 16)	0.7504	0.9575	1	Child doesn't want to stay on (age 16)	0.7230	0.9634
	2	Father's no qualifications, age 5	0.6772	0.9862	2	Parents' expect child to stay on in education, age 10	0.6226	0.9911
	3	Mother SES 4 & 5, age 10 Mother SES 4 & 5, at birth	0.6186	0.9948	3	Number in class requiring remedial help in reading Number in class requiring remedial help Locus of control (age 16)	0.4163	0.9939
15 Low criminality	1	Locus of control (age 16)	0.0662	0.1157	1	Child smokes at 16 (lowest 25%)	0.3253	0.5510
	2	Participation in baby-sitting Father's unemployed, at birth	0.0073	0.1077	2	Parents approve activities age 16-5 School peer group age 10, low average attainment	0.2191	0.5907
	3	Father over concerned (teacher rated) Externalising behaviours age 10 (top 20%) High anti-social behaviour amongst acquaintances, age 16	0.0078	0.2725	3	participation in clubs / societies Frequency of going to community centre (lowest 25%)-25 Participation in school holidays	0.1624	0.6328
16 High criminality	1	Human figure drawing, age 5 (lowest 25%)	0.0060	0.0125		none		
17 Racial intolerance	1	Self efficacy (age 16)	0.0227	0.0315	1	Father's unemployed, age 5	0.0686	0.0879
	2	Teacher's rating of self esteem at age 10 (lowest 25%)	0.0189	0.0353	2	Parent's anti-TV age 5 (top 20%)	0.0656	0.0945
	3	Reading score at age 5 (lowest 25%)	0.0164	0.0374	3	Age 5 height, lowest 20% CM engaged in gen. antisoc acts (top 20%)-75	0.0528	0.1151
18 Not voting	1	Child doesn't want to stay on (age 16)	0.3292	0.4165	1	Internalising behaviours age 10 (top 20%)	0.3959	0.4088
	2	Father's unemployed, at birth Mother SES 4 & 5, age 10	0.3187	0.5629	2	MR behaviour: internalising at 16 (top 20%)-75 No participation in school concerts / plays	0.2421	0.3926
	3	Mother SES 4 & 5, age 10 Parents' approve activities	0.2521	0.6034	3	participation in clubs / societies Does not read often per week cmregb9	0.1743	0.4420
19 Low self-efficacy	1	Neighbourhood poor education	0.2177	0.3475	1	Father does not have qualifications, age 0	0.2315	0.2686
	2	Parents' approve activities	0.1615	0.3847	2	Age 5 height, lowest 20% Locus of control (age 16)	0.1301	0.3444
	3	Teacher's rating of self esteem at age 10 (lowest 25%) Reading score at age 10 (lowest 25%)	0.1274	0.5327	3	Teacher's rating of peers at age 10 (lowest 25%) Low school reading set	0.1213	0.4967
20 Low income	1	Low school reading set	0.2375	0.5274	1	Mother's age (lowest 20%)	0.1227	0.1534
	2	Teacher's rating of socialising skills, age 5 (lowest 25%)	0.2131	0.6224	2	Participation in day trips	0.0910	0.1340
	3	Father's unemployed, at birth	0.1692	0.5208	3	Locus of control age 10 (lowest 20%)	0.0799	0.1928
21 On benefits	1	No father present at birth	0.1267	0.2488	1	Father's no qualifications, age 5	0.0559	0.0969
	2	Father's unemployed, at birth	0.1230	0.2488	2	Human figure drawing, age 5 (lowest 25%)	0.0460	0.1356
	3	Father's unemployed, age 10	0.1088	0.4084	3	Externalising behaviour at age 10 (top 20%)	0.0420	0.1332
22 Workless household	1	No father present at birth	0.1087	0.2066	1	Child smokes at 16 (lowest 25%)	0.0743	0.1218
	2	Child doesn't want to stay on (age 16)	0.0884	0.2701	2	Does not read often per week MR behaviour: internalising at 16 (top 20%)-75	0.0467	0.1578
	3	Reading score at age 10 (lowest 25%) Father's unemployed, at birth	0.0691	0.3495	3	Father's no qualifications, age 5 Mother SES 4 & 5, age 10 Participation in other school activities	0.0170	0.2447
23 Workless household w children	1	No father present at birth	0.0796	0.1784	1	Reading score at age 10 (lowest 25%)	0.0123	0.0449
	2	Father's unemployed, at birth	0.0776	0.1784				
	3	Father's unemployed, age 10	0.0680	0.3116				
24 Absent child		none			1	Mother's age (lowest 20%)	0.0401	0.0478
					2	Father's no qualifications, age 5	0.0389	0.0616
					3	Does not read often per week	0.0342	0.0801
25 Mental health problem	1	Participation in day trips	0.2966	0.2997	1	Low school reading set	0.1597	0.2068
	2	Reading score at age 5 (lowest 25%)	0.2926	0.3294	2	Father's SES 4 & 5, age 5	0.1493	0.2215

Table 13: p2 of 3

Outcome	Female				Male			
	Level	Variables	Predicted probabilities		Level	Variables	Predicted probabilities	
			Minimum	Maximum			Minimum	Maximum
	3	Father's no qualifications, age 5 Father did not stay on in education Parent's anti-TV age 5 (lowest 25%)_p75	0.2352	0.4200	3	Number in class requiring remedial help Number in class requiring remedial help in reading	0.1094	0.2413
26	Depression	1 Parents' expect child to stay on in education, age 10 2 Father's no qualifications, age 5 Locus of control (age 16) 3 Human figure drawing, age 5 (lowest 25%) School peer group age 10, low average SES Child doesn't want to stay on (age 16)	0.2649 0.2201 0.2018	0.3096 0.3413 0.3797	1 2 3	Birthweight, lowest 20% Reading score at age 5 (lowest 25%) Participation in other school activities Mother did not stay on in education	0.1242 0.1167 0.0909	0.1547 0.1625 0.1730
27	Other mental health issue	1 No father present age 10 2 No father present age 5 Peer group, age 10: Low levels of high SES 3 Participation in day trips	0.1003 0.0943 0.0870	0.1110 0.1371 0.1612	1 2 3	Low school reading set Internalising behaviours age 10 (top 20%) Father's SES 4 & 5, age10 Father's SES 4 & 5, age 5	0.0706 0.0659 0.0545	0.0872 0.1158 0.1201
28	Addition to drug/alcohol	none			1 2	Participation in other school activities Father's no qualifications, age 5	0.0237 0.0216	0.0297 0.0415
29	NEET18	1 Neighbourhood poor education	0.0714	0.1303	1 2 3	Father ethnicity West Indies Child smokes at 16 (lowest 25%) Frequency of going to sports centre (lowest 25%)~25	0.0642 0.0534 0.0309	0.1111 0.1746 0.4422
30	NEET30	1 Reading score at age 10 (lowest 25%) 2 Father's unemployed, at birth	0.1148 0.1123	0.2967 0.3854	1	Father's SES 4 & 5, at birth	0.0380	0.0793
31	Overcrowding	1 No father present at birth 2 Teacher's rating of attendance at age 10 (lowest 25%) 3 Father's SES 4 & 5, age10	0.0495 0.0492 0.0451	0.0576 0.0984 0.1358	1 2 3	Mother did not stay on in education Mother does not have qualifications Copying score age 5 (lowest 25%) Teacher's rating of attendance at age 10 (lowest 25%) Age 10 height, lowest 20%	0.0186 0.0176 0.0146	0.0436 0.0669 0.1062
32	Financial problem	1 Child has been in care by age 5 2 Play truant at age 16 (lowest 25%) 3 Number in class requiring remedial help Father's no qualifications, age 5	0.0745 0.0566 0.0432	0.0976 0.1283 0.1560	1 2 3	Participation in day trips Participation in baby-sitting MR behaviour: internalising at 16 (top 20%)~75	0.0511 0.0478 0.0394	0.0516 0.0613 0.0649

Note: BCS data excludes sweeps at 22 and 42 months.

Maximum and minimum predicted probabilities for each level are calculated running a probit model: a) for level 1 using only level 1 covariates; b) for level 2, using level 1 and 2 covariates and c) using covariates for all 3 levels.

Table 14: First 3 levels for Regression Tree Analysis, using only selected measures, BCS (for outcomes where algorithm collapses)

Outcome	Level	Variables	Predicted probabilities	
			Minimum	Maximum
Females				
6 Teen parent	1	Child smokes at 16 (lowest 25%)	0.047	0.144
<i>observed mean for females = .0586</i>	2	Mother does not have qualifications	0.021	0.186
	3	Vocabulary scores at age 5 (lowest 25%)	0.015	0.365
		Child stay on in education (lowest 25%) Father's age (youngest 20%) No father at home, age 5		
7 Single parent	1	Persons per room at birth (top 20%)	0.074	0.141
<i>observed mean for females = .1045</i>	2	Father's SES 4 & 5, age 5	0.051	0.214
		Child smokes at 16 (lowest 25%)		
	3	Mother's age (youngest 20%)	0.032	0.440
		Parents approve of activities at age 16~5 No father at home, age 5 No father at home, at birth		
16 High criminality	1	Human figure drawing, age 5 (lowest 25%)_p25	0.006	0.012
<i>observed mean for females = .0077</i>	2	Close friends engaged in antisoc acts_pal~75	0.004	0.038
	3	Number in class requiring remedial help in reading	0.002	0.053
23 Workless household w children	1	Persons per room at birth (top 20%)	0.048	0.128
<i>observed mean for females = .0833</i>	2	Child smokes at 16 (lowest 25%)	0.020	0.221
		Parents approve of activities at age 16~5		
	3	Father's unemployed, at birth	0.017	0.545
		Human figure drawing 2, age 5 (lowest 25%)_p25 Neighbourhood poor education Father's SES 4 & 5, at birth		
24 Absent child	1	No father at home, at birth	0.002	0.002
<i>observed mean for females = .0017</i>	2	Frequency of going to uniformed youth clubs (lowest 25%)_p25	0.001	0.148
		Parent's anti-TV age 5 (top 20%)		
	3	Externalising behaviours at age 5 (top 20%)	0.000	0.210
		Mother-rated (MR) behaviour: internalising at age 16_m~75 Mother did not stay on in education		

28 Addiction to drug/alcohol	1	Child smokes at 16 (lowest 25%)	0.007	0.025
	2	Number in class requiring remedial help No older siblings, at birth	0.003	0.053
	3	Mother-rated (MR) behaviour: clumsy naughty at age 16_m~750.000 Frequency of going to uniformed youth clubs (lowest 25%)_p25		0.103
<i>observed mean for females = .0118</i>				
Males				
6 Teen parent	1	Child stay on in education (lowest 25%)	0.007	0.025
	2	Child has bee in care by age 16 Mother-rated (MR) behaviour: internalising at age 16	0.006	0.041
	3	Mother's age (youngest 20%) Birthweight, lowest 20% Father's unemployed, at birth	0.006	0.088
<i>observed mean for males = .0087</i>				
7 Single parent	1	Child has been in care by age 5	0.006	0.014
	2	Father's unemployed, at birth Father does not have qualifications	0.002	0.072
	3	Participation in school concerts / plays Mother's age (youngest 20%)	0.003	0.646
<i>observed mean for males = .0066</i>				
16 High criminality	1	Play truant at age 16 (lowest 25%)	0.046	0.279
	2	Parents approve of activities at age 16 No father at home, at birth	0.014	0.475
	3	Father's SES 4 & 5, at birth Mother's age (youngest 20%)	0.012	0.401
<i>observed mean for males = .0867</i>				
23 Workless household w children	1	Reading scores at age 10 (lowest 25%)	0.012	0.045
	2	Child smokes at 16 (lowest 25%) Mother's age (youngest 20%)	0.008	0.129
	3	Mother's SES 4 & 5, at birth Father's unemployed, at birth Copying scores at age 5 (lowest 25%) Mother's SES 4 & 5, at birth	0.006	0.221
<i>observed mean for males = .0234</i>				

24 Absent child	1	Father's ethnicity, West Indies	0.039	0.104
	2	Child locus of control age 16 (lowest 25%)	0.021	0.189
	3	No mother at age 5	0.017	0.307
<i>observed mean for males = .0395</i>				
28 Addiction to drug/alcohol	1	Child wets in bed at age 5	0.033	0.105
	2	Child locus of control age 16 (lowest 25%) Maths scores at age 10 (lowest 25%)	0.010	0.109
	3	Teacher's rating of self esteem at age 10 (lowest 25%) spg4AcL_p25	0.008	0.137
<i>observed mean for males = .0368</i>				

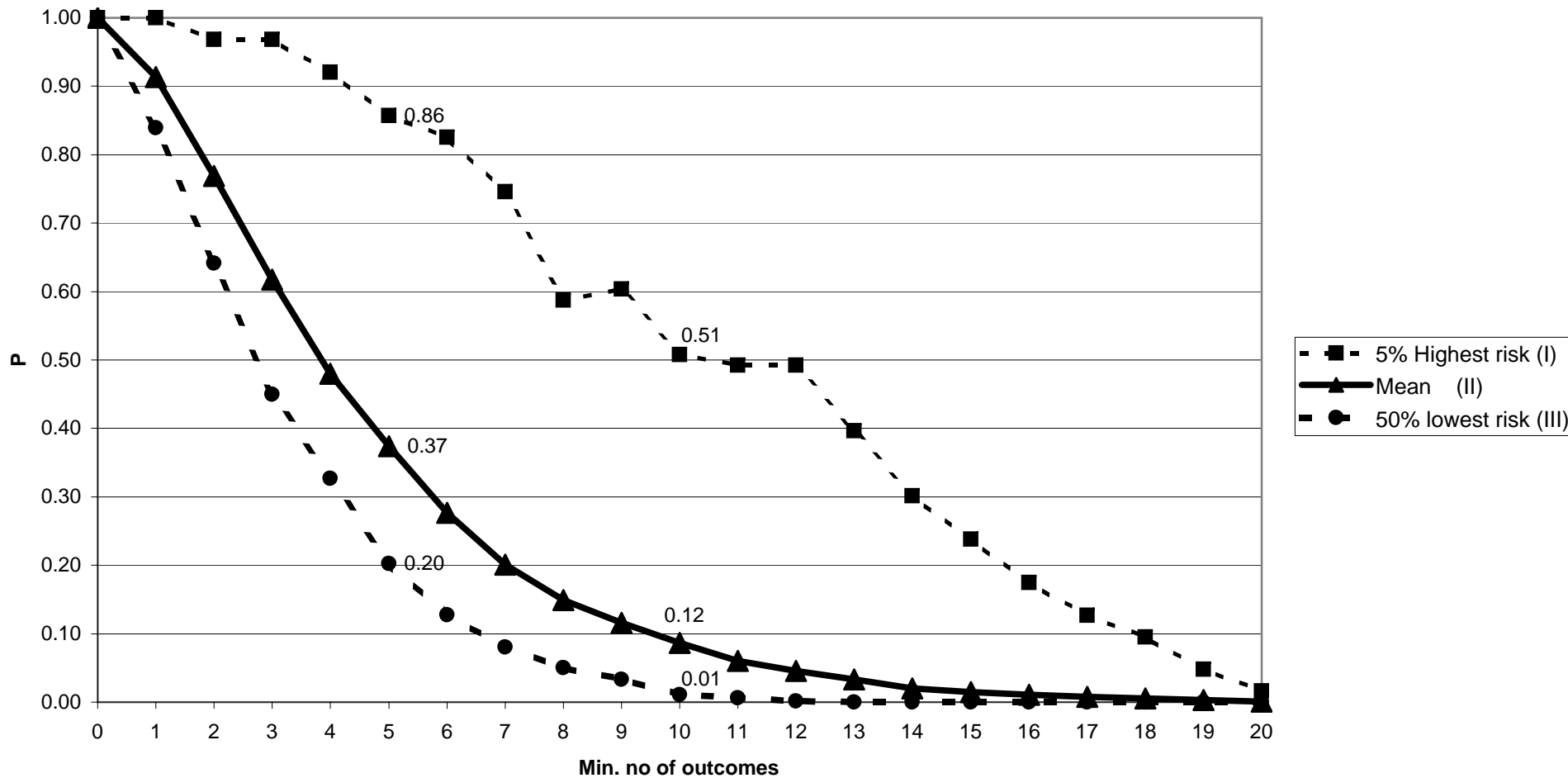
Table 15: First 3 levels for Regression Tree Analysis, NCDS (1958 Cohort)

Outcome	Level	Variables	Predicted probabilities	
			Minimum	Maximum
Females				
33 Unskilled or semi-skilled occupation <i>observed mean for females = .2016</i>	1	Teacher's rating of progress age 16 (lowest 25%)	0.113	0.472
	2	Persons per room, age 7, (over 2) CM expect to continue in learning (age 11)	0.092	0.652
	3	reading teacher rating at 7 Mother smoke during pregnancy CM in low ability class age 11	0.067	0.854
34 Literacy or numeracy problem <i>observed mean for females = .111</i>	1	reading teacher rating at 7	0.062	0.308
	2	Maths scores at 7 (lowest 25%) Reading scores at age 11 (lowest 25%)	0.043	0.389
	3	number slow or very slow says teacher age 11 Rutter parental scale age 11 (top 20%) Reading scores at 7 (lowest 25%) Father's age (youngest 25%)	0.028	0.466
35 Had child before 20 <i>observed mean for females = .1203</i>	1	Maths scores age 11 (lowest 25%)	0.079	0.227
	2	Mother smokes regularly CM does not have good attitudes to school~25	0.052	0.297
	3	Parents do not have interest in school (age 16) Persons per room at birth (over 2) BSAG age 11 (top 20%)	0.028	0.482
38 Depressed <i>observed mean for females = .1099</i>	1	Rutter parental scale age 11 (top 20%)	0.086	0.162
	2	CM reads (sometimes) Number of children at home at 7 (over 4)	0.053	0.295
	3	Mother's family size (over 4) condition or handicap that interferes permanently at 11	0.028	0.670

Males

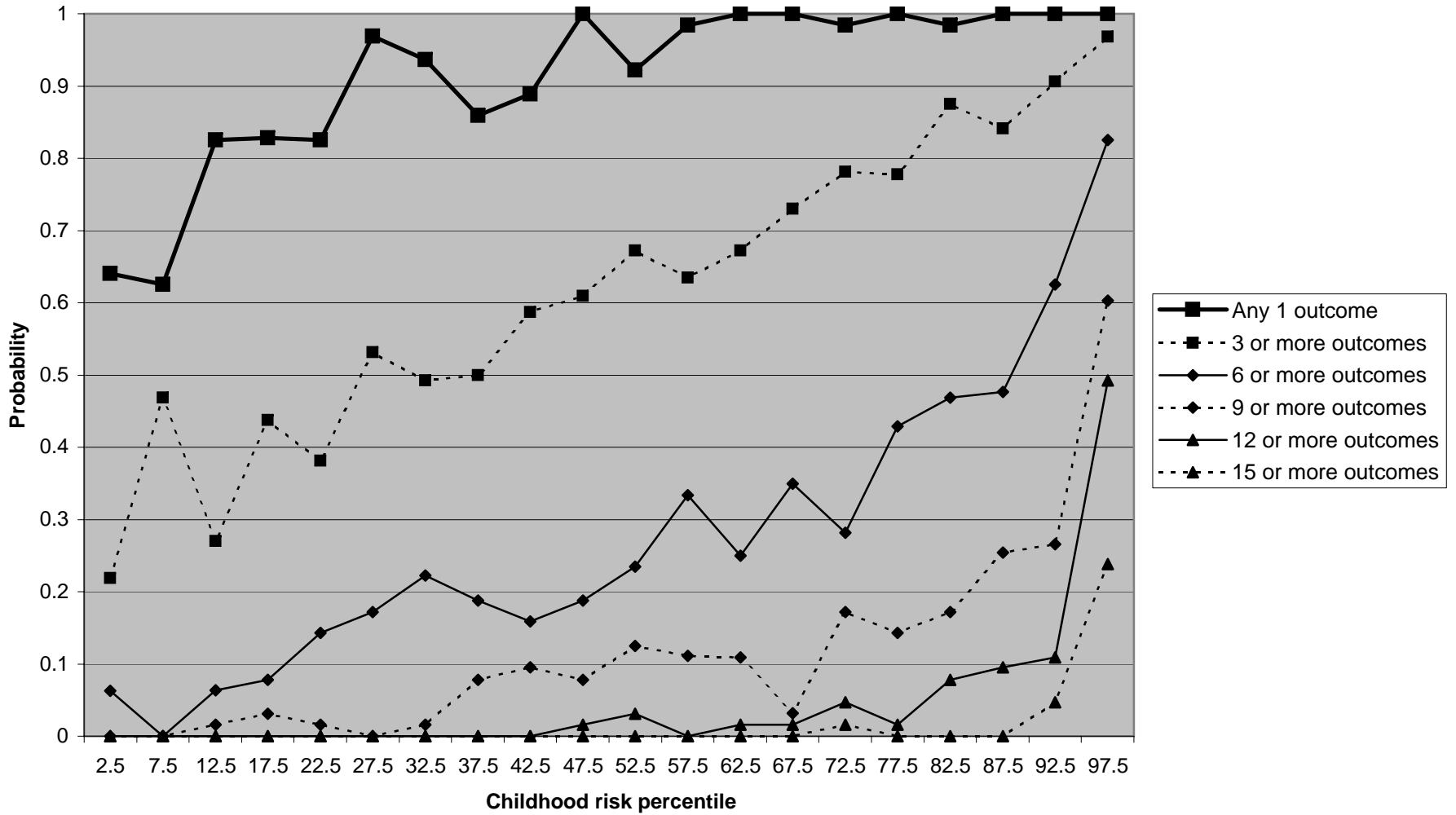
		Predicted probabilities		
Level	Variables	Minimum	Maximum	
33 Unskilled or semi-skilled occupation	1	Reading, teacher rating at 7	0.153	0.334
	2	Father did not stay on in education Father's SES 4 & 5, age 7	0.099	0.418
	3	General knowledge limited or handicap, teacher-report, 11 BSAG, age 7 (top 20%) Father's SES 4 & 5, at birth	0.092	0.547
<i>observed mean for males = .2145</i>				
34 Literacy or numeracy problem	1	use of books below average or very poor says teacher age 11	0.070	0.381
	2	Reading scores age 16 (lowest 25%) Reading scores at 7 (lowest 25%)	0.048	0.460
	3	reading teacher rating at 7 Persons per room at birth (over 2) Rutter parental scale age 11 (top 20%)	0.040	0.532
<i>observed mean for males = .1564</i>				
35 Had child before 20	1	English is not spoken at home at age 7	0.029	0.070
	2	Father's SES 4 & 5, age 7	0.022	0.109
	3	Class size at 7 (large) Weight age 16 (top 20%)	0.017	0.086
<i>observed mean for males = .0316</i>				
38 Depressed	1	Vocabulary scores, age 11 (lowest 25%)	0.024	0.079
	2	Good school attendance Rutter parental scale age 11 (top 20%)	0.015	0.221
	3	Financial hardship at 16 Class size at 7 (large) Mother smoke during pregnancy Mental backwardness age 16	0.007	0.446
<i>observed mean for males = .0413</i>				

Figure 2: Probability of crossing outcome threshold by level of risk, 1970 Cohort



Note: Full information sample used, N=1271

Figure 3: Probability of multiple adult deprivation by level of childhood risk



Note: Full information sample used, N=1271