Supplementary Information

Measuring the predictability of life outcomes using a scientific mass collaboration

Abstract

This file contains supplementary information for "Measuring the predictability of life outcomes using a scientific mass collaboration." It is designed to be used as a reference for readers seeking information on specific topics. It is not designed to be read from beginning to end.

Contents

S1 Design and execution of the Fragile Families Challenge

S1.1 Data

The background and outcome data used in the Fragile Families Challenge come from the Fragile Families and Child Wellbeing Study (hereafter Fragile Families study). The Fragile Families sample was a multi-stage, stratified random sample of hospital births between 1998 and 2000 in large U.S. cities (more than 200,000 residents), with a 3:1 oversample of births to non-married parents [22].

Data collection for the Fragile Families study occurred in six waves: at the birth of the focal child and when the focal child was approximately age 1, 3, 5, 9, and 15. Each wave consisted of different data collection modules (Figure 1, main text). Each data collection module is made up of approximately 10 sections, where each section includes questions about a specific topic. The full list of data collection modules between birth and year 9 and the topics included in each module are presented in Table S1. The survey codebooks—which include information about question text, response options, and question order—are available at `https://fragilefamilies.princeton.edu/documentation`.

The background dataset used in the Fragile Families Challenge was a specially constructed version of the Fragile Families data collected between the child's birth and age nine. To create the background dataset, we 1) combined the Fragile Families data into a single file, 2) dropped observations that were obtained in 2 out of the 20 cities of birth because these were pilot cities where some questions were asked differently, and 3) made changes to the data to promote the privacy of respondents and reduce the risk of harm in the event of re-identification [18]. Ultimately, the background dataset had 4,242 rows—one for each family—and 12,943 columns—one for each variable plus an ID number for each family. Of the 12,942 variables, 2,358 were constant (i.e., had the same value for all rows). Some of these constant variables were caused by our privacy and ethics redactions [18]. In addition to the data, metadata described the contents of the data [15].

Of the approximately 55 million possible entries in the background dataset ($4,242 \times 12,942$), about 73% of possible entries did not have a value (Fig. S1). The Fragile Families study uses a variety of codes to denote reasons that a data entry might not have a value, including not being in the survey wave (about 17% of all possible entries), refusal to answer certain questions (about 0.1% of entries), intentional skip patterns within the survey design (about 25% of entries), and redaction for privacy and ethics [18] (about 6% of entries).

The largest source noted above—intentional skip patterns—are not an indicator of poor data quality but are instead an immediate consequence of the survey design. For example, questions about relationships and parenting behaviors were often asked separately for resident versus non-resident parents, with the valid

Data module	Child age	Sections
Mother	Birth	A) Child health and development, B) Father-mother relationships, C) Fatherhood, D) Marriage attitudes, E) Relationship with extended kin, F) Environmental factors and government programs, G) Health and health behavior, H) Demographic characteristics, I) Education and employment, J) Income
Father	Birth	A) Child health and development, B) Father-mother relationships, C) Fatherhood, D) Marriage attitudes, E) Relationship with extended kin, F) Environmental factors and government programs, G) Health and health behavior, H) Demographic characteristics, I) Education and employment, J) Work activities, K) Income
Mother	1	A) Family characteristics, B) Child well-being and mothering, C) Father-child relationship, D) Mother's relationship with father, E) Current partner, F) Demographics, G) Mother's family background and support, H) Environment and programs, J) Health and health behavior, K) Education and employment, L) Income
Father	1	A) Family characteristics, B) Child well-being and fathering, C) Mother-child relationship, D) Father's relationship with mother, E) Current partner, F) Demographics, G) Father's family background and support, H) Environment and programs, J) Health and health behavior, K) Education and employment, L) Income
Mother	3	A) Family characteristics, B) Child well-being and mothering, C) Father-child relationship, D) Mother's relationship with father, E) Current partner, F) Demographics, H) Mother's family background and support, I) Environment and programs, J) Health and health behavior, R) Religion, K) Education and employment, L) Income
Father	3	A) Family characteristics, B) Child well-being and fathering, C) Mother-child relationship, D) Father's relationship with mother, E) Current partner, F) Demographics, H) Father's family background and support, I) Environment and programs, J) Health and health behavior, R) Religion, K) Education and employment, L) Income
Primary care giver and in-home observation	3	A) Health and accidents, B) Family routines, C) Home toy and activity items, D) Nutrition, E) Food expenditures, F) Housing/building characteristics, G) Parental stress, H) Parental mastery, J) Discipline, K) Informal social control and social cohesion and trust, L) Exposure to violence, M) Child's behavior problems, P) Observation checklist, Q) Common areas, R) Interior of house or apartment, S) Child's appearance, T) Home scale, U) Child emotion and cooperation, V) Ending
In-home activities with child and mother	3	A) Height and weight, B) Child's Peabody Picture Vocabulary Test or TVIP, C) Walk-A-Line, D) Q-Sort, E) Mothers Peabody Picture Vocabulary Test or TVIP, F) Child Care/Employment History Calendar
Child Care Provider Survey (for center-based care)	3	A) Care provided at the center, B) Care provided for focus child (Information from director or teacher), C) Care provided for focus child (Information from teacher), E) Teacher-parent relationship, F) Teacher beliefs, G) About the childcare teacher
Child Care Center Observations	3	No clear section headings but contents include: Space and furnishings, personal care routines, language-reasoning, activities, interaction, program structure, parents and staff
Family Care Provider Survey (for family-based care)	3	A) Care provided, B) Child care routine and program, D) Provider-parent relationship, E) Child care provider beliefs, F) About the child care provider
Family Care Provider Observations	3	No clear section headings but contents include: Space and furnishings for care and learning, basic care, language and reasoning, learning activities, social development
Child Care Study Post-Observation Form	3	A) Observation checklist, B) Common areas, C) Interior of building, D) Home scale, E) Post-visit rating by interviewer
Mother	5	A) Family characteristics, B) Child well-being and mothering, C) Father-child relationship, D) Mother's relationship with father (for mothers who are or were in a relationship) E) Current partner, F) Demographics, H) Mother's family background and support, I) Environment and programs J) Health and health behavior, R) Religion K) Education and employment, L) Income
Father	5	A) Family characteristics, B) Child well-being and fathering C) Mother-child relationship D) Father's relationship with mother (for fathers who are or were in a relationship), E) Current partner, F) Demographics, H) Father's family background and support I) Environment and programs J) Health and health behavior, R) Religion K) Education and employment, L) Income
Primary care giver and in-home observation	5	A) Health and accidents, B) Family routines, C) Home toy and activity items, D) Nutrition, E) Housing/building characteristics, F) Parental stress and mastery, G) Discipline, H) Exposure to violence, J) CPS contact, K) Food expenditures, L) Child's behavior, N) Activities, P) Observation checklist, Q) Common areas, R) Interior of house or apartment, S) Child's appearance, T) Home scale, U) Child emotion and cooperation, V) Ending
In-home activities with child and mother	5	A) Weight/height, B) Peabody Picture Vocabulary Test with child, C) Woodcock-Johnson Letter-Word activity with child, D) Attention sustained task, E) Child care employment history calendar, F) Five-minute speech sample, G) Peabody Picture Vocabulary Test with mother
Teacher	5	A) Information specific to the participating child, B) Academic skills specific to the participating child, C) Classroom behavior and social skills specific to the participating child, D) Classroom characteristics, E) Class resources and activities, F) School climate and environment, G) General information about teacher
Mother	9	A) Core mother interview: Family characteristics, household roster, marital, and fertility history, B) Bio father contributions and resources, C) Mother's relationship with father, D) Current partner, E) Mother's family background and support, F) Environment and programs, G) Health and health behavior, H) Religion, I) Education and employment, J) Income, K) Secondary caregiver
Father	9	A) Core father interview: Family characteristics, household roster, marital, and fertility history, B) Bio mother and bio father contributions and resources, C) Father's relationship with mother, D) Current partner, E) Father's family background and support, F) Environment and programs, G) Health and health behavior, H) Religion, I) Education and employment, J) Income, K) Secondary caregiver
Primary care giver	9	A) Introduction to non-parental caregiver survey, B) Mother-child relationship, C) Father-child relationship, D) Demographics, E) Income, education, and employment, F) Health and wellbeing, G) Environment, H) Health and accidents, I) Family routines and home life, J) Nutrition, K) Parental stress and mastery, L) Child's education, M) Child's neighborhood
Interviewer observation	9	A) Observation checklist, B) Common areas, C) Interior of house or apartment, D) Child's appearance, E) Home scale, F) Child emotion and cooperation, G) Ending
Child	9	A) Parental supervision and relationship, B) Parental discipline, C) Sibling relationships, D) Routines, E) School, F) Early delinquency, G) Task completion and behavior, H) Health and safety, I) Closing
In-home activities with child and primary care-giver	9	No clear section headings but activities include: Consent, Child assessment (PPVT, Digit span, Woodcock-Johnson Tests 9 and 10), Primary caregiver self-administered questionnaire, Health measures, Saliva sample, Biological mother weight, Child weigh/height, Primary caregiver open-ended responses
Teacher	9	A) General information, B) Classroom behavior and social skills specific to the participating child, C) Information specific to the participating child, D) Parent/guardian involvement, E) Classroom characteristics, F) School climate and environment, G) General information about teacher

Table S1. Summary of information collected in the Fragile Families and Child Wellbeing Study between child birth and age 9. Section letters are not always consecutive in the questionnaires. Full questionnaires are available at `https://fragilefamilies.princeton.edu/documentation`.

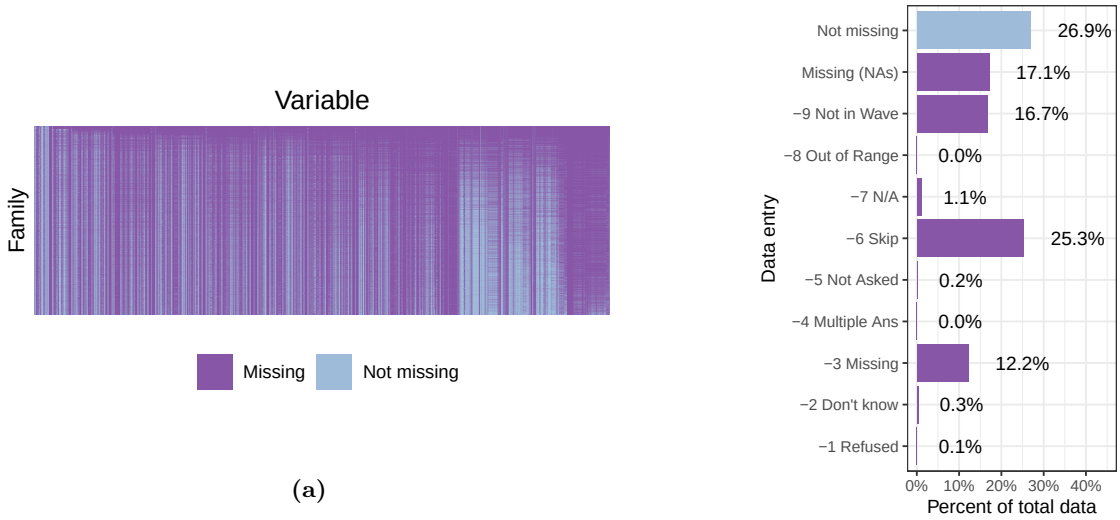


Fig. S1. Missing entries in the Fragile Families Challenge background dataset. The Fragile Families Challenge background dataset had 4,242 rows and 12,942 columns (plus an ID number). Of the approximately 55 million distinct data entries, about 73% were missing. There were many types of missing entries.

values stored in separate columns for each of these subgroups. Another source of intentional skips is that some questions were asked only of the subpopulation to whom they applied. For example, only mothers not married to the child's father at the birth were asked the likelihood that they would marry him. It is also the case that some of the entries coded as "not in wave" correspond to intentional skip patterns that caused some individuals to be excluded from certain data collection modules entirely. For instance, those who did not meet a threshold of hours spent with a child care provider are coded as "not in wave" on the childcare provider survey modules. In summary, some of the missingness in the data matrix is a consequence of the survey design—gathering an enormous amount of information from respondents over many years by only asking the relevant questions—rather than an indicator of poor data quality.

Of the missing data that were not an intentional component of the survey design, the share from survey nonresponse is arguably low compared to what one might expect for a study following a longitudinal research design. For example, of the 4,898 mothers that began the study, 3,515 (72%) were interviewed in wave 5, 9 years after they were initially recruited to participate. The study minimizes nonresponse by employing multiple strategies to reach families, including phone tracking, neighborhood canvassing, social media outreach, and using alternate contacts from families and friends. Data collection in each city typically last 9 months to 1 year.

The outcome data used in the Fragile Families Challenge was measured at child age 15. The full list of data collection modules at year 15 are presented in Table S2. The six outcomes that we selected

Data module	Child age	Sections
Mother/Primary care giver	15	A) Non-parental caregiver, B) Youth health and behavior, C) Youth education, D) Family life and parenting, E) Household structure and family relationships, F) Nonresident biological parent, G) Coresidental biological father or coresideential/nonresidental partner H) Primary care giver health and behavior, I) Social environment and informal support J) Housing and programs K) Education, employment, and income
In-home assessment	15	A) Observation checklist, B) Common areas, C) Interior of house or apartment, D) Youth's appearance, E) Home scale, F) Youth emotion and cooperation, G) Ending
Child	15	A) Introduction, B) Education, C) Family relationships, D) Health and health behavior, E) Neighborhood, F) Risky behaviors - sexual activity and illegal drug use

Table S2. Summary of information collected in the Fragile Families and Child Wellbeing Study at child age 15. Full questionnaires are available at `https://fragilefamilies.princeton.edu/documentation`.

to be the focus of the Challenge were: 1) child grade point average (GPA), 2) child grit, 3) household eviction, 4) household material hardship, 5) primary caregiver layoff, and 6) primary caregiver participation in job training. We selected these six variables for substantive and methodological reasons. Substantively, we picked variables from a variety of domains where we thought that good predictions would be useful for subsequent empirical research. Methodologically, we wanted a variety of variable types—such as binary or continuous and about the child, household, or primary caregiver—so that we could study the relationship between the type of outcome, its overall predictability, and the best methods for predicting it. The variables we refer to as continuous actually take a discrete set of numeric values (Table S3). We did not pick these six outcomes because we thought that they would be especially easy or difficult to predict.

The operationalization of each of these six outcomes varies across the scientific literature, and Table S3 describes the operationalization we used. Two outcomes in particular—eviction and grit—warrant further discussion. The measure of eviction in the Fragile Families study includes eviction for nonpayment of rent or mortgage, regardless of whether a court ordered the eviction or a landlord carried it out informally [7, 17]. Other research, however, focuses on formal court-ordered evictions for any reason [6]. Also, the Fragile Families measurement of grit is different from the measure proposed in [8]. More specifically, [8] proposes a grit scale consisting of six items related to consistency of interest and six items related to perseverance of effort. The Fragile Families study scale is shorter—four items—and was designed with adolescent school outcomes in mind. Two items ("I finish whatever I begin"; "I am a hard worker") are exactly as in the original scale for perseverance of effort. One item on the Fragile Families study scale ("Once I make a plan to get something done, I stick to it") is a simplified version of one of the original items about consistency of interests ("I have difficulty maintaining my focus on projects that take more than a few months to complete"). Likewise, the Fragile Families study scale includes an item focused on schoolwork ("I keep at my schoolwork until I am done with it"), which is a more targeted version of an item from the original perseverance scale ("I am diligent"). A final difference is that [8] propose a scale with five answer choices ("not at all like me" to "very much like me") whereas the Fragile Families study scale involves four choices

("strongly disagree" to "strongly agree").

Following standard practice for the common task method, we divided the outcome data into three disjoint sets: training, leaderboard, and holdout. To split the outcome data, we started by acquiring the information needed to construct the six outcome variables (Table S3); these data were available only to members of the Fragile Families team. Next, we randomly split the outcome data using systematic sampling [25]. More specifically, we first sorted all observations by city of birth, parents' relationship status at the birth, mother's race, whether at least one outcome was non-missing, and then the outcomes in the following order: eviction, layoff, job training, GPA, grit, and material hardship. In the sorted data, we grouped observations into sets of 8 sequential observations. Then, we randomly assigned four, one, and three of the observations to the training, leaderboard, and holdout sets.

Figure S2 shows the distribution of outcomes in the training data, and Table S4 shows the number of non-missing cases in each of the training, leaderboard, and holdout sets. Cases with missing outcomes were not used when measuring the mean squared error of the predictions in the holdout set. In the leaderboard set only, we imputed missing values on the outcome variables by taking a random sample (with replacement) from the distribution of observed outcomes in the leaderboard set. Because these random draws are unpredictable by construction, we could assess whether respondents were overfitting to the leaderboard (i.e., submitting numerous queries and updating their models based on leaderboard score) by measuring how well participants could predict these random values.

Age 15 outcome	Age 9 questions	Response values	Reporter	How aggregated
GPA	At the most recent grading period, what was your grade in 1. English or language arts? 2. Math? 3. History or social studies? 4. Science?	1. A 2. B 3. C 4. D or lower	Child	Reverse-coded and averaged. Marked NA if any item missing due to no grade, pass/fail, refusal, don't know, homeschooled, or not interviewed.
Grit	Thinking about how you have behaved or felt during the past four weeks, please tell me whether you strongly agree, somewhat agree, somewhat disagree, or strongly disagree with the following statements. 1. I keep at my schoolwork until I am done with it. 2. Once I make a plan to get something done, I stick to it. 3. I finish whatever I begin. 4. I am a hard worker.	1. Strongly agree 2. Somewhat agree 3. Somewhat disagree 4. Strongly disagree	Child	Reverse-coded and averaged. Marked NA if any item missing due to refusal, don't know, or not interviewed.
Material hardship	We are also interested in some of the problems families have making ends meet. In the past twelve months, did you do any of the following because there wasn't enough money? 1. Did you receive free food or meals? 2. Were you ever hungry, but didn't eat because you couldn't afford enough food? 3. Did you ever not pay the full amount of rent or mortgage payments? 4. Were you evicted from your home or apartment for not paying the rent or mortgage? 5. Did you not pay the full amount of gas, oil, or electricity bill 6. Was your gas or electric services ever turned off, or the heating oil company did not deliver oil, because there wasn't enough money to pay the bills? 7. Did you borrow money from friends or family to help pay bills? 8. Did you move in with other people even for a little while because of financial problems? 9. Did you stay at a shelter, in an abandoned building, an automobile or any other place not meant for regular housing, even for one night? 10. Was there anyone in your household who needed to see a doctor or go to the hospital but couldn't go because of the cost? 11. Was your telephone service (mobile or land line) cancelled or disconnected by the telephone company because there wasn't enough money to pay the bill?	0. Event did not occur 1. Event occurred	Child's primary caregiver	Averaged. Marked NA if any response missing due to refusal, don't know, or not interviewed.
Eviction	1. In the past twelve months, were you evicted from your home or apartment for not paying the rent or mortgage? 2. (If no above:) Since [month and year of interview at approximately child age 9], were you evicted from your home or apartment for not paying the rent or mortgage?	0. No 1. Yes	Child's primary caregiver	If no to both questions, 0. If yes to either question, 1. Marked NA if missing due to refusal, don't know, or not interviewed.
Layoff	Since [month and year of interview at approximately child age 9], have you been laid off from your employer for any time?	0. No 1. Yes	Child's primary caregiver	Marked NA if missing due to refusal, don't know, or not interviewed.
Job training	Since [month and year of interview at approximately child age 9], have you taken any classes to improve your job skills, such as computer training or literacy classes?	0. No 1. Yes	Child's primary caregiver	Marked NA if missing due to refusal, don't know, or not interviewed.

Table S3. Outcome variables measured at child age 15.

Outcome	Training	Leaderboard	Holdout
GPA	1,165	304	886
Grit	1,418	362	1,075
Material hardship	1,459	375	1,099
Eviction	1,459	376	1,103
Layoff	1,277	327	994
Job training	1,461	376	1,104
Total possible	2,121	530	1,591

Table S4. Number of non-missing cases for each outcome in the training, leaderboard, and holdout sets.

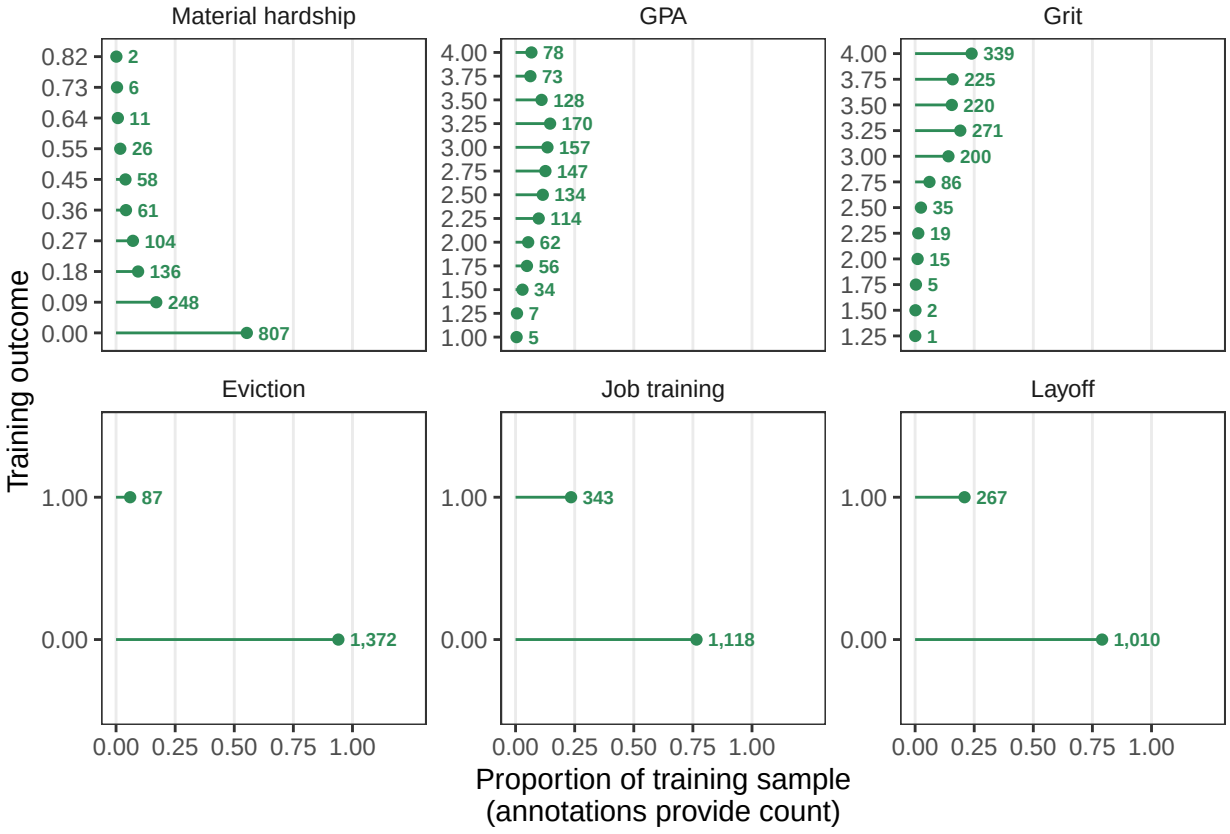
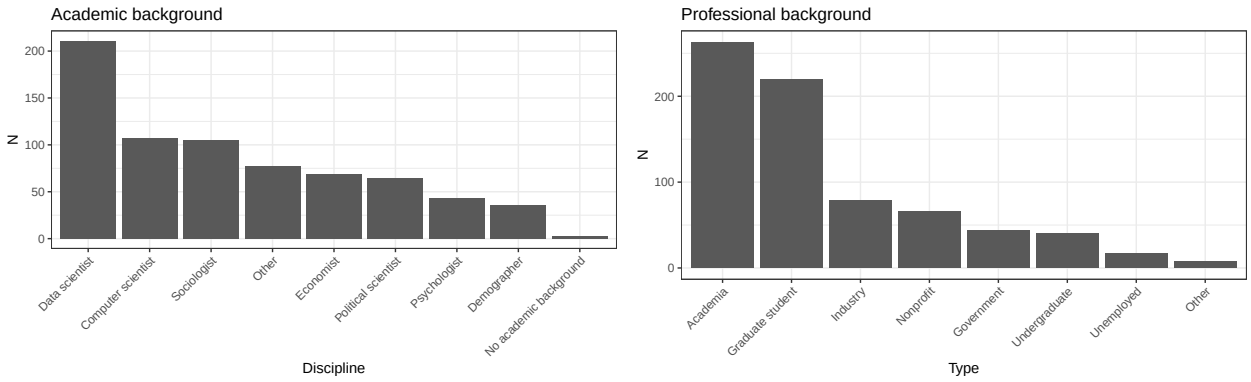


Fig. S2. Distribution of outcomes in the training set. The number of missing cases for each outcome varied (Table S4) and are excluded here.

S1.2 Evaluation metric: Mean squared error

There are many potential metrics by which to evaluate predictive performance, and different metrics may lead to different conclusions [13]. For the Challenge, we wanted an evaluation metric that met three criteria: 1) familiar to participants, 2) applicable to both binary and continuous outcomes, and 3) aligned with how we plan to use the predictions in further research. Mean squared error (MSE) meets these criteria. MSE is widely used for outcomes that are continuous (e.g., ordinary least-squares regression minimizes squared error) and binary (e.g., it is deeply related to the Brier score [5]). Additionally, we plan to use these predictions to identifying families with particularly unexpected outcomes so that we can collect more information from these families. MSE is suited to this goal because it heavily penalizes large errors, thereby encouraging predictions that produce large errors only when a case is truly unexpected.

We present results in terms of R^2_{Holdout} (Eq. 1, main text), a transformation of MSE that increases interpretability and comparability across outcomes. There are many definitions of R^2 in the literature, and our definition is based on the recommendation of [16]. We normalized the squared prediction errors by the squared prediction error from a null model: predicting the mean of the training data. An alternative approach would normalize by the squared prediction error when predicting the mean of the holdout data. Both approaches produce similar results because the training and holdout sets are similar because of the way we created them.



(a) **(b)**

Fig. S3. Self-reported disciplinary affiliation and professional background of applicants. Each applicant could select as many of these as applied. See [18] for a copy of the application.

S1.3 Procedures of the Fragile Families Challenge

We recruited participants to the Fragile Families Challenge through a variety of approaches including: contacting colleagues, working with faculty who wanted their students to participate, and hosting getting started workshops at universities, in courses, and at scientific conferences. Anyone who wanted to participate in the Challenge needed to apply for data access [18]. During the application process, participants provided informed consent to the procedures of the Fragile Families Challenge.

Applicants came from a wide range of disciplinary backgrounds (Fig. S3a). By far the most common self-reported disciplinary affiliation reported was "data scientist", and there was also a substantial representation from social science disciplines and other disciplinary backgrounds. 64% of participants reported affiliation with more than one discipline. Applicants also came from a variety of sectors and career stages (Fig. S3b). Finally, applicants reported a variety of motivations such as general interest in the topic and to improve the lives of disadvantaged children (Fig. S4).

In order to make a submission to the Challenge, each team needed to create an account on our Challenge platform, which was a customized instance of CodaLab[1], open-source software designed to manage research projects using the common task method. Each submission was required to include three main elements: predictions for all six outcomes for all 4,242 cases, the code used to generate those predictions, and a narrative explanation of the strategy used to generate the predictions. When a submission was uploaded, it was automatically assessed to see if it met the submission guidelines. Submissions meeting the guidelines were automatically scored with the leaderboard data. Each account was permitted to upload 10 submissions per day. Although participants were required to upload predictions for all six outcomes,

[1] `http://codalab.org/`

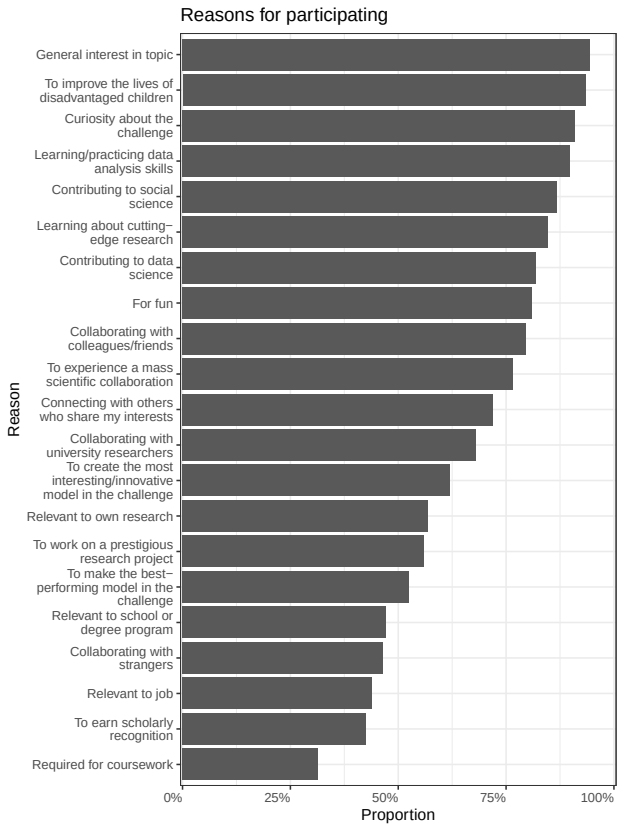


Fig. S4. Self-reported motivations of Fragile Families Challenge applicants. Each applicant could select as many of these as applied. See [18] for a copy of the application.

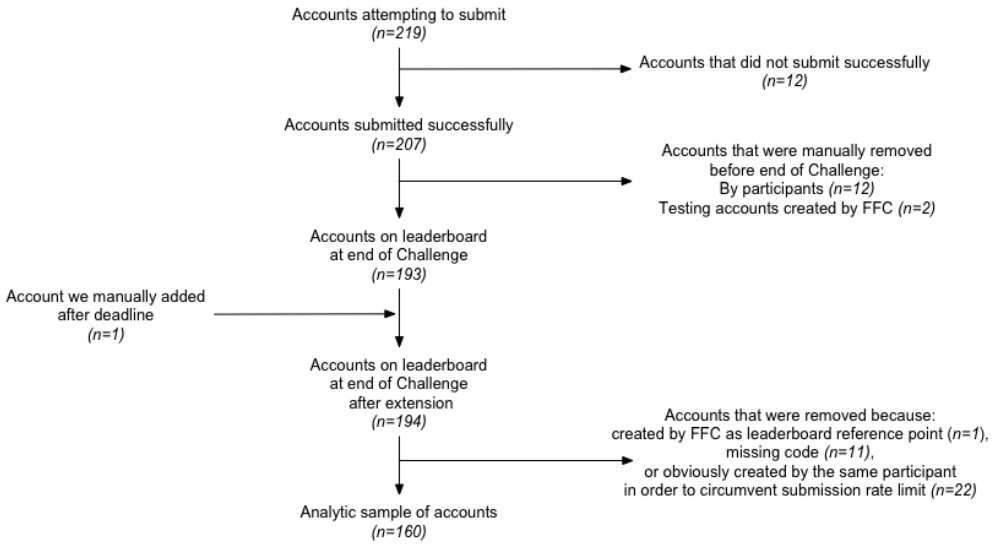


Fig. S5. Flow diagram of criteria producing the sample of 160 accounts that made a valid submission.

some participants chose to focus on a subset of outcomes (e.g., just GPA or just continuous outcomes) and uploaded the mean of the training data for the other outcomes.

At the end of the Challenge, each team selected one submission to be on the leaderboard as their final submission, with the default choice being the most recent submission. Our analyses began with these final submissions. All participants agreed—during the application processes—that all their submissions would be open-sourced at the end of the Challenge [18].

The first submissions to the Challenge were on March 5, 2017, as part of a controlled roll-out in a class at Princeton University. The Challenge officially opened to the larger research community on March 21, 2017 and ended on August 1, 2017. We opened the holdout data and scored these final submissions on September 11, 2017.

There were a total of 219 accounts that attempted to submit predictions to the Challenge, of which 160 accounts were considered valid submissions. Some of our analysis uses a subset of these accounts that made a qualifying submission, which is defined to be a valid submission that scored better than the mean of the training data. Figure S5 provides information about why some accounts were excluded. Table S5 provides additional information about the valid submissions and qualifying submissions. The set of accounts does not have a one-to-one relation to the set of participants. Some accounts are associated with teams of participants and a small number of participants contributed to more than one account.

Restriction	Material hardship	GPA	Grit	Eviction	Job training	Layoff
Valid submissions	160	160	160	160	160	160
Submissions different from the mean of the training data	122	128	121	111	117	112
Qualifying submissions (better than the mean of the training data)	92	98	65	48	42	42

Table S5. Submission restrictions. Many of our analyses use either the full set of 160 valid submissions, the restricted set of submissions for which at least one prediction was at least 10^{-4} away from the mean of the training data, or the restricted set of qualifying submissions, which are the submissions that were more accurate than the mean of the training data.

S2 Analysis of the best performance

S2.1 Construction of confidence intervals

The best R^2_{Holdout} scores observed in the Challenge are descriptive quantities known with certainty. However, we may also view the holdout set as one random sample from the population of families that were eligible for the Fragile Families and Child Wellbeing Study. In this case, the observed scores can be considered estimates that would vary from holdout sample to holdout sample. The 95% confidence intervals in Fig. 3 capture this variability by bootstrapping. More specifically, we mimic a simplified version of the Fragile Families and Child Wellbeing Study sampling process by drawing 10,000 simple random samples with replacement from the holdout set. Within each replicate sample, we score all submissions and record the highest observed performance: $\max_j \hat{R}^{2*}_j$ (using $*$ to denote evaluation within a bootstrap sample). Finally, we report the .025 and .975 quantiles of the $\max_j \hat{R}^{2*}_j$ scores to produce a 95% confidence interval.

S2.2 Winning submissions slightly outperform benchmark models but many submissions do not

The winning submissions in the Fragile Families Challenge did not produce accurate predictions as measured by R^2_{Holdout}. To provide a comparison, we worked with domain experts to create a simple benchmark model that used standard methods and a small number of commonly used variables. More specifically, we created a benchmark model that used linear regression (for continuous outcomes) and logistic regression (for binary outcomes) and four predictors: mother's race/ethnicity (black, Hispanic, or white/other), mother's level of education (less than high school, high school degree, some college, college), and mother's marital status at the birth of the child (married, cohabiting, or other), and a measure of the outcome—or a closely related proxy—collected at child age nine (Table S6). Further, to provide additional points of comparison, we also assessed other closely related benchmark models, including those that use other statistical learning procedures or subsets of the predictors.

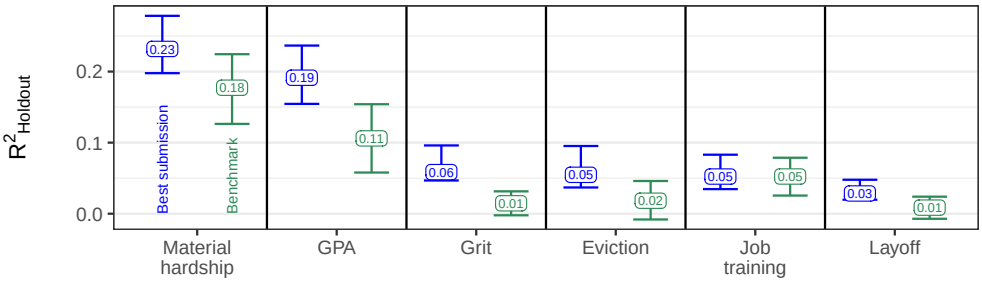
To construct our benchmark models, we imputed missing data using the `Amelia` package [14]. In the imputation model, we treated mother's race and mother's marital status as nominal variables, mother's education as an ordinal variable, and the lagged outcome variable as a continuous variable. We included the six outcomes in the imputation model in order to improve efficiency, but observations with missing outcomes were excluded from model fitting. After preparing the predictor variables, for each of the six outcomes separately, we learned the relationship between the predictors and the Challenge training data using linear regression for continuous outcomes and logistic regression for binary outcomes. After fitting the models, we used them to make predictions about the outcomes for the holdout data. Next, we truncated predictions to the range of possible outcome values (e.g., between 0 and 1 for eviction).

Figure S6 Panel A plots the performance of the best model in the Challenge and the main benchmark model. These values were substantively similar and can be compared in several ways. The absolute difference between the values $\left(R^2_{\text{Holdout, Best}} - R^2_{\text{Holdout, Benchmark}}\right)$ is presented in Fig. S6 Panel B. The percentage of the gap between the benchmark model and perfect prediction that was closed by the best model $\left(\frac{R^2_{\text{Holdout, Best}} - R^2_{\text{Holdout, Benchmark}}}{1 - R^2_{\text{Holdout, Benchmark}}}\right)$ is presented in Fig. S6 Panel C. Although the preferred metric for comparing $R^2_{\text{Holdout, Best}}$ and $R^2_{\text{Holdout, Benchmark}}$ varies depending on context, we conclude that the differences between the performances were small in absolute terms. Small absolute improvements can be consistent with large relative improvements when the benchmark model has very poor performance (e.g., grit, layoff). The best submission for layoff, for instance, achieved R^2_{Holdout} four times that achieved by the benchmark. However, in these cases, the absolute performance of the best submission was still quite poor. Overall, we conclude that the best submission had poor predictive performance that was only slightly better than a simple benchmark.

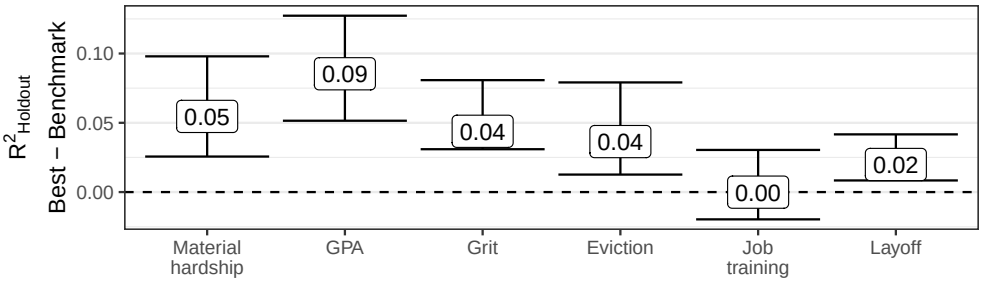
Age 15 outcome	Age 9 questions	Response values	Reported by	How aggregated
GPA	Overall, how would you rate this child's academic skills in each of the following areas, compared to other children of the same grade level? 1. Language and literacy skills 2. Science and social studies 3. Mathematical skills	1. Far below average 2. Below average 3. Average 4. Above average 5. Far above average	Teacher	Averaged
Grit	1. Child persists in completing tasks. 2. Child fails to finish things he or she starts. 3. Child does not follow through on instructions and fails to finish homework.	1. Never 2. Sometimes 3. Often 4. Very often	Teacher	Averaged with (2) and (3) reverse-coded.
Material hardship	Same as age 15 questions	0. Event did not occur 1. Event occurred	Mother if child lived with mother at least half the time. Otherwise, father if child lived with mother at least half the time. Otherwise, a non-parental primary caregiver.	Averaged.
Eviction	In the past twelve months, were you evicted from your home or apartment for not paying the rent or mortgage?	0. No 1. Yes	Mother if child lived with mother at least half the time. Otherwise, father if child lived with mother at least half the time. Otherwise, a non-parental primary caregiver.	NA
Layoff	Last week, did you do any regular work for pay? Include any work you might have done in your own business or military service where you got a regular paycheck.	0. No 1. Yes	Mother if child lived with mother at least half the time. Otherwise, father if child lived with mother at least half the time. Missing if child has a non-parental primary caregiver.	NA
Job training	During the last four years, have you taken any classes to improve your job skills, such as computer training or literacy classes?	0. No 1. Yes	Mother if child lived with mother at least half the time. Otherwise, father if child lived with mother at least half the time. Missing if child has a non-parental primary caregiver.	NA

Table S6. Lagged outcome variables—or closely related proxies—measured in the age 9 survey and used in the benchmark models.

A) Performance of benchmark and best submissions.



B) Absolute improvement of the best over the benchmark.



C) Proportion of unpredicted component closed by best.

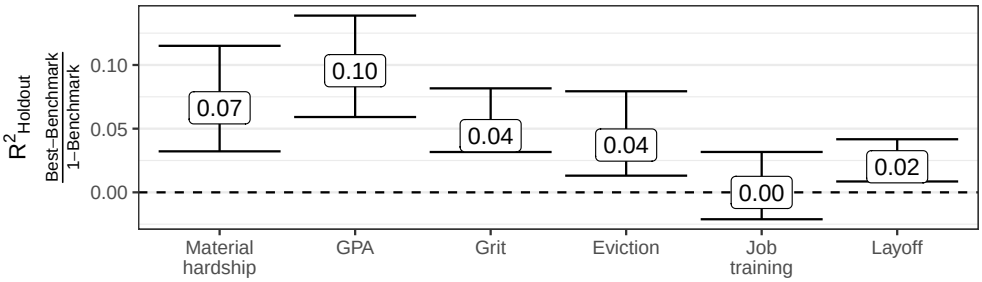


Fig. S6. Maximum R^2_{Holdout} relative to benchmark models. Horizontal lines indicate the value that would be realized if the best submission and the benchmark had equal performance.

To provide additional points of comparison, we also considered benchmarks with different predictor sets (e.g., just information collected at birth or just the lagged outcome) and with a different method of statistical learning (e.g., random forest). Figure S7 shows the predictive performance of these alternative benchmarks. For some outcomes (e.g., material hardship), the predictive performance of the benchmark can be almost matched by a subset with only one predictor (a proxy variable measured in the prior wave), and adding three demographic predictors only adds minimally to prediction. For other outcomes (e.g., GPA), the full benchmark can almost be matched by a subset with only the three demographic predictors, and adding a lagged proxy adds very little. For simplicity, the main text presents only the benchmark with all four predictors.

We note that a small difference in predictive performance, as measured by R^2_{Holdout}, does not imply that the best submission and the benchmark made similar predictions for any given observation (Fig. S8). The correlation between the benchmark and best predictions ranged from 0.22 (Eviction) to 0.64 (GPA).

Finally, we note that although the best submissions to the Challenge performed better than the benchmark, many submissions fell short of this benchmark model. Between 31% (layoff) and 100% (job training) of qualifying submissions were worse than the benchmark model (Fig. S9), despite the fact that many of these submissions were quite complex.

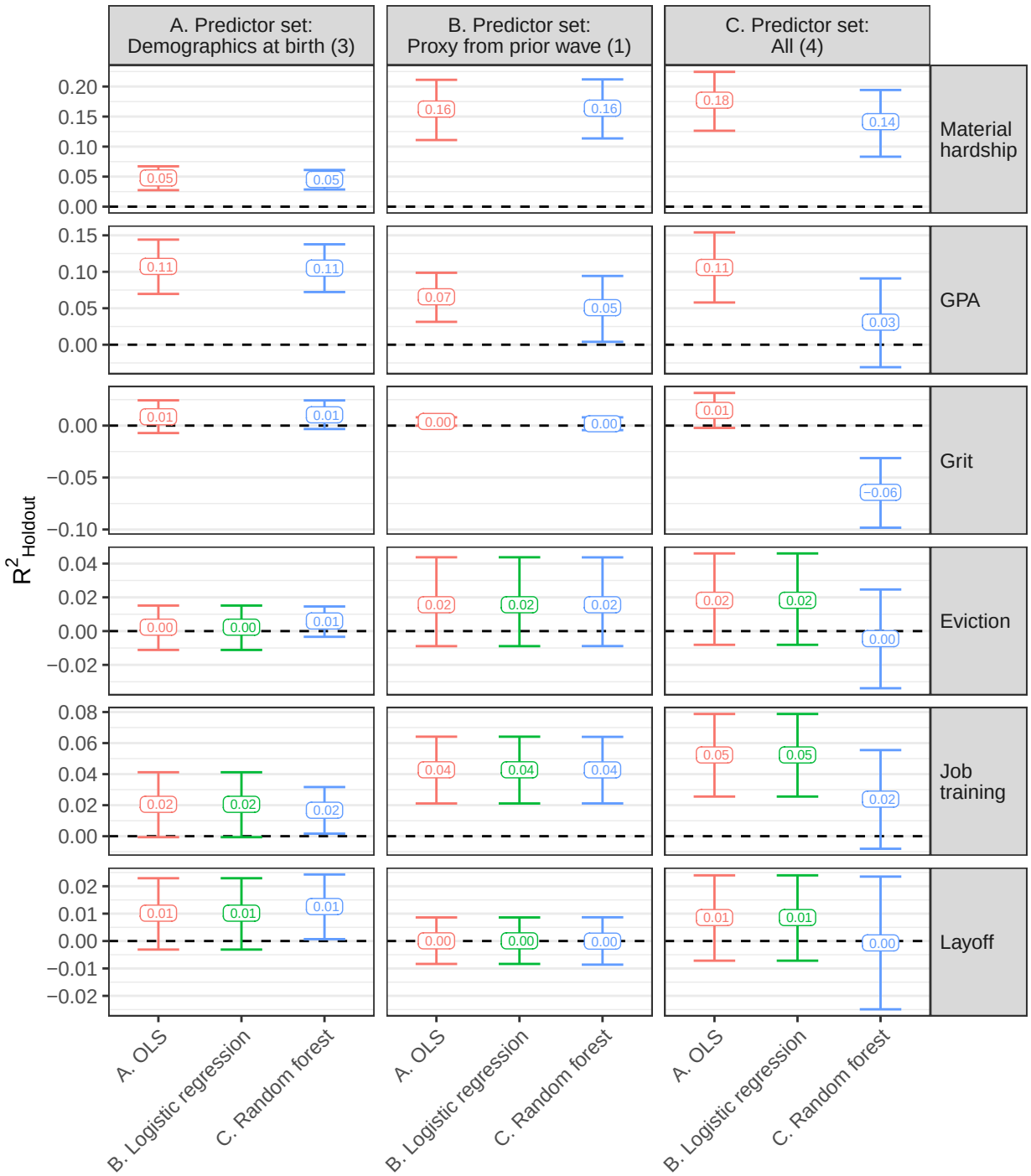


Fig. S7. Alternative benchmark models. The benchmark results reported in Figure S6 use the full predictor set (4 variables) and OLS (continuous outcomes)/logistic regression (binary outcomes).

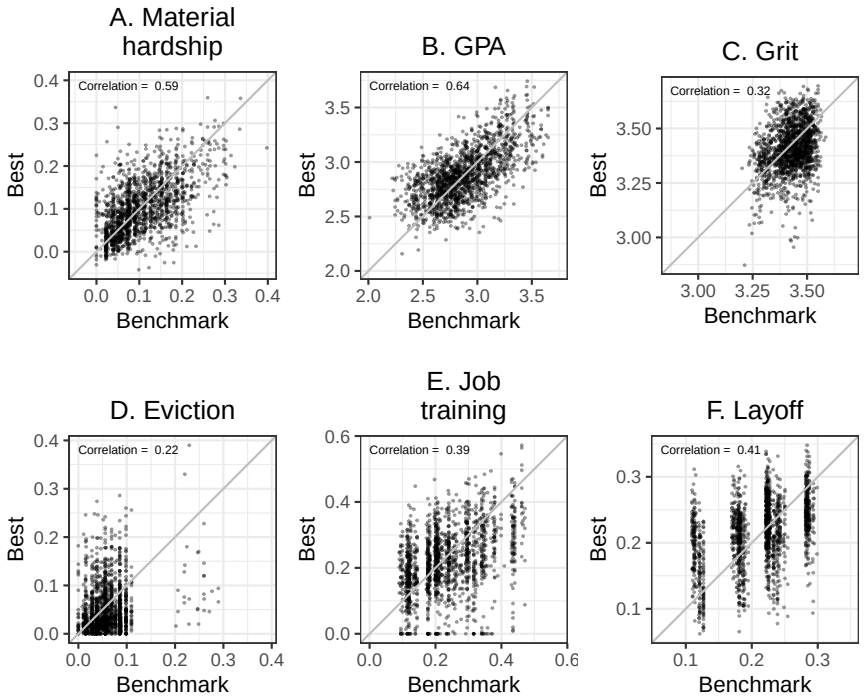


Fig. S8. Comparison of individual-level predictions for the best submissions and benchmark models. Although the best submissions and benchmark models have similar R^2_{Holdout}, they do not have equivalent individual-level predictions.

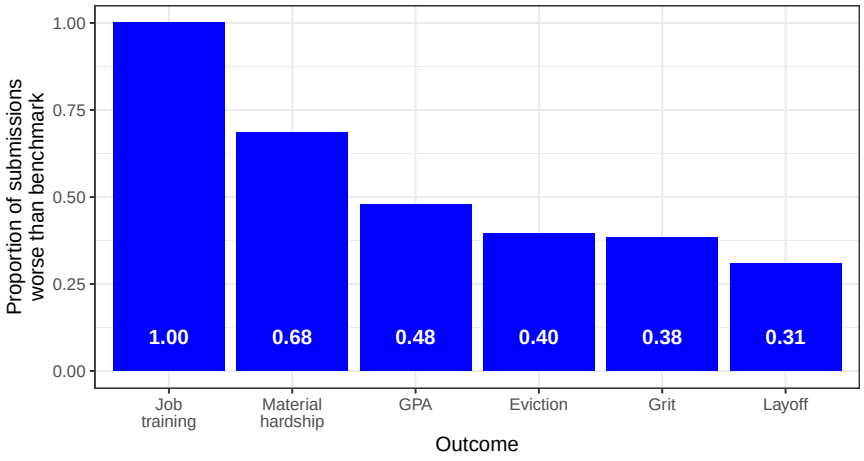


Fig. S9. Many qualifying submissions have worse predictive performance than the benchmark model.

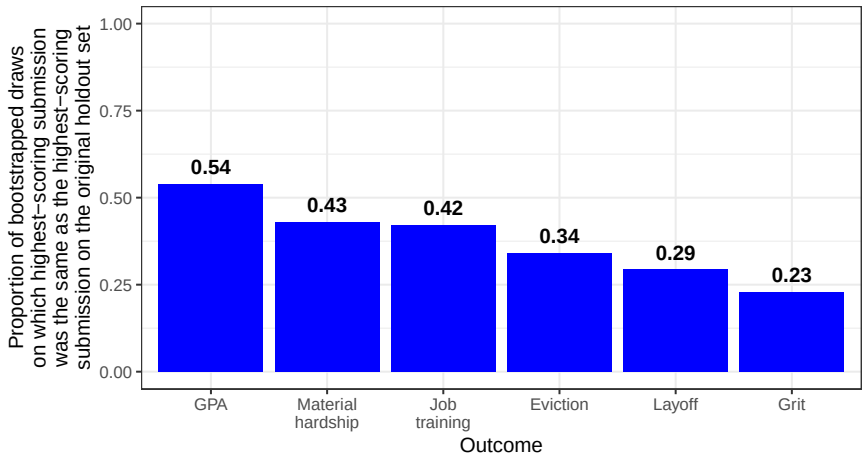


Fig. S10. Winner is uncertain. The account with the best score on the holdout set does not always have the best score on bootstrap samples of the holdout set, suggesting the winner may be lucky in this particular draw of the holdout set.

S2.3 Winner may be different in a new holdout set from the population

The winning submission to the Challenge is an unambiguous descriptive quantity: the submission with the best score on the holdout set. However, the chosen winner might have been different if we had evaluated on a different holdout from the same data generating process. To simulate this possibility, we constructed 10,000 bootstrap samples of the holdout set. Figure S10 shows the proportion of bootstrap samples for which the best-scoring submission was the same submission that scored the best on the Challenge holdout set. This proportion ranges between 23% and 54%, suggesting that the chosen winner depended partly on the luck of the holdout set.

S2.4 The best holdout score is optimistic for out-of-sample performance

In the main text we reported the R^2_{Holdout} of the submission that scored best in the holdout set. In other words, we evaluated all the submissions in the holdout set and picked the one with the best score. While this approach accurately describes what happened in the Fragile Families Challenge, it provides an optimistic estimate of the best performance that would be measured if we could evaluate the submitted models on the full population from which the holdout set was drawn.

To illustrate this concern, consider a setting in which five researchers contribute models that are all different but equally good: they would all achieve $R^2 = 0.5$ on average across possible draws of a holdout sample from the population. In one draw of the holdout set, we might observe the scores $\hat{R}^2 = \{0.4, 0.45, 0.47, 0.52, 0.56\}$, and we would report 0.56 as the best score in this specific holdout set. Next, imagine that we drew a new holdout set from same population and again scored the submissions from these five model. We would again expect to get $\hat{R}^2$ values of around 0.5, with the best submission slightly above 0.5. However, the submission that originally scored 0.56 is likely to have a score closer to 0.5 in this new holdout set. If we could evaluate the models on a very large holdout set, all the scores would converge to their true value of 0.5. When the holdout set is of limited size, however, the best $\hat{R}^2$ in that set is likely to be at least partly lucky in the sense that it happened to predict well for the observations in that particular holdout set. Likewise, the the maximum R^2_{Holdout} in the Challenge is likely optimistic for the best performance that would be achieved if models were evaluated on the full population.

In order to assess the possible magnitude of this issue in the Challenge, we developed and used an alternative procedure for estimating the maximum performance, which is likely pessimistic and which has higher variance. The alternative procedure splits the holdout set in half in order to select the best submission on a selection sample and then assess the performance of that best submission on an independent evaluation sample. In the simple five-submission example above, regardless of which model was selected in the selection sample, in expectation this model would achieve $\hat{R}^2_{\text{Evaluation}} = 0.5$. Using separate samples for selection and evaluation allows us to combat a "winner's curse" phenomenon in which the submission that happens to do the best in the selection sample is likely to regress toward the mean in a new sample.

Our alternative procedure targets the following estimand: the expected predictive performance of a procedure that uses a random half of the holdout set to select the best model and uses observations not used for training or selection to evaluate performance. We estimate this quantity by randomly splitting the Challenge holdout set into two halves: a selection set (795 observations) and an evaluation set (796 observations). We select the best submission in the selection set and then evaluate its $\hat{R}^2_{\text{Evaluation}}$ in the evaluation set. Because the selection and evaluation sets are independent samples from the population, the

performance in the evaluation set of the submission chosen in the selection set is an unbiased estimator of the ability of this submission to predict new observations.

To avoid an unlucky split of the holdout set, we repeat this procedure 100 times, under different partitions of the holdout observations to selection and evaluation, and average the results.

$$\hat{\tau} = \frac{1}{100} \sum_{i=1}^{100} \hat{R}^2_{\text{Evaluation},i} \tag{S1}$$

To produce a conservative confidence interval for this estimand, we calculate the analytical sample variance of the mean squared error within each split. We then average over splits to produce an estimated variance, which is upwardly biased for the true variance of the estimator because it does not account for the reduction in the variance achieved by averaging over 100 splits. We used this estimated variance (which is conservative) along with the typical normal approximation to produce a conservative 95% confidence interval around the overall point estimate.

Figure S11 (lines A and B) presents our results. For all six outcomes, this alternative procedure produces slightly lower point estimates. Two possible sources of this difference are: 1) with only half of the holdout set, we are more likely to select an inferior model and 2) when selecting and evaluating on the full holdout set, our original estimator is optimistically biased for out-of-sample performance. For all six outcomes, the alternative procedure also produces wider confidence intervals. Two possible sources of this difference are: 1) with a smaller evaluation set we are more uncertain about performance and 2) a conservative bias in how we estimates variance when averaging over split.

To conclude, in the main text we reported the R^2_{Holdout} of the submission that scored best in the holdout set. This quantity describes what happened in the Challenge but is optimistic for the best performance that would be measured if we could evaluate the submitted models on the full population. To assess the magnitude of this problem in the Challenge, we developed and deployed an alternative procedure that is likely pessimistic and produces estimates with higher uncertainty. Both procedures produce essentially the same results (Fig. S11 (lines A and B)). We chose to present the results from a single holdout set in the main paper for two reasons: 1) they are simpler and 2) an optimistic bias—predictive performance tends to look better than it is—created by selecting and evaluating on one holdout set runs counter to the main claim of the paper that predictions were inaccurate. If anything, the chosen best submission would be—in expectation—even more inaccurate if re-evaluated on a new holdout set from the same population.

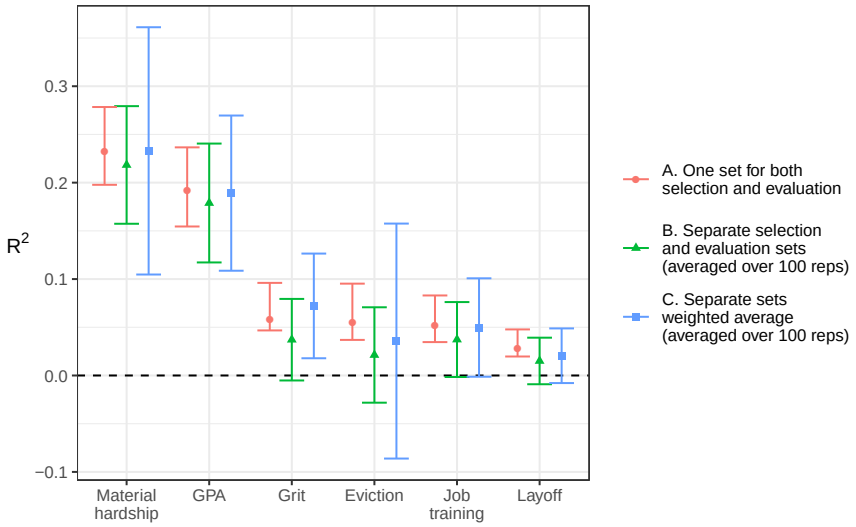


Fig. S11. Maximum R^2_{Holdout} relative to alternative estimators of out-of-sample error. Model (A) is the best-scoring submission to the Challenge, evaluated on the same set used to select it. Error bars approximate the sampling distribution of this estimate by the middle 95% of bootstrap draws, with selection and evaluation performed within each bootstrap draw. Model (B) splits the holdout set into selection and evaluation samples, selects the best model on the selection sample, and evaluates it on the evaluation sample, and averages over 100 repetitions of this procedure as described in Section S2.4. This approach avoids overfitting in the model selection stage. Model (C) uses the same sample splitting procedure as (B) but creates a weighted average of submissions as described in Section S2.5. Confidence intervals in (B) and (C) are analytic as described in Sections S2.4 and S2.5.

S2.5 A weighted average of submissions does not perform better

Rather than focusing on the best submission, it might be possible to achieve better performance by combining several submissions. We consider one simple strategy: combining submissions by a weighted average, a technique sometimes called "stacking" [4, 40]. A weighted average would improve predictive performance if the truth lies somewhere between the submitted predictions, or formally within their convex hull [4]. As a concrete example, this would be true if one submission tended to under-predict the same observations that another submission over-predicted, so that an average of the two would be close to the truth. However, because the predictions were very similar across submissions, an ensemble combining several submissions is unlikely to substantially outperform the best individual. Nonetheless, this section reports results for a weighted average.

We construct a weighted average with weights $\vec{\beta}$ learned with a constraint that all weights are non-negative and sum to 1 (Eq. S2). We impose the regularization constraint $\lambda = 0.01$ to avoid problems due to highly correlated predictors.

$$\overbrace{\hat{\vec{\beta}}}^{\substack{\text{Estimated}\\ \text{stacking}\\ \text{weights}}} = \underset{\vec{\beta}:\vec{\beta}\geq 0,|\vec{\beta}|_1=1}{\operatorname{argmin}} \left(\left| \vec{y} - \hat{\mathbf{y}}^{\text{Submitted}} \vec{\beta} \right|_2 + \lambda |\vec{\beta}|_2 \right) \tag{S2}$$

The stacked predictor is the weighted average of all submissions that results.

$$\hat{\vec{y}}_{\text{Stacked}} = \hat{\mathbf{y}}^{\text{Submitted}} \hat{\vec{\beta}} \tag{S3}$$

Because there is a risk of overfitting by learning the weights in the same holdout set used to evaluate the final model, we conduct this entire procedure within the selection-evaluation split procedure as described in Section S2.4. As in that section, we average results over 100 stochastic splits and calculate a conservative 95% confidence interval. After performing this procedure, we find that the performance of this weighted average is comparable to that of the overall best individual submission (Fig. S11). In other words, combining submissions in this way did not substantially improve predictive performance.

S3 Patterns in predictions and prediction errors

In the main text, we report a number of patterns about the predictions and predictions errors. In this section, we provide additional information about those patterns.

Figure S12 compares all qualifying submissions (those that beat the baseline submission) for each outcome. It shows that the distance between the most divergent submissions was less than the distance between the best submission and the truth. In other words, the submissions were much better at predicting each other than at predicting the truth.

Figure 4 in the main text presents a heatmap of the squared prediction error for all qualifying submissions for each outcome. Here we now provide further description of three patterns of prediction errors for these submissions.

First, for each outcome, the squared prediction error is strongly related to the family being predicted and weakly related to the technique used to make that prediction. This pattern is apparent visually in Figure 4 in the main text: for each outcome, many observations are well predicted by all submissions and some observations are poorly predicted by all submissions. One way to quantify the pattern in squared prediction errors is to compare the model fit of two linear regression models, one with fixed effects for each family:

$$e_{ijk}^2 = \alpha_i + \epsilon_{ijk} \tag{S4}$$

and one with fixed effects for each submission:

$$e_{ijk}^2 = \eta_j + \delta_{ijk}. \tag{S5}$$

where e_{ijk}^2 is the squared error for family i, submission j, and outcome k

$$e_{ijk}^2 = (\hat{y}_{ijk} - y_{ik})^2. \tag{S6}$$

The R^2 values for these two models are reported in Table S7 and show that for each outcome the squared prediction error is strongly related to the family being predicted and weakly related to the technique used to the make the prediction.

A second pattern in the prediction errors is that the observations that are hardest to predict are those that are far from the mean of the training data. One way to summarize the difficulty of predicting a particular

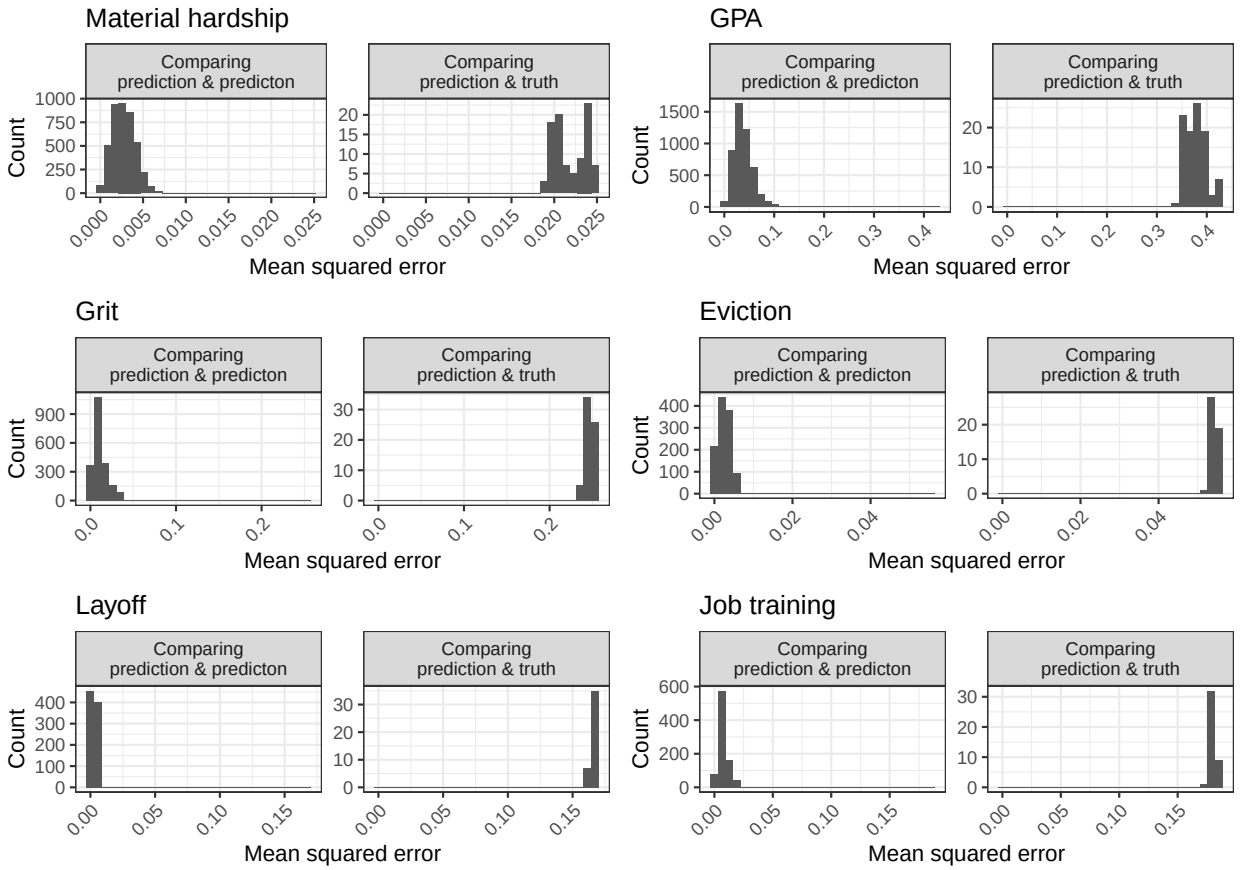


Fig. S12. Comparing submissions to each other and the truth. For each outcome, the histograms show the mean squared error between all pairs of submissions (left panel) and the mean squared error between all submissions and the truth (right panel). For all six outcomes, the submissions were much closer to each other than they were to the truth, where distance is measured by mean squared error. These results only include qualifying submissions and are restricted to observations with non-missing outcome data.

Outcome	Family fixed effects model	Accounts fixed effects model
Material Hardship	0.925	0.002
GPA	0.897	0.002
Grit	0.968	0.0001
Eviction	0.986	0.00002
Job training	0.938	0.0001
Layoff	0.982	0.00002

Table S7. Model fit (measured by R^2) for linear regression models of squared prediction error. The model with fixed effects for each family fits the data very well, and the model with fixed effects for each account fits the data poorly. These results quantify the visually apparent pattern in Figure 4 that squared prediction error is strongly related to the family being predicted and weakly related to the technique used to the make the prediction.

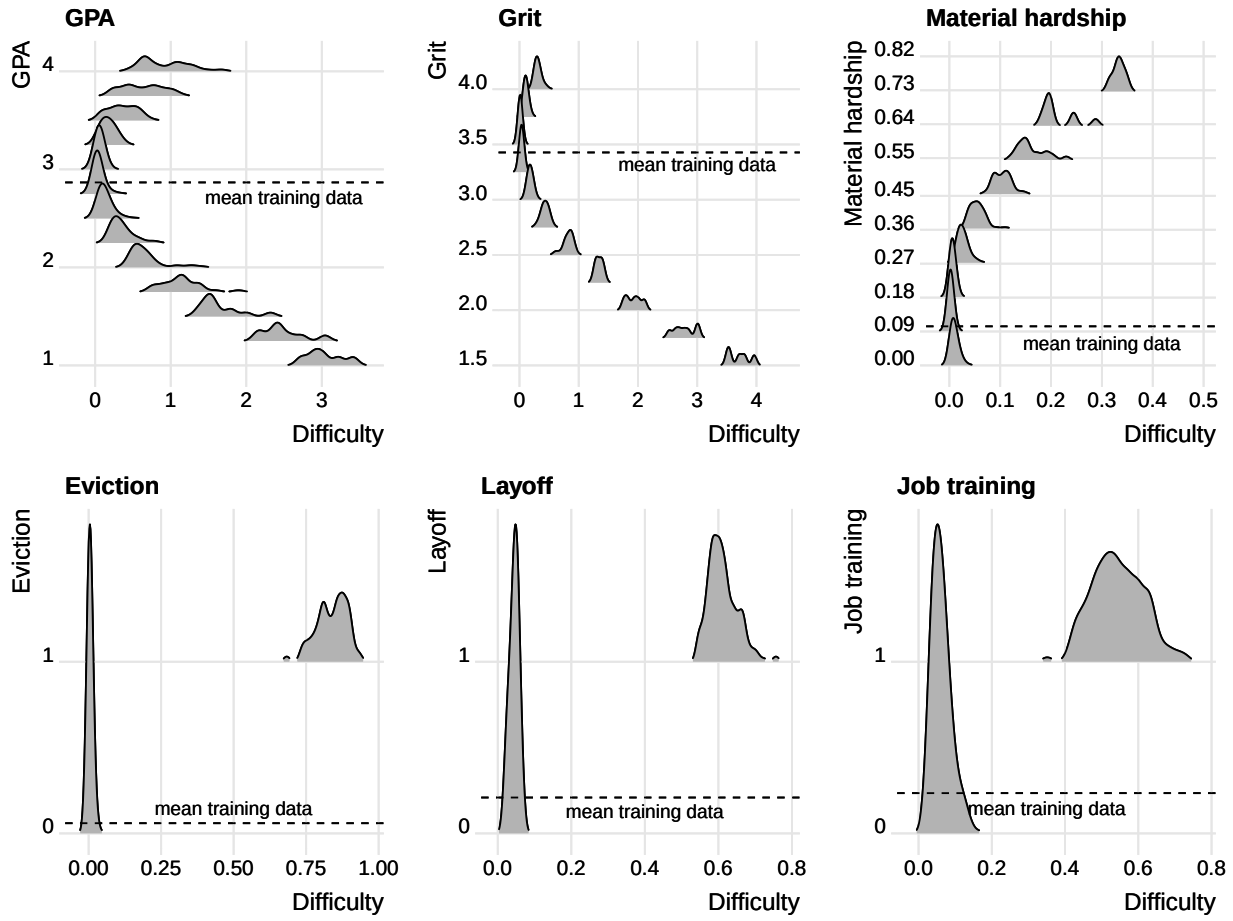


Fig. S13. Observations that are furthest from the mean of the training data are the hardest to predict. Each ridge indicates the smoothed density for the difficulty of predicting cases with a specific outcome value (e.g., GPA of 4.0). Difficulty is defined to be the mean of the squared errors predicting that case averaged over all qualifying submissions (Eq. S7).

observation, y_{ik}, is mean square error for that observation averaged across all qualifying submissions

$$d_{ik} = \frac{\sum_j (\hat{y}_{ijk} - y_{ik})^2}{n_{jk}}, \tag{S7}$$

where n_{jk} is the number of qualifying submissions for outcome k (Table S5). It turns out that this difficulty value (d_{ik}) is larger for observations that are far from the mean of the training data (Fig. S13). While one might expect that unusual observations will be the hardest to predict, this pattern is not a guaranteed to arise. For example, a child with a GPA of 3.0 (close to the mean of the training data) could have been predicted to have a GPA of 1.0 by all submissions, which would have produced a case close to the mean of the training data with large difficulty value (d_{ik}).

A third pattern in the prediction errors is that families with difficult-to-predict values for one outcome

do not, in general, have difficult-to-predict values for the other outcomes. For example, adolescents with a difficult-to-predict value for GPA do not also tend to have a difficult-to-predict value for grit (Fig. S14). But, there are three exceptions. First, families where eviction was difficult to predict also tended to have a material hardship value that was difficult to predict ($r = 0.46$). This pattern may be created, in part, because eviction is one element of the material hardship scale (Table S3). Also, families where the layoff of the primary caregiver was difficult to predict also had difficult-to-predict values for eviction ($r = 0.17$) and material hardship ($r = 0.12$). These patterns may be created, in part, because of the causal relationships between these three variables. For example, it could be that unexpected layoffs increase the chance of unexpected eviction and unexpected material hardship.

In conclusion, for each of the six outcomes: the squared prediction error is strongly related to the family being predicted and weakly related to the technique used to make that prediction (Table S7); the observations that are hardest to predict tend to be far from the mean of the training data (Fig. S13); and families with observations that are difficult to predict for one outcome generally do not have observations that are difficult to predict for other outcomes (Fig. S14).

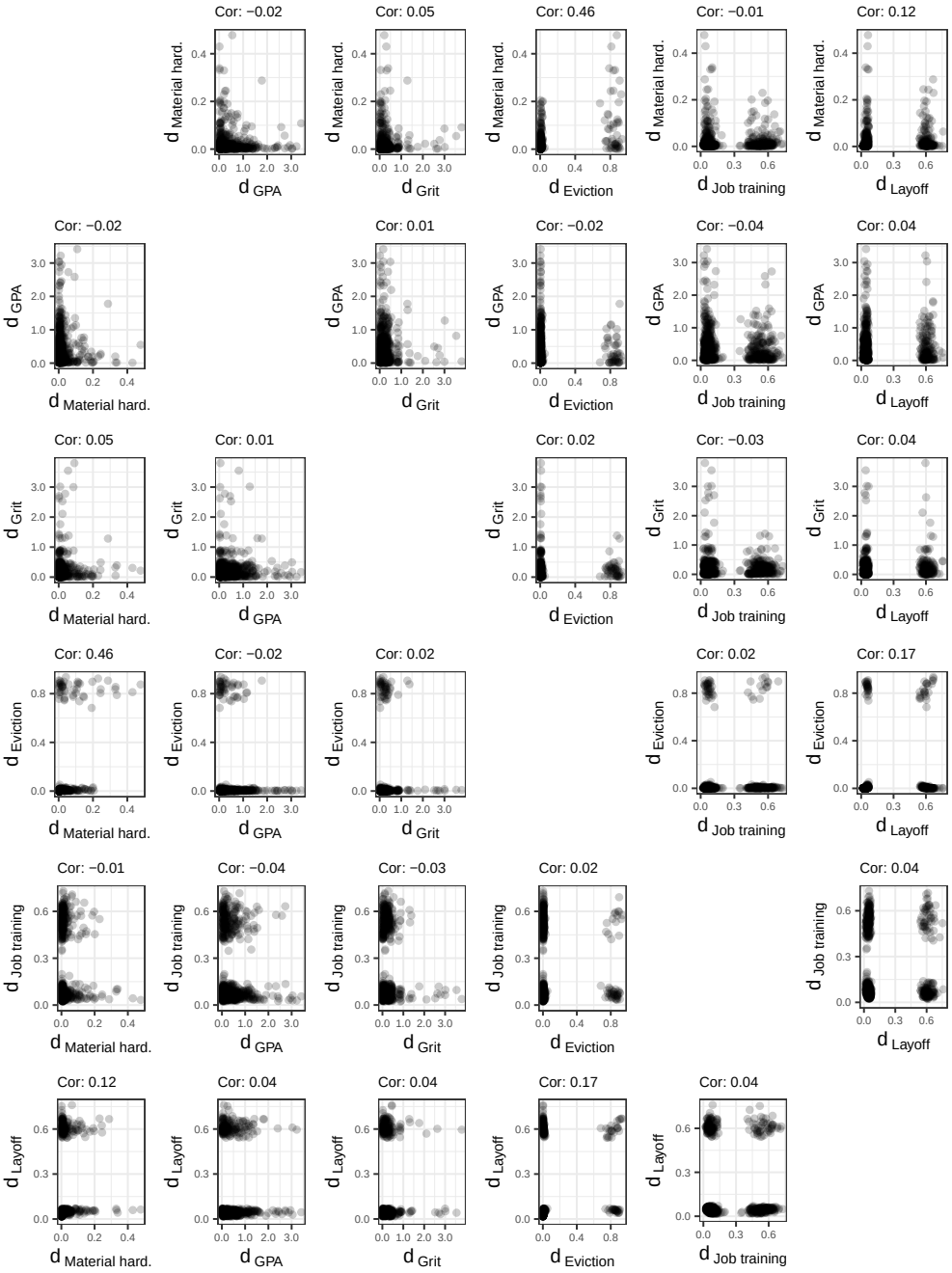


Fig. S14. Families with observations that are difficult to predict for one outcome generally do not have observations that are difficult to predict for other outcomes. Each scatter plot shows the relationship between the difficulty values (Eq. S7) for two outcomes. The difficulty of prediction is not strongly correlated across outcomes with three exceptions: eviction-material hardship ($r = 0.46$), layoff-eviction ($r = 0.17$), and layoff-material hardship ($r = 0.12$). This figure only includes cases with non-missing data for all six outcomes.

S4 Approaches used by teams to generate predictions

S4.1 Prediction approaches as demonstrated in submitted code files

Many of the submissions in the Challenge had similar predictive performance (Fig. S15). This was not because all of the teams used the same approaches to generating predictions. Here we parse the code that accompanied all 160 valid submissions. This included scripts in the following languages: Python (230 scripts), R (145 scripts), Stata (28 scripts), Matlab (9 scripts), and SPSS (2 scripts). In some cases, there was more than one script per submission.

To analyze the code accompanying each submission, we first concatenated all the scripts in the submission into a single file. We then extracted the set of functions used in each file. For submissions in Python, R, Stata, and Matlab, we extracted the function automatically using Pygments[2]; for submissions in SPSS, this step was done manually. Next, we identified the full set of functions that were used in the Challenge across all languages. We then manually categorized each function into one of three mutually exclusive and exhaustive categories: data preparation (e.g., subsetting the data or imputing missing data); statistical learning; or interpretation (e.g., visualization and extracting predictions). Further, we subcategorize statistical learning functions into 11 mutually exclusive and exhaustive categories (Table S8). At the end of this procedure, for each submission, we have a count of the number of times each function was used. However, we do not know which of these function calls were associated with each particular outcome (i.e., we know how many times a particular submission used linear regression, but we cannot associate these function calls with a particular outcome).

[2] http://pygments.org/

Statistical learning label	Meaning	Example functions
tree	Partitioning and tree-based methods	`ctree, rpart`
regular	Regularized models	`bayesglm, spikeslab`
nn	Neural networks	`nnet, sklearn.neural_network`
logistic	Binary response models	`logit, probit`
linear	Linear models	`lm, reg`
kernel	Kernel methods	`sklearn.gaussian_process`
flex	Flexible functional form models	`gam, svm`
factor	Latent variable models and factor analysis	`chol, svd`
ensemb	Ensemble models	`xgboost, randomforest`
cv	Cross-validated models	`sklearn.cross_validation`
classif	Classification and clustering models	`sklearn.naive_bayes, knn`

Table S8. Classification scheme for statistical learning procedures. A single submission could include more than one statistical learning procedure.

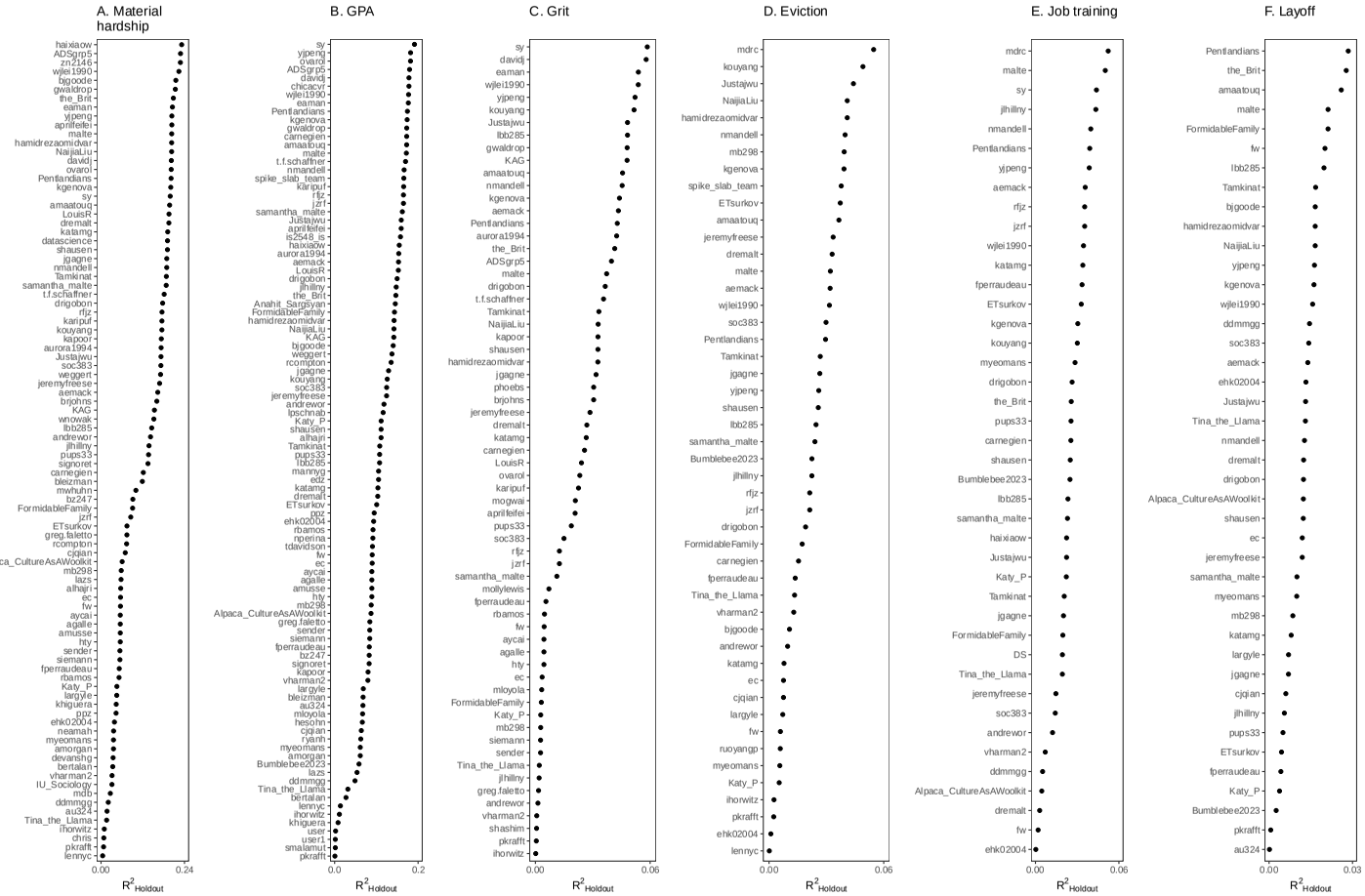


Fig. S15. The R^2_{Holdout} of many submissions are similar. The best model is not enormously better than the next-best model. Figure shows all qualifying submissions.

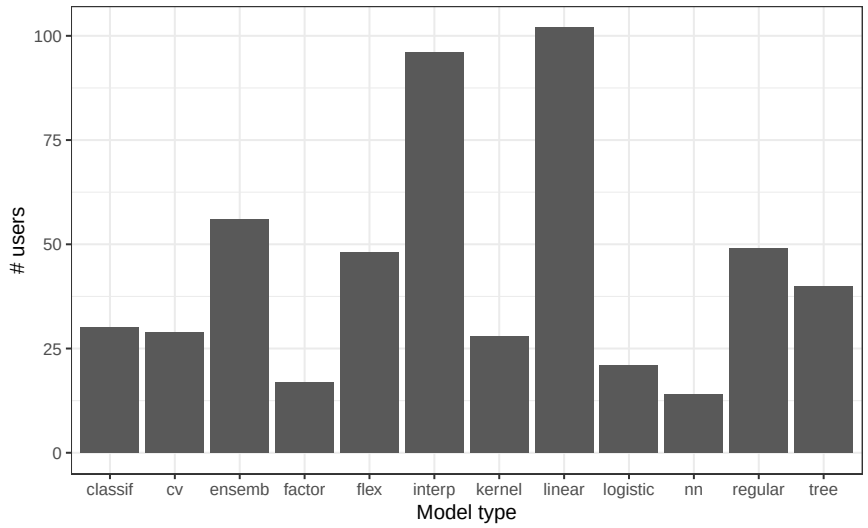


Fig. S16. Statistical learning procedures used in submissions. Participants constructed predictions using a wide range of functions. The classification scheme is described in Table S8.

Overall, the analysis of the submitted code produced three main results. First, submissions included a wide variety of statistical learning procedures (Fig. S16). Second, if we restrict our attention to qualifying submissions (those that performed better than predicting the mean of the training data), then there is not a strong relationship between the amount of data preparation and R^2_{Holdout} (Fig. S17). Third, again restricting attention to qualifying submissions, there is not a strong relationship between the statistical learning functions used in the submission and R^2_{Holdout} (Fig. S18).

We emphasize that our automated analysis of submission code potentially misses important pieces of each submission, such as the order that the functions were executed, user-created functions, and specific information about how each statistical learning approach was applied (e.g., procedure for estimating hyperparameters). Also, some submissions might include code that was not actually executed to generate predictions, or might exclude code that was executed (e.g., if it was commented out subsequent to the execution that generated the submitted predictions). Finally, these results do not provide causal evidence that one procedure leads to better outcomes. Submissions using a particular learning method might have differed along other dimensions that affect predictive performance, such as the quality of the data preparation. Despite these important caveats, the automated analysis shows that a wide range of statistical learning approaches were used in the Challenge. Further, this analysis provides suggestive evidence that the predictive performance was not strongly related to the techniques used to produce each submission.

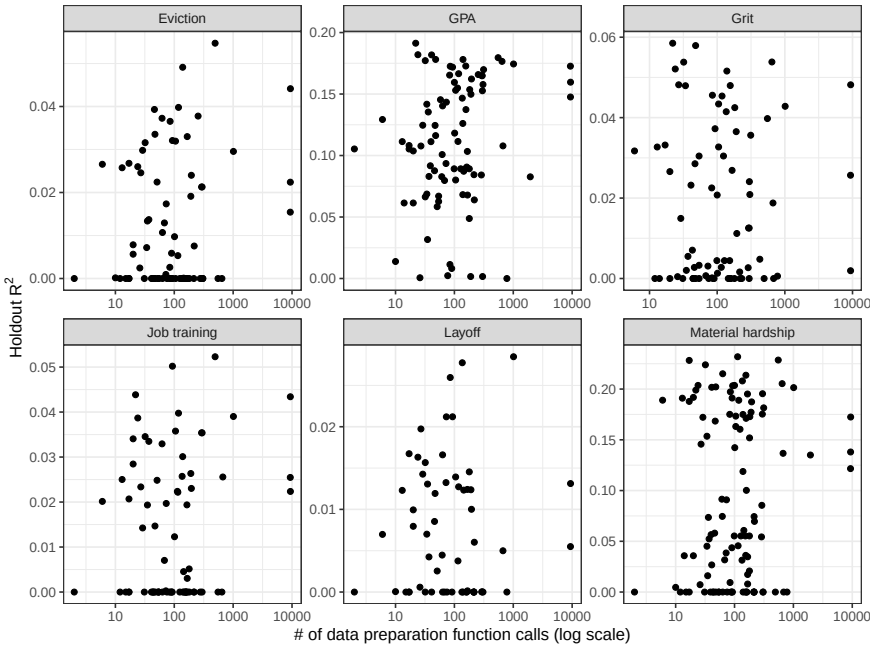


Fig. S17. Performance of submissions by amount of data preparation. There is not a strong relationship between the log number of data preparation function calls and R^2_{Holdout}, for submissions with $R^2_{\text{Holdout}} > 0$.

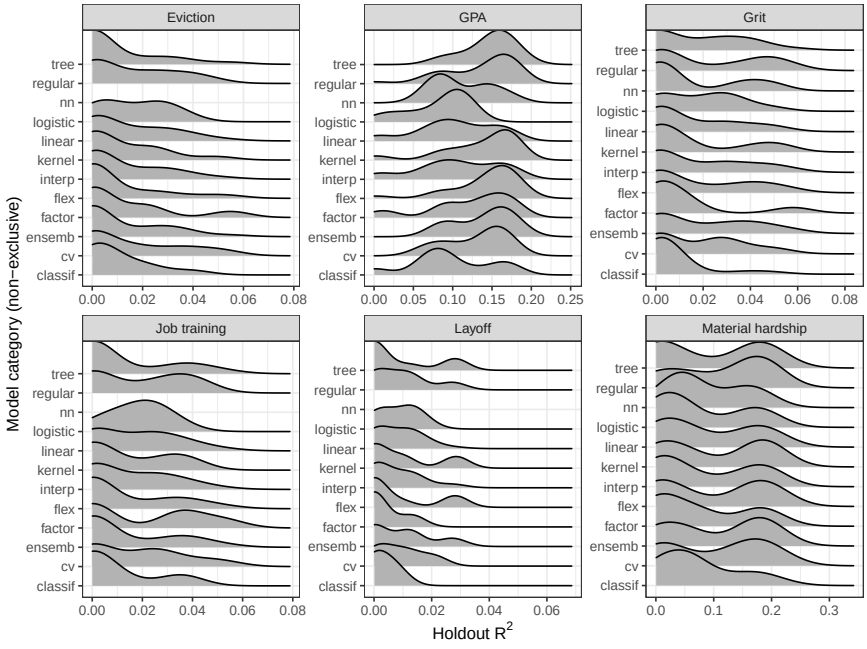


Fig. S18. Performance of submissions by type of statistical learning procedure. Each ridge indicates the smoothed density of R^2_{Holdout} among submissions that used procedures of the corresponding type and had $R^2_{\text{Holdout}} > 0$. A missing ridge indicates that a statistical learning procedure of that type was not used in any submission where $R^2_{\text{Holdout}} > 0$ for that outcome. X-axis scales vary by facet.

S4.2 Prediction approaches as reported in categorical survey responses

To give a sense for the range of methodological approaches used in the Challenge, participants authoring this paper provided summaries of their methodological choices in the form of yes/no survey responses and brief narratives. This section presents the survey responses as four figures corresponding to the following steps: data preparation (Fig. S19), variable selection (Fig. S20), learning algorithm (Fig. S21), and model interpretation (Fig. S22).

Fig. S19. Participant reports of data preparation steps.

Fig. S20. Participant reports of feature selection steps.

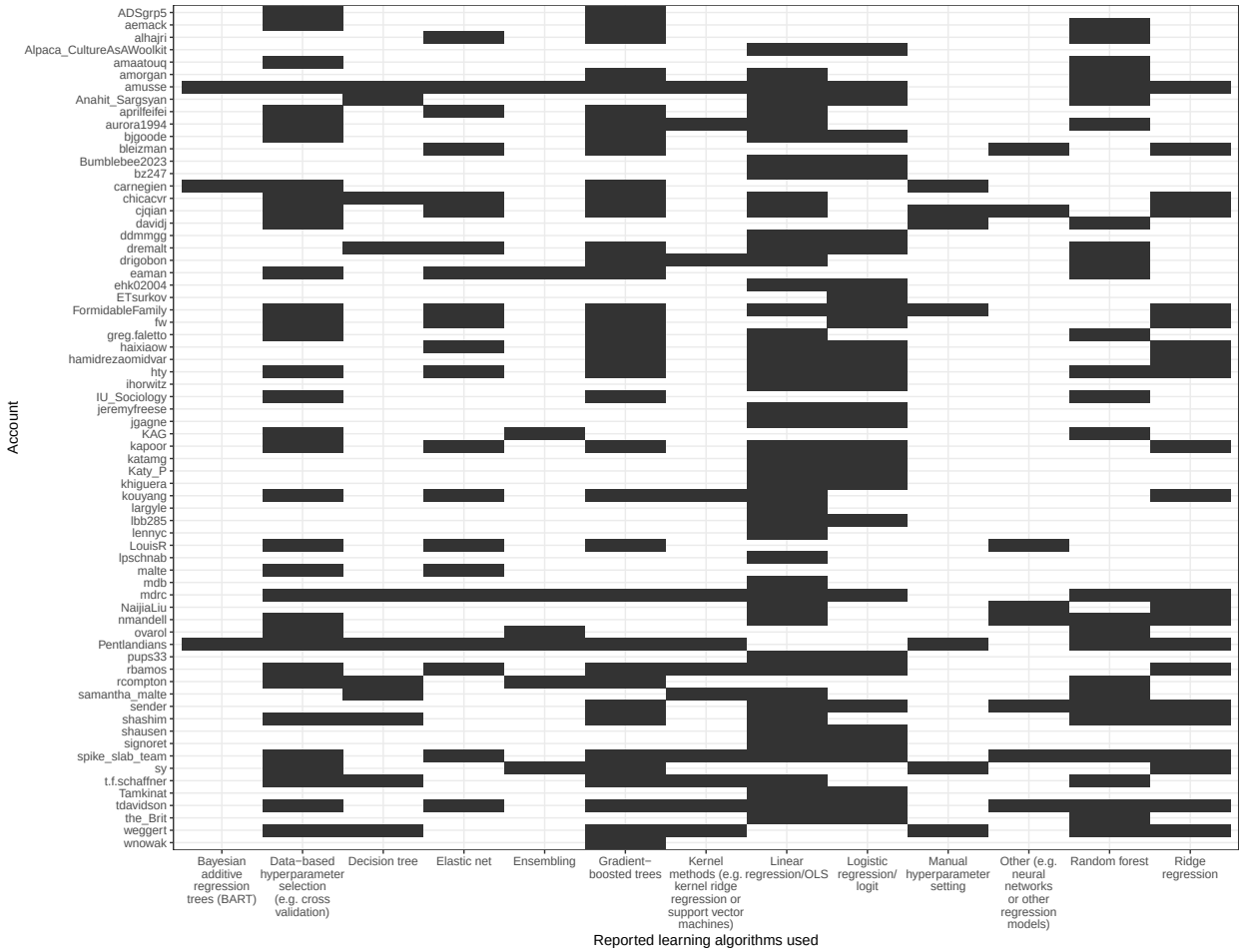


Fig. S21. Participant reports of learning algorithms used.

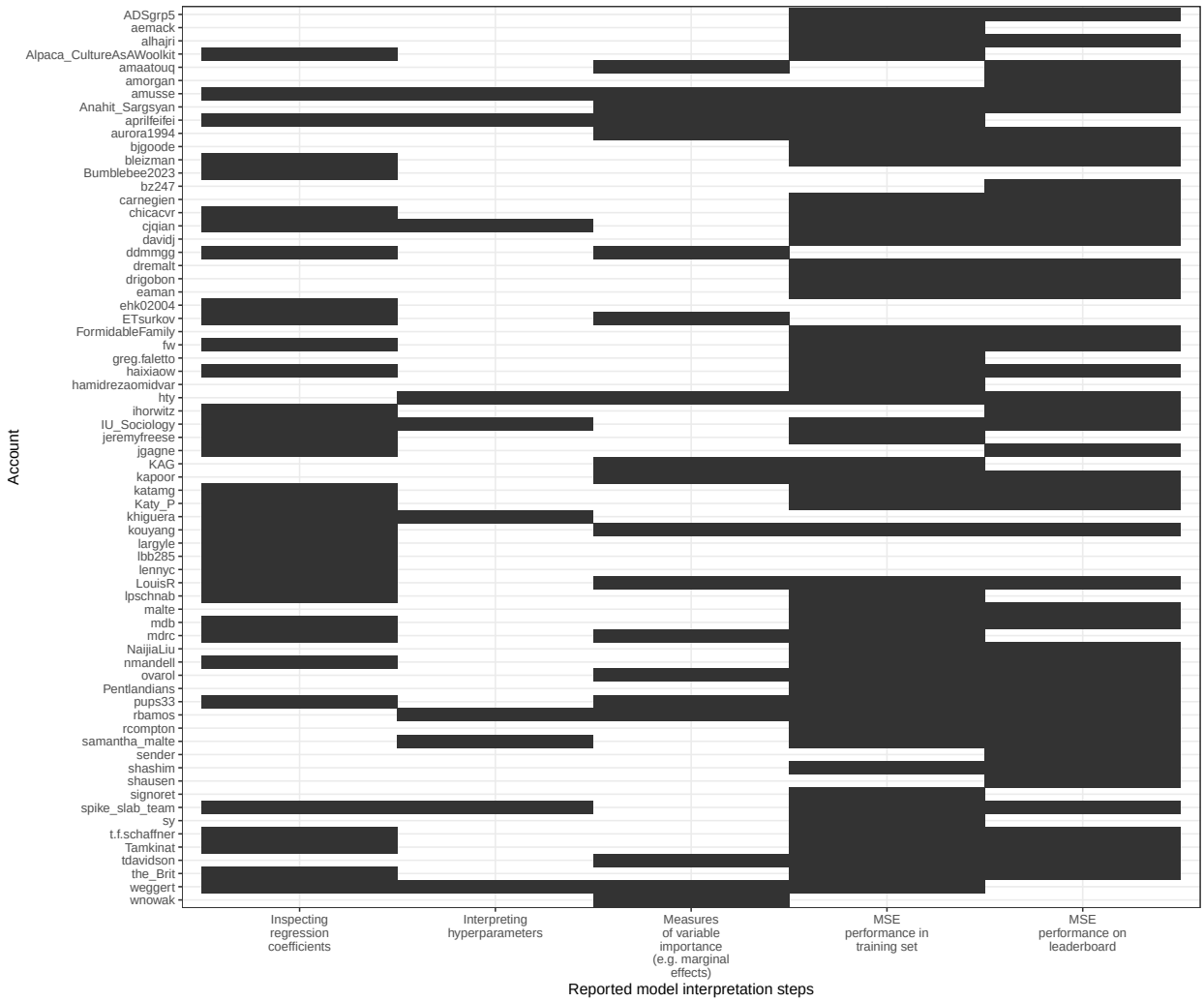


Fig. S22. Participant reports of model interpretation steps used.

S4.3 Prediction approaches as described by narrative summaries

This section presents narrative summaries of the approaches teams used to produce submissions to the Challenge, provided by teams for whom a participant is an author of this paper. We group together authors who indicated that they are describing the same account submission. We provide account usernames next to each narrative summary.

- **Caitlin E. Ahearn and Jennie E. Brand (pups33)**

  Our approach to the Fragile Families Challenge was to draw on empirical results from the large literature on the determinants of job loss to generate a simple model to predict caregiver layoff. We chose to focus exclusively on layoff as the outcome because Jennie Brand, a member of our team, is an expert in that area. We used prior research on layoff to build a model predicting layoff, and adjusted the model based on model fit statistics of different iterations of predictors. Our models included subsets of sociodemographic, employment, family, psychosocial, and family background characteristics. We ultimately submitted a few model specifications to the Challenge: the first was based on prior research on job loss for the general population; the second was based on prior research on job loss among disadvantaged mothers (the primary population of the Fragile Families data); and the third was a more parsimonious model that included a few key covariates. We also tested various methods of imputation, including single and multiple imputation. The results and predictions of each of these models were all quite similar. Our best performing model, and final submission, was one of our early model specifications based on covariates for predicting layoff in the general population. We chose a simple estimations strategy, and did not expect different model specifications to produce much variation, as prior research has repeatedly shown that job loss is a relatively exogenous shock.

- **Khaled Al-Ghoneim (KAG)**

  Using the same feature selection as the team (Pentlandians), I combined multiple random forests using weighted averaging. The weights are the out-of-bag performance score for each forest.

- **Abdullah Almaatouq (amaatouq)**

  In this submission, we ran a Random Forest regressor for continuous outcomes and Random Forest classifier for the categorical outcomes. In the case of categorical outcomes, we predict the probability of positive examples rather than the binary class. All of the models in this submission were trained on the top 600 untransformed features selected by mutual information that was performed during the pre-processing step. Due to the high potential of overfitting, we ran 300 Random Forests in a nested cross-validation fashion. This means, we used a series of train/validation/test set splits, where in the

inner loop, the score is maximized by fitting a model to each training set, and then maximized in selecting hyper-parameters over the validation set. In the outer loop, generalization error is estimated by averaging test set scores over several dataset splits. We then weighted each model predictions based on this outer loop score.

- **Drew M. Altschul (dremalt)**

  I first removed many variables with low-variance, so that I could then select particular features of interest with generalized boosting. Once I had a separate set of features for each outcome variable, I made a single data subset and multiply imputed missing values. Prior to imputation I added some variables to the dataset either because they were of special theoretical interest, or because I wanted to use them to add power to the imputations. With the imputed datasets I fitted elastic net and more standard linear regression models to the dataset, and using these models I generated the predictions that I submitted to the challenge.

- **Nicole Bohme Carnegie and James Wu (carnegien)**

  The first step in our process was data cleaning. Coming from a sociology-influenced statistics background, we felt that it was important to account for skips, etc. from the survey instruments in a logical manner. We spent a great deal of time cleaning and recoding data to reflect implied answers from the survey structure, and force categorical responses to be treated as such. We also dropped "administrative" variables, like time of survey and sample weights, and any variables that were constant across traning observations. Once this was done, we used one of four methods to reduce the number of variables used in predictive modeling: LASSO, Bayesian GLM, Horseshoe, and Bayesian Additive Regression Trees (BART). All predictive models were fit using BART. We fit many combinations of variable selection methods and hyperparameter settings for BART, in order to explore which of these would be related to final predictive performance.

- **Ryan James Compton (rcompton)**

  Our method involved cleaning, balancing, and then splitting the data set to ensure a more generalizable model. Due to the high number of variables within the data set, we used Principal Component Analysis as a feature engineering method to reduce both the number of variables and redundant information. After conducting a parameter search for how many components would be best for each dependent variable, modeling was conducted through Cross Validation and Random Forests. The best model found (through MSE performance) would then be used to make predictions for the Challenge test set.

- **Debanjan Datta and Brian J. Goode (bjgoode)**

Our approach to the the Challenge was primarily focused on survey structure and variable construction. In the former, we recognized when either dropouts would occur or the information would not be available due to a previous answer. These data were then imputed using the nearest data point in terms of year, relation, or question type. In the latter strategy, variables were constructed by counting scale responses forming the Material Hardship, Grit, and GPA outcomes.

- **Thomas Davidson (tdavidson)**

  I initially started by experimenting with all six outcomes and a range of models but decided to focus on GPA. I used simple heuristics to deal with missingness and identify variable types. I then standardized continuous variables and one-hot transformed categorical variables. To make predictions I used feed-forward neural networks, varying the depth, breadth, and activation function used.

- **Anna Filippova, Connor Gilroy, Ridhi Kashyap, Antje Kirchner, Allison C. Morgan, Kivan Polimis, and Adaner Usmani (FormidableFamily)**

  Recent applications in computer science have sought to incorporate human knowledge into machine learning methods to address overfitting during prediction tasks, where data sets are incomplete, and have a high ratio of variables to observations. To address these issues, we implement a "human-in-the-loop" approach in the Fragile Families Challenge. First, we try several different approaches for imputing missing responses: mean imputation, regression based approaches, and multiple imputation. Next, we use surveys to elicit knowledge from experts and laypeople about the importance of different variables to different outcomes. This strategy gives us the option to subset the data before prediction or to incorporate human knowledge as scores in prediction models, or both together. We incorporate this variable information and imputed data into regularized regression models. What we find is that human intervention is not obviously helpful. Human-informed subsetting reduces predictive performance, and considered alone, approaches incorporating scores perform marginally worse than approaches which do not. However, incorporating human knowledge may still improve predictive performance, and future research should consider new ways of doing so.

- **Eaman Jahani (eaman)** I worked as an individual member of a group. The group generated a single pipeline for data cleaning, imputation, and variable transformation which we all used for our own independent statistical learning step. The group prediction was an ensemble of our individual model predictions. Our data pipeline first determined which variables are categorical and which are continuous based on the number of unique values they take. Then it converted all categorical variables to dummies, including missingness dummies. It also dropped variables with low variance or high rate of missingness. The final cleaned data had more than 20,000 features. So our individual models

had to do aggressive feature selection. To reduce the number of features prior to model learning, I performed univariate feature selection and reduced the number of features to the top most predictive 100, 300, 1000 and 1500 features. Prior to model building and feature selection, I also added various transformations of the continuous variables, e.g. log or square or square-root, to the data matrix. I only attempted to predict the continuous variables. My submission used a multi task elastic net which predicted all dependent variables, GPA, grit and material hardship, at the same time. The multi-task elastic net could be more efficient in case there is significant correlation between the dependent variables. The grit prediction came from this multi task model. For GPA, I took the average of two models. The first model was an elastic net (trained only on GPA) using the top 1500 features selected first through univariate feature selection. The second model was a random forest regressor on top 1000 features. In both models, features were normalized. Elastic net was also trained on gpa-squared since this transformation of the dependent variable gave better results. For material hardship, I used an elastic net on the top 300 features.

- **Stephen McKay (the_Brit)**

  Subject specific knowledge was used to identify a long list of those variables most likely to have associations with the outcomes of interest, including some of the scales available in the survey and values of the outcomes in earlier waves. Statistical measures (R-squared, MSE, regression coefficients) were then used, alongside subject expertise, to produce final models from among that list. Continuous outcomes were modeled using relatively small random forests, and binary outcomes using logistic regression.

- **Allison C. Morgan (amorgan)**

  We chose only "constructed" variables – those derived from the originally collected data by domain experts – to train our models on. These observations were more or less complete, meaning the issue of handling missing data was less relevant for us here. In processing our data, we maintained discrete categorical variables and turned continuous variables into discrete variables by binning them into quartiles. This binning was done for later use with a different ML approach that required categorical variables. In retrospect, it would have been wise to allow our models to learn the appropriate divisions for a continuous variable. We predicted all outcome variables using a combination of regularized linear regression (for continuous outcomes) and a random forest (binary outcomes). Results were evaluated based on our ranking on the leadership board at the time of submission.

- **Alex Pentland (Pentlandians)**

  The Pentlandians submission is an ensemble prediction, where it aggregated four individual sets of

predictions (i.e., one from elastic net, two by random forest, and another from a GBoost tree). In particular, the ensemble prediction consists of a weighted team average, in which the weights were determined by relative ranking on the leaderboard (i.e., the weight vector for the top three performing predictions for each outcome was given by (1/2, 1/3, 1/6) for first, second, and third, respectively). Predictions performing worse than 30th on the leaderboard were not included in this averaging.

- **Louis Raes (LouisR)**

  My approach in making predictions was based on a cursory reading of literature on the Fragile Families and Child Wellbeing study, combined with a lot of trial and error.

- **Daniel E. Rigobon (drigobon)**

  Pre-processing of the data was done by removing features with low-variance, performing one-hot-encoding on all categorical features, and mean-imputing all missing continuous and ordinal features. Due to the large amount of covariates produced by this process, a sparse linear regression was run for each outcome to identify important features. A regularization parameter was selected to ensure that the regression's R^2 value was close to an ad-hoc value of 0.4. Following feature selection, various learning algorithms were evaluated on splits of the training data: Principal Components Regression, Kernel Support Vector Machine, and Random Forest. The Random Forest algorithm consistently had the best performance for all outcomes. Its' hyperparameters were selected by cross-validation, and final predictions were made with the full training set.

- **Claudia V. Roberts (chicacvr)**

  We divided the project into two steps. In step 1, we used a completely automatic approach that does not consider the data (the norm in data mining) to fit 124 models for GPA prediction. In step 2, we attempt to improve upon our results. We use a strategy that combines engineering-centric statistical analysis techniques with classical, more manual social science methodologies: we examined each variable in the codebook, manually selecting the ones believed to be predictive of academic achievement based on a non-expert reading of domain-specific research. Results indicate that it in most cases, it pays off for engineers to "make friends" with the FFCWS codebooks. We were able to improve the predictive accuracy of 6 of the 10 top step 1 models, of which 4 saw significant improvements. However, manual variable selection did not improve the predictive ability of the 2 most accurate models from step 1. We tried many different approaches to data pre-processing. We tried almost all combinations of 4 different decisions: 2 types of automatic variable selection (F-test and mutual information) using 2 thresholds (10% and 20%), 2 types of imputation strategies (median and mode), and 2 standardization approaches (no standardization and standardization).

- **Yoshihiko Suhara (sy)**

  My approach was training Gradient-boosted Tree models on imputed features without feature selection, with intensive hyper-parameter search based on Grid Search. The hyperparameter candidates were manually crafted based on my experience in data science competitions. The approach was fully data-driven; I made the best use of computation resource for the hyper-parameter search to reduce the risk of overfitting, and I chose the predictive model that performed best in the cross-validation evaluation.

- **Erik H. Wang and Diana M. Stanescu (haixiaow)**

  Our approach consists of the following steps. First, we do early house-cleaning by dropping variables with more than 60 percent missing and dropping variables with standard deviation smaller than .01. Second, we mean-impute the data and perform LASSO regressions of the outcome variables on all remaining covariates. We then drop any covariate whose coefficient is zero. Third, for the remaining covariates, we identify their originals (i.e., before mean imputation), and apply multiple-imputation using Amelia (which employs EM algorithms). We apply LASSO again for variable-selection using the Amelia-imputed dataset. When applying Amelia, we set M = 5 and pick the third dataset.

- **Muna Adem, Andrew Halpern-Manners, Patrick Kaminski, Helge Marahrens, Landon Schnabel, and Zhi Wang (IU_Sociology)**

  Our approach rests on a combination of social science theory and machine learning methods. We first developed a theoretically-informed list of variables we expected to be important. We then augmented this list with highly predictive variables selected by a LASSO regression. All variables in the augmented list were verified using domain knowledge. Finally, using the complete list, we trained a random forest regressor / classifier, and tuned its hyperparameters with cross-validation.

- **Abdulla Alhajri (alhajri)**

  Model performance for the leaderboard and holdout sets was determined by looking at the improvement over the baseline - or relative accuracy improvement.

- **Anahit Sargsyan, Areg Karapetyan, Bedoor AlShebli, and Wei Lee Woon (Anahit_Sargsyan)**

  The employed approach resorts to machine learning techniques for devising a predictive model for GPA with a particular focus on explicability of the results produced when considering the nuanced variations between subjects. To facilitate the analysis of the data, a number of pre-processing steps were carried out: (i) all missing and negative values were replaced by NaN and the columns with 0 variance were removed, (ii) columns with more than 400 NaN values were dropped, (iii) the variant of kNN (k-Nearest Neighbors) imputation algorithm was leveraged to estimate the NaN values. Next, a manifold of filter-

and wrapper-based methods, including Principal Component Analysis, Ridge regression, Lasso, and Gradient Boosting Regression, were attempted in search of the most informative feature subset of reasonable cardinality. These methods were applied to the extracted pool of features, both recursively and explicitly, and probed under diverse parameter settings. The acquired subsets were then evaluated for their predictive accuracy across various models trained. The target subset of optimally descriptive features, as revealed by extensive experiments, was obtained by the following means. Feature importances were estimated by the Extra Trees Regressor algorithm and Randomized Lasso, and the top 500 features were retained from each. For the latter, two different values were considered for the regularization parameter, thus resulting in two separate feature subsets. The intersection of these three subsets, containing 69 features, led to maximized GPA prediction accuracy. More concretely, with the Random Forest algorithm, a mean squared error of approximately 0.363 was achieved, allowing these results to be placed in the top quartile of the final FFC scoreboard.

- **Redwane Amin (spike_slab_team)**

  Before using machine learning techniques to predict the outcomes, preparing the data and selecting features are crucial steps. In addition to statistical techniques, investing time in analyzing the study documentation allowed to filter out variables which would have otherwise added noise or led to overfitting. Then for each chosen statistical learning method, we tuned hyper-parameters (where applicable) on the training set using cross validation and evaluated their performance on the held-out validation set.

- **Ryan B Amos and Guanhua He (rbamos)**

  Our most successful insight was the use of feature selection. We tried a variety of feature selection techniques, and found k-means to be the most effective technique. We found using 50 clusters provided the best results, which means most of the data can be well represented by just 50 variables. The most effective machine learning techniques on the clustered data were elastic net regularization for continuous outcomes and a support vector machine trained with stochastic gradient descent for discrete outcomes. We tried a variety of imputation techniques, but ultimately we found that the naive method of imputing the data to the mode was just as effective as more targeted imputation methods.

- **Lisa Argyle (largyle)**

  I used subject area expertise and hypothesized that income at birth and household income growth over time would be correlated with the outcome variables, especially layoffs and material hardship. I conducted basic data cleaning of the predictor variables (child gender, and income/poverty at birth), and generated a new variable indicating the change in household income from wave 1 to wave 5. I then

used a linear OLS model to predict the six outcome variables. I imputed the mean value for any case dropped due to missing data.

- **Livia Baer-Bositis (lbb285)**

  The models to predict each of the six outcome variables were built around the concept that the past predicts the future. The key explanatory variables were all constructed from data collected in year 9 of the study including a scaled measure of hardship and selected via trial and error.

- **Moritz Büchi (mdb)**

  The first step was to obtain a complete data set using multiple imputation by chained equations. The variable selected as the outcome was material hardship. The approach in this submission was to show that a simple linear model may produce smaller mean squared errors than benchmark models even when the selected predictors are theoretically uninterpretable, pointing to the often opposing analytical goals of prediction versus explanation.

- **Bo-Ryehn Chung and Flora Wang (fw)**

  Our methods and models were parsimonious in complexity, but still managed to perform higher than average in at least one outcome. We first removed variables with low variance (mainly those of string types) to reduce the dataset dimensionality. We then performed median and mode-based imputation on variables with missing or certain no response codes, as most outcomes were of skewed distributions. We then evaluated various regularized regression methods that selected important features, and applied median importance weights to these features with cross-validation for the final model. The elastic net model with cross validation performed the best based on metrics as MSE and confidence intervals of the cross validation scores. We made sure to assess the features selected and their coefficients using literature reviews and the study documentation. Other regularized regression methods either overfit the data (ridge regression) or did not select enough variables that made intuitive sense (due to the random nature of LASSO feature selection for correlated variables). We found that the elastic net model had a good balance of finding correlated variables abundant in our longitudinal dataset and selecting enough features for the final model.

- **William Eggert (weggert)**

  It was interesting to see that classifying Eviction on the validation set yielded an accuracy of greater than 90% out of the box; in fact PCA and K-Best feature selection often made the accuracy worse. Additionally, this challenge was very susceptible to choosing hyperparameters that produced excellent accuracy, but trivial results (e.g. predicting GPA but only producing middle-of-the-road GPAs).

Therefore, careful feature space reduction is of utmost importance. Voting classifiers were the most robust against this pitfall. However, a Gaussian Mixture Model Process served to be the most promising avenue for worthwhile results. A combination of domain-expert collaboration to reduce the feature space, with a GMM, is predicted to produce the best results.

- **Gregory Faletto (greg.faletto)**

  I only trained models for the continuous responses. I relied on the constructed variables. I trained a lasso model and a principal components model on each continuous response. For the lasso model, I chose the tuning/penalty parameter by cross-validation. For the principal components regression, I chose the number of principal components to include by cross-validation. Finally, I chose which of these models to use by comparing the mean squared error of each model.

- **Zhilin Fan (ADSgrp5)**

  Given all the background data from birth to year 9 and some training data from year 15, we infer six key outcomes (GPA, grit, material hardship, eviction, layoff, job training) in the year 15 test data. In the data cleaning process, we deal with categorical variable and continuous variables separately, for continuous variable, we replace the NA with the median value of that variable, and create a new categorical variable to indicate the NAs (where the NAs may contain information to some degree), attach the new indicating categorical variable to the original categorical features. For categorical features, we replace the NAs with a number that doesn't exist in original data set and transform every categorical to a dummy matrix, for every dummy matrix whose elements are either 0 or 1, we choose the 2nd to last column to avoid col-linearity. Given the cleaned data, our team work on different directions, one team work on different features and one team work on different machine learning tools, since we got to know the xgboost apparently outperforms other methods, we together work on features selected from various angles. For case 1: data obtained when children are at age 9 and only consider the continuous variables. Case 2: data obtained when children are at age 9, use categorical variables. Then we bag them by using the weighted average. We use the same strategy to other continuous outcomes (grit, material hardship).

- **Jeremy Freese (jeremyfreese)**

  I did two things. More seriously but less time-consumingly, I just fit models that seemed to make some intuitive sense to me as models. I did not do anything fancy here, nor did I have any illusions that these would rise to the top, but I was using the challenge for pedagogical purposes. Also, as a lark, I played around with generating a bunch of prediction sets with small differences and seeing if I could infer the values of outcomes in the quiz set. This worked for the rare outcome. I thought this

might provide some great advantage–basically it would allow me to generate predictions using both the testing + quiz set–but after the challenge I was informed by MS that there was something that they had done (I forget exactly what) that thwarted this strategy.

- **Josh Gagné (jgagne)**

  GLM with mean imputation and predictions shrunk toward the training set mean.

- **Sonia P. Hashim and Viola Mocz (shashim)**

  We worked as a team to predict gpa, grit, and materialHardship. We experimented with multiple models, using ordinary linear regression, lasso regression and ridge regression after conducting imputation, feature engineering, and feature selection on the expanded feature set. To impute missing values, we tested median single imputation, K-Nearest Neighbors single imputation, and multiple imputation using the Amelia package in R. Also, if there were columns with one value in addition to NA we converted these into binary variables where 0 indicated missing data and 1 indicated present data. We engineered features by taking the mean of matching inputs and by using maximum pooling in order to combine features using similar questions asked across years. We also conducted feature selection by removing features with a low frequency of observations and low variance, features that were highly correlated with each other, and features with low random forest importance. Five-fold cross-validation was used to evaluate the efficacy of our models on the available training data. Our final submission used ordinary linear regression with median imputation, missing data indicators, engineered features, and feature selection using variance.

- **Sonia Hausen (shausen)**

  I cleaned and recoded the data, experimenting with all 6 outcomes using OLS and logit. Coming from a sociology background and studying well-being, I looked for an all-encompassing variable, like overall life satisfaction, which would capture many of the other variables within it (like abuse, income, job status, mental health etc.). I hypothesized that the variable "mother's overall life satisfaction at year 9, self-reported" might be a good predictor given the strong influence mothers have on child outcomes. I used MSE performance on the leaderboard as a guide; my best performing models included the variable mom_sat.

- **Kimberly Higuera (khiguera)**

  Initially, I started by focusing on how long seeded childhood characteristics could predict long term outcomes by using low birth weight as the main dependent variable. I had attended a talk on low birthweight and it's links to test scores when I was undergraduate and I wanted to investigate whether

the link was robust with the fragile family data. When the link based on the coefficients and t scores did not seem statistically or socially significant across different model types and different covariates and outcomes, I decided to shift into looking at factors that were arguably even more deeply seeded than birthweight: mother's characteristics. I considered these more deeply seeded because they existed before the birth and also because they had long term interaction and thus potential long term effects on the respondents. Plus mothers seem much more impactful in that they are seem less likely to be missing from the raising of a child than a father. Following that I ended up looking at how mother's characteristics affected the likelihood of getting laid off.

- **Ilana M. Horwitz (ihorwitz)**

  I looked at all 6 outcomes. Based on my prior knowledge, I chose the following explanatory variables: whether the mother received welfare, the length of time the mother looked for a job, whether the home had peeling paint, frequency of drinking alcohol, the father's influence on school, a child's sense of belonging in school, and whether the child saw his father in the last year. In some cases, I converted the variable into a binary outcome. The explanatory variable I used varied based on the outcome of interest. I then ran logits and OLS models to predict outcomes.

- **Lisa M. Hummel (Bumblebee2023)**

  I relied on knowledge from sociology and psychology about what factors impact outcomes for children and families and attempted to capture those in the models to predict the results.

- **Naman Jain and Ahmed Musse (amusse)**

  In this work we build machine learning models to predict the 6 key outcomes in the children's development: GPA, grit, material hardship, eviction, layoffs, and job training. We first imputed the missing values from the survey by replacing the missing values with the mode for that category. Then, to predict the 3 binary outcomes, we performed chi-squared feature selection to get the 1000 best features. We use various binary classifiers such as Logistic Regression, K-nearest neighbors, Random Forest and other ensemble methods to predict eviction, layoffs and job training. For these outcomes, we found a tuned Random Forest Classifier to perform best given its ensemble nature and enhanced ability to restrict over-fitting. To predict the 3 continuous outcomes, we first did chi-squared feature selection to get the 1000 best features. To predict GPA, grit and material hardship, we conduct Support Vector, Lasso, Ridge and Gaussian Process regressions. Here, the simplicity (and run-time efficiency) of Lasso regression and its in-built ability to conduct model selection made it the preferred method. Our results corroborate past research showing that children in stable two parent households fare better but establish correlations with more novel features as well. We see that housing conditions in early years

especially in years 1, 3 and 5 after birth are particularly predictive for all 6 outcomes. Therefore, an appropriate policy response might be to focus efforts to promote better access to housing for families with young children.

- **Kun Jin and Xiafei Wang (aprilfeifei)**

  We worked as a group. During the data pre-processing, we deleted variables with 70% missing values and imputed the missing values of the rest variables with the mean value. For our best model, we conducted lasso regression with L1 regularization upon 4574 variables for all six outcomes to select features. To be specific, we first obtained the coefficients of the regression model using 5-fold cross-validation and the elastic method with $\alpha = 0.5$; further determined the largest regularization coefficients such that the mean saqured error (MSE) is within one standard error of the minimum MSE; finally, corresponding features with larger coefficients are selected and used to train the regression model. Feature normalization by scaling was followed by feature selection. Prior to our final model, we also fit linear regression, SVM and Linear Discriminant Analysis model, but neither of them yielded better results than lasso regression.

- **David Jurgens (davidj)**

  My approach used a Random Forest (RF) classifier for categorical attributes and RF regressor for numeric attributes. All categorical features were one-hot encoded. During the initial design phase, I tested using PCA and SVD transformations of the features, which worsened performance and were left out of the final model. Similarly, I also examined using mean-value imputation of missing data, which also led to worse performance with the RF models. Both design choices were tested using 5-fold cross-validation within the training data. Since the categorical tasks had imbalanced numbers of instances for each label, I used SMOTE to oversample rare classes using synthetic instances until all classes had an equal number of instances. This oversampling lowered the error in all of my cross-validation tests on the training data. Finally, I optimized the hyperparameters of the model using a sweep across multiple values; ultimately, the only two hyperparameters that substantially affected performance was the minimum number of instances per leaf in the decision tree and the total number of decision trees. Ultimately, I chose a minimum of 10 instances per the leaf, which I suspect prevents the model from overfitting by identifying more robust predictors that apply to multiple subjects, and an unnecessarily-large number of trees (10,000) to account for the large number of possible feature combinations.

- **E. H. Kim (ehk02004)**

  Matthew Desmond, according to his talk on his book Eviction, argued that having children was posi-

tively related to being evicted, implying that the disruptiveness/chaos that comes with having children and the inherent lack of calm in the household was responsible for a family being evicted. As such, in addition to looking at how ratings of how calm the atmosphere of the house is when the child is 9, I include an index variable of the child exhibiting appropriate behavior at age 9 (as our continuous variable) and I include the binary variable for whether or not a childÍs behavioral/social problems was discussed with the teacher during the last school year when the child is 9, as both seemed likely to shed light on the disruptiveness/chaos in the household that might be affiliated with eviction. In addition, it was assumed that the father's race (thinking of race literature), the mother's age at childbirth, and how often the children had a tendency to move were affiliated with layoff - the latter variables included based off of intuitive assumptions.

- **Ben Leizman and Catherine Wu (bleizman)**

  In this dataset, we imputed values for missing data by using the feature's highest frequency positive value and then normalized all values. We created datasets for each of the six outcomes, using only sample where the outcome is not NA. We reduced the feature set of each dataset from 12,805 to 100 using mutual information-based feature selection. To predict continuous outcomes such as GPA, Grit, and Material Hardship, we trained LASSO, Ridge, and Elastic Net regression models. To predict categorical outcomes such as Job Layoff, Job Training, and Eviction, we trained Ridge, K-Nearest Neighbors, and Multi-layer Perceptron classification models. We used default hyperparameters and selected an optimal model using k-fold cross validation. The regression models were evaluated on mean squared error and R^2, while the classification models were evaluated on precision, recall, and F1 score.

- **Naijia Liu (NaijiaLiu)**

  We developed an iterative feature selection method using Ridge regression.

- **Andrew E. Mack (aemack)**

  We trained random forest, gradient boosting and LASSO regression models using various hyper-parameters. We also used F-statistics to screen variables, with the number of variables used counting as an additional hyper-parameter. In total, we had 9 potential models for each outcome. For each of the 6 outcomes, we selected the best model using 10-fold cross-validation.

- **Mayank Mahajan (kapoor)**

  Feature selection was performed using a combination of randomized LASSO and RFE algorithms. For the continuous variable outcomes, OLS, ridge regression, LASSO, and Elastic Net models were all

tested for their training and test accuracy for predicting grit, GPA, and material hardship. For the binary outcomes, regularized logistic regression, AdaBoost regression, and Multinomial NB classifers were tested for their classification accuracy.

- **Noah Mandell (nmandell)**

  For a dataset as large and raw as this one, how the data is processed and cleaned can be crucial. We immediately dropped all feature columns with zero variance, along with any column containing strings. We also dropped columns with more than 30% of the data missing. This reduces the number of columns significantly, but it also reduces the ratio of missing to non-missing data, which is the goal of this step. We choose the threshold of 30% missing per column because this produces a dataset with only 20% of the data missing across all columns. We then attempted to distinguish categorical variables from continuous ones by noting that the numerical coding for the categorical variables used only integers and did not use values in the range 20-99. This resulted in 90% of the variables being labeled categorical. We then used a one-hot encoding for the categorical data, and we filtered out one-hot-encoded columns with low variance. We then used Ridge regression to simultaneously predict all six outcomes. We used a nested cross-validation procedure, with an inner cross-validation loop for fitting hyper-parameters, and an outer loop for evaluating model performance.

- **Malte Möser (malte)**

  For preprocessing, I removed highly correlated features as well as those with low variance. Then, I added indicator variables for questions that were skipped or where the respondents refused to answer. I imputed all missing values in the dataset, with mean imputation for numerical features and mode imputation for categorical features, and standardized the numerical features. For prediction, I used a generalized linear model as provided by the R package 'glmnet', with hyperparameter tuning based on repeated cross-validation using the R package 'caret'.

- **Katariina Mueller-Gastell (katamg)**

  I used existing sociological theory to identify features that would plausibly be correlated with other parental characteristics. I then included these features in simple OLS and logit models, taking care not to overfit the training data by using my own train/test split. For example, I found that whether the mother had breastfed any of her children was a fairly good predictor of child outcomes.

- **Qiankun Niu and Kengran Yang (hty)**

  In this project, we first explored the data from the survey with large feature sets. We observed that there are many missing values in the data so we first cleaned the data according to missing value, constant

values and perfect co-linearity. Then we explored different imputation methods. After cleaning and imputing the data, we adopted various regression and classification methods and chose random forest due to the high dimensionality and complexity of variable relationship.

- **William Nowak (wnowak)**

  Built ensemble models using only independent features from mother, father, and other individual contributors.

- **Hamidreza Omidvar (hamidrezaomidvar)**

  First we used MICE imputation method to fill missing data. In the next step, we calculated the correlation of all features for each outcome. After this step, we selected top correlated features (either negative or positive correlation) as the main features. Finally, we used linear (for continuities outputs) and logistic (for binary outputs) regressions to predict the outputs.

- **Karen Ouyang and Julia Wang (kouyang)**

  At the time of our submission, the Fragile Families dataset was the largest and sparsest with which we had worked, so we focused on sanitizing the dataset and engineering useful features. In addition to imputing missing data, we iteratively tested different Pearson correlation coefficient thresholds to select features. Of the six models that we trained and tuned, the elastic net regression model consistently made the most accurate predictions.

- **Katy M. Pinto (Katy_P)**

  In the submissions, my approach to the Challenge was to focus on established relationships between variables to create models for the six outcomes based on prior research. I focused mainly on constructed variables. I focused on parental background (e.g. parent's education, race/ethnicity), household structure (e.g. marital status, income, number of children) and child's individual characteristics (e.g hours spent on homework, gifted) as predictors of GPA, Grit, Material Hardship, Eviction, Job Loss, and Job Training. The attempts included OLS for the continuous outcome variables, logistic regression for the binary outcome variables. I attempted one multiple imputation technique that did not provide better fit in my model compared to recoding missing variables to mean/median/mode and including a flag for missing in models. In the end, the final models submitted were much simpler in approach, compared to some of the early models I attempted with more predictor variables. I also compared my submissions based on the leadership board and the simpler models with fewer predictors seemed to perform better than models I submitted with more predictors.

- **Ethan Porter (lennyc)**

I only included variables for which I had a substantive reason to believe might affect the outcome.

- **Kristin E. Porter and Tejomay Gadgil (mdrc)**

  MDRC applied several analytic steps in our predictive analytics framework to the Fragile Families Challenge (FFC). Those focused on data processing, creating and curating measures, and modeling methods. The following describes the underlying premises that guided our analyses: (1) Invest deeply in measure creation, combining both substantive knowledge and automated approaches. (2) "Missingness" is informative and should not be "imputed away." (3) Eliminate unhelpful measures (those with very little variation, those that are redundant and those that did not apply to the primary care giver). (4) Evaluate "learners" based on out-of-sample performance, using cross-validation. In MDRC's predictive analytics framework, we define a "learner" as some combination of (a) a set of predictors, (b) a modeling method or machine learning algorithm, and (c) any tuning parameters for the corresponding machine learning algorithm. (5) Combine results from different learners with ensemble learning.

- **Crystal Qian and Jonathan D. Tang (cjqian)**

  We preprocessed the Fragile Families data through pruning based on answer ratio (i.e., features that were missing more often were regarded as less important) and mapping all string-based features to integers to make them suitable for regression. We performed imputation by training a regressor on the labeled data, with no missing values, assigned to corresponding training classes. To address and take advantage of data sparsity (approx. 17% of cells in the dataframe were empty), we eliminated approximately 25% of the 13,000 features that had the lowest ratio of non-NA responses to total responses, encoded string responses, and used mean value/regression-based imputation to further prune NA responses. Afterwards, we used it to predict the missing values within the training set, as well as predicting the values for the entirety of the test classes. We concatenated the now-filled training classes and test classes, making a final prediction array. Then, we applied regression-based prediction techniques (LASSO, etc) for both the discrete and continuous predictors, in part incentivised by Brier score leaderboard scoring. We tested various regressors, including Lasso, Lasso with Least Angle Regression, Elastic net, and Ridge regression, for effectiveness in predicting GPA, Grit, Material Hardship, Eviction, Layoff, and Job Training. We did k-folds cross validation (k=5) in order to locally evaluate our different models. Ultimately, we found that Lasso regression performed the best for us on this dataset (the questions asked in this study could be highly correlated, explaining the success of Lasso regression). Using imputation in order to generate our own value-completed training data was extremely helpful. Interestingly, the three most predictive features to our models included the questions, "In past year, you shouted, yelled, or screamed at child.", "Is there someone you could

count on to co-sign bank loan for $5000?", and "There were problems with neighborhood safety."

- **Tamkinat Rauf (Tamkinat)**

  I predicted GPA, grit, material hardship, eviction, job training, and layoff. I drew on past research for initially selecting predictors, and then used MSE as well as leaderboard positions to tweak my models. In general, I used the most parsimonious models possible. I used logit models for binary outcomes and OLS for continuous outcomes. My participation in the Fragile Families Challenge was part of an exercise for a statistical methods course. While I independently conducted and submitted my predictions, this was really a joint effort with my professor, Jeremy Freese, and 15 other colleagues who took the course. We shared code for constructing variables and frequently discussed our modeling strategies.

- **Thomas Schaffner and Andrew Or (t.f.schaffner and andrewor)**

  This submission consists of work by two participants. We evaluated multiple data imputation strategies and predictive models, learning predictors for each outcome variable separately. After evaluating several indicative measurements (R^2, MSE, precision, recall, accuracy, and F1 scores), we selected an imputation strategy that replaced missing data and removed string-valued variables in conjunction with random forest predictors. We then programmatically tuned the random forest hyperparameters to arrive at our final predictions.

- **Landon Schnabel (lpschnab)**

  I produced individual predictions and also worked with a group (the IU_Sociology team). Initially, I developed more complex models using a large number of variables driven largely by what seemed to matter in the training data. I ultimately decided, however, to use a simpler and more theoretically-driven approach with just a few variables and basic methods. In my final submission, I used just the following variables and linear regression to predict GPA: parental education at baseline, the child's earlier score on a Woodcock-Johnson test, and the child's earlier grit.

- **Bryan Schonfeld (signoret)**

  We read through the literature to find substantively important variables. We divided the data into a test set and training set, and used a variety of regression and statistical learning tools (logistic regression, linear regression, LASSO, etc) to find the best predictors.

- **Ben Sender (sender)**

  Our study leveraged seven prediction models: Linear Regression, Lasso Regression, Ridge Regression, Logistic Regression, Random Forest, Neural Network, and Naive Bayes. We imputed missing features

using the mode from other years, and selected features for each outcome using chi-squared. To tune and evaluate the models we created a test set with 10% of the families from the original training set. We evaluated binary models based on accuracy, and continuous models on mean squared error and R-squared. The results for our test set were similar to the results for our Fragile Families submission.

- **Emma Tsurkov (ETsurkov)**

  My approach to the Fragile Families Challenge was based on utilizing my background in law to try and create parsimonious but effective model, focusing on the eviction outcome. Although eviction is a mostly exogenous shock, I wanted to examine it as an outcome of an institutional process. Eviction, more than the other outcomes in the Fragile Families Challenge is a result of legal action. Accordingly, I have conducted research into eviction law. I found that smoking is prohibited in public housing and that smoking even inside one's housing unit without causing a fire or any damage can serve as grounds for eviction. Additionally, I found that many states and localities have strict anti-smoking laws in multi-unit buildings, and that even if there is no law prohibiting smoking in the unit, landlords can prohibit smoking and use smoking as grounds for eviction. This led me to believe that mother's smoking, might be a good predictor of eviction, whether as a genuine reason or a pretext used by landlords trying to remove undesirable tenants. I tested different model specifications and chose the best performing model, with basic covariates of the mother's education and race.

- **Austin van Loon (Alpaca_CultureAsAWoolkit)**

  I used background knowledge to select a small set of variables that seemed likely to matter for the outcomes. I would iteratively (a) predict missing values of the dependent variable using my set of variables (b) find which of the remaining independent variables had the most missing values (c) remove that variable from my set of variables and (d) repeat.

- **Onur Varol (ovarol)**

  My approach consists of feature categorization and model selection. First, I parsed all codebooks to extract information about the panel, survey respondents, and keyword-based labels. Later I selected different feature groups and evaluated their performance on cross-validated random-forest models. Both feature imputation and filtering missing values are tested, and removal of the missing values are performed for all of the analysis. Features having high importance score and models having success on the leaderboard kept for the next iteration of model construction.

- **Samantha Weissman (samantha_malte)**

  We took a systematic approach towards identifying relevant features for predicting outcomes in the

Fragile Families Challenge, using a combination of imputation, feature selection, and cross-validated model selection. To impute the data we code skipped and refused answers through new binary feature vectors, replaced categorical values with the mode and numerical values with the mean of each feature, removed low variance and highly correlated features (based on Spearman correlation), converted all categorical values into indicator variables and scaled continuous variables to have a mean of 0 and unit variance. To further reduce the dimensionality of the data we used Lasso regression and elastic net regression with cross-validation to find the best hyperparameters, as well as cross-validated recursive feature elimination (RFE) based on a support vector machine with a linear kernel and a step size of 5. We implemented 5 different classifiers and regressors (AdaBoost, Gaussian Process, Linear regression, random forest, and SVM). Finally for cross-validation and hyperparameter tuning we employed different techniques per classifier/regressor, including 10-fold cross-validation and grid search, and evaluate both a linear and a Gaussian kernel.

- **Yue Gao, Jingwen Yin and Chenyun Zhu (aurora1994)**

  We first solved the missing data issue by making NA a special level in categorical features and imputing the missing value with median in continuous features. After data cleaning and missing data imputing, we separated the features into categorical variables and continuous variables. We used random forest based feature selection method to select a few significant features for each outcome. We used Boruta Package to conduct feature selection which works as wrapper algorithm around Random Forest. Various machine learning algorithms were evaluated based on the selected features and the best algorithm for each outcome was selected using MSE. We tried a series of models including linear regression, full tree, pruned tree, random forest, conditional inference trees, stochastic gradient boosting, support vector machine, linear bagging, ensemble linear regression and random forest, ensemble support vector and random forest, linear discriminant analysis, C5.0, and KNN. Based on the root mean squared error metric for continuous outcome variables and accuracy metric for binary outcome variables, we predicted the final results by using random forest for GPA, eviction, job training, using stochastic gradient boosting for grit, material hardship, and doing LDA for layoff. Finally we retrained the models using full dataset and submitted to the leaderboard.

- **Bingyu Zhao, Kirstie Whitaker, Maria K Wolters, and Bernie Hogan (bz247)**

  The submission of our team adopted a few simple steps to make predictions. First, basic data cleaning was carried out, which involved selecting only the continuous variables as the predictors (removing all categorical variables), imputing the empty entries by the mean and removing constant (0 variance) columns. This left 1,771 variables remaining in the dataset. In the second step, Principle Component

Analysis was conducted and the top 50 principle components were kept as the model covariates. In the last step, multi-variable linear regression was used to model the continuous dependent variables and logistic regression to model the binary outcomes. In the end, our submission performed better than the benchmark data in two out of the six outcome variables, including the material hardship and GPA, while both are continuous outcomes.

S5 Specificity and generality of the Fragile Families Challenge

The predictability of life outcomes observed in the Fragile Families Challenge is specific to this prediction task. We expect that predictability would have been higher if the Fragile Families study had 1) more families, 2) more predictors, 3) less measurement error, or 4) data in a format that was easier to use [15]. Likewise, we expect that predictability would have been lower if the Fragile Families study had 1) fewer families, 2) fewer predictors, 3) more measurement error, or 4) data in a format that was harder to use. We do not know the size of the changes that would be needed to produce a qualitative change in predictability. Further, these same four issues would arise, in varying degrees, in research using any high-quality, longitudinal survey data, because these studies all use similar data collection techniques and face similar budget constraints.

Despite the broad similarity between the Fragile Families study and other longitudinal survey studies, we speculate that two specific features of the Fragile Families study may lead to lower predictability: study timing and study population. Three aspects of timing may decrease predictability: the six year gap between waves 5 and 6, a large social disruption (the Great Recession [10]) during this gap, and the collection of wave 6 data when the child was 15 years old, which may be a particularly turbulent time for children and families. Further, the Fragile Families study population—a largely urban and disadvantaged group living in the contemporary United States—may have more unpredictable lives than other groups.

In addition to the characteristics of the Fragile Families study, the results of the Challenge may also depend in part on the decisions we made when designing the prediction task. For example, it is possible that the results of the Challenge would have been qualitatively different if we selected different outcomes from wave 6 (age 15). In fact, we did observe that some outcomes (e.g., material hardship and GPA) were more predictable than others (e.g., grit, eviction, job training, layoff). It is also possible that the results of the Challenge would have been qualitatively different if we picked a different evaluation metric [13] or if we made different decisions in our privacy and ethics audit [18]. Ultimately, we think that these and other questions about the specificity and generality of the results of the Challenge are important questions for future empirical research.

S6 Fragile Families Challenge provides building blocks for future research

The open-sourced submissions to the Challenge provide important building blocks for future research about the predictability of life outcomes. Here we sketch just two examples of such research.

First, researchers may wonder whether the results of the Challenge would have been qualitatively different if we selected different outcomes from wave 6 (age 15), if we picked a different evaluation metric [13], or if we made different decisions in our privacy and ethics audit [18] (see Sec. S5). These questions could be addressed empirically by re-purposing participants' code and techniques to predict all outcomes in wave 6 (age 15). Further, this analysis could be expanded to include predictors we chose not to share in the context of a mass collaboration for privacy reasons (e.g., geographic information) [18].

In addition to establishing robustness to our design choices, the submissions to the Challenge can also be used to focus the search for important, unmeasured predictors. The predictions made by each team can be used to identify families that are especially difficult to predict, given existing data and methods (Eq. S7 in Sec. S3). Researchers can then conduct in-depth interviews with these especially difficult-to-predict families in order to search for important variables that if collected might improve predictive performance [26]. These examples provide just two ways that we imagine that the open-sourced submissions can be used to advance future research on the predictability of life outcomes.

S7 Computing environment

The results in this paper were generated by code written in R (v 3.5.1) [21] using a number of packages: `Amelia` (v 1.7.4) [14], `broom` (v 0.4.2) [23], `corrr` (v 0.4.0) [24], `dplyr` (v 0.8.3) [35], `forcats` (v 0.4.0) [34], `foreach` (v 1.4.3) [1], `foreign` (v 0.8.67) [20], `ggplot2` (v 2.2.1.9000) [30], `ggridges` (v 0.4.1) [39], `grid` (v 3.3.3) [21], `gridExtra` (v 2.2.1) [2], `haven` (v 1.1.0) [38], `here` (v 0.1) [19], `magrittr` (v 1.5) [3], `mvtnorm` (v 1.0.6) [11], `quadprog` (v 1.5.5) [28], `ranger` (v 0.9.0) [41], `readr` (v 1.1.1) [37], `readstata13` (v 0.8.5) [9], `reshape2` (v 1.4.3) [29], `scales` (v 0.5.0) [31], `stargazer` (v 5.2) [12], `stringr` (v 1.2.0) [32], `tidyr` (v 0.7.2) [36], `tidytext` (v 0.1.7) [27], `tidyverse` (v 1.2.1) [33].

Automated function extraction from code submissions was performed in Python 3 using the Pygments (`http://pygments.org/`) syntax highlighting library and the rpy2 (`https://rpy2.readthedocs.io/en/version_2.8.x/`) package. Computations were done on a machine running Mac OSX 10.13.3.

S8 Author information

S8.1 Authorship

We offered participants in the Fragile Families Challenge two opportunities to publish their results based on an authorship policy set at the beginning of the Challenge. First, two us (Salganik and McLanahan) were guest editors of a Special Collection of the journal *Socius* where all participants could submit a manuscript describing their approach to the Challenge. These manuscripts were all peer-reviewed and evaluated independently of the predictive performance of the approach they described. Second, we promised participants who made a meaningful contribution to the Challenge that they would have the opportunity to be a co-author of this paper, if they wished. We operationalized this promise by offering co-authorship to participants if they were part of an account that made a qualifying submission for one or more outcomes. There were 115 such accounts, and these are the accounts that are presented in Figure 4 of the main text. Participants who wished to be a co-author needed to fill out an online feedback form that collected information about their submission and their overall assessment of the manuscript. The online feedback form also offered participants the opportunity to upload detailed feedback about the manuscript.

Our outreach to the participants in the Challenge about co-authorship of this paper proceeded in two stages. In the first stage, we invited all 31 authors of papers in the *Socius* Special Collection to be co-authors of this paper. All of these authors were part of an account that made a qualifying submission. 25 authors representing 16 accounts initially agreed and completed the feedback form. For the 6 authors that did not respond, we sent them one or more follow-up emails as part of the second stage of outreach (described more below). Of these 6, 2 completed the feedback form, and 4 did not respond.

In the second stage, we invited all other participants. Because the application process and the Challenge scoring process were run on different platforms (Google Forms, Qualtrics, and CodaLab), we do not have a precise mapping between participants and accounts. Further, for each submission, we do not have a list of all participants who created the submission (we began asking for this information in the middle of the Challenge). Therefore, we emailed all 429 applicants who were not authors of papers in the *Socius* Special Collection, and we offered them a chance to be a co-author of this paper if they contributed to the submission from one of the qualifying accounts. Of these emails, 55 bounced back to us, and we attempted to reach these participants through other means (e.g., other email addresses, social media, former employers, etc.). We received 77 completed feedback forms by the deadline in our email. We also received a number of other kinds of responses: 8 people responded that they did not wish to be a co-author; 14 responded that they were not eligible to be co-authors; 1 was deceased; 7 responded to our email in some way but did not

complete the feedback form (e.g., asking a question). We sent a follow-up message to 332 participants (those that responded in some way but did not complete the form or those that did not respond) with an extended deadline. In response to the follow-up emails, we received 18 additional completed feedback forms. Of the 92 completed responses[3] to the feedback form in the second stage, 75 responses corresponded to one of the qualifying accounts. These 75 responses cover 53 accounts. Some of the invalid responses on the feedback form were caused by participants associated with accounts that did not qualify and some were caused by a coding error on our part that led us to initially use an incorrect (and overly large) list of qualifying accounts. Overall, across both stages, we received feedback from 100 co-authors describing 69 accounts.

During this process of inviting co-authors, we also identified 5 people who contributed to a submission, but for whom we do not have a record of them applying for data access. This could occur for a variety of reasons, such as participating in a large team (which might not require direct data access) or an error in our records. We investigated these 5 cases and found that in each case, the co-author participated as part of a team with at least one other person who had applied for and been granted data access. Further, in each case, we would have granted data access if the person had applied. We emailed these co-authors with an update about this situation and explained the reasons behind our application process.

[3]This count excludes 2 opt-outs and 1 erroneous submission.

S8.2 Funding

This study was supported by the Russell Sage Foundation, NSF (1760052), and NICHD (P2-CHD047879). Funding for FFCWS was provided by the NICHD (R01-HD36916, R01-HD39135, R01-HD40421) and a consortium of private foundations, including the Robert Wood Johnson Foundation. These authors would like to acknowledge additional funding:

- Caitlin E. Ahearn and Jennie E. Brand: The author benefited from facilities and resources provided by the California Center for Population Research at UCLA (CCPR), which receives core support (P2C-HD041022) and training support (T32HD007545) from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD).
- Khaled Al-Ghoneim: Work was conducted while visiting the Media Lab, MIT.
- Drew M. Altschul: MRC Mental Health Data Pathfinder award (MC_PC_17209).
- Barbara E. Engelhardt: NIH R01 HL133218 and an NSF CAREER AWD1005627.
- Anna Filippova, Connor Gilroy, Ridhi Kashyap, Antje Kirchner, Allison C. Morgan, Kivan Polimis, and Adaner Usmani were supported in part by the Russell Sage Foundation. Support for the computational resources for this research came from a Eunice Kennedy Shriver National Institute of Child Health and Human Development research infrastructure grant (P2C HD042828) to the Center for Studies in Demography and Ecology at the University of Washington.
- Josh Gagné: Institute of Education Sciences Grant R305B140009.
- Brian J. Goode and Debanjan Datta: Partially supported by DARPA Cooperative Agreement D17AC00003 (NGS2).
- Moritz Hardt: NSF Career Award (1750555).
- Eaman Jahani: This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1122374. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation.
- Patrick Kaminski: National Science Foundation grant 1735095, NRT-Interdisciplinary Training in Complex Networks and Systems.
- Alexander T. Kindel: National Science Foundation Graduate Research Fellowship.

- Arvind Narayanan: National Science Foundation grant IIS-1704444.
- Alex Pentland: Toshiba Professor of Media Arts and Sciences Chair.
- Katy M. Pinto: The author benefited from the facilities and resources provided by the UCLA, Chicano Studies Research Center during her Visiting Scholar appointment at the center.
- Kirstie Whitaker, Maria K. Wolters, and Bingyu Zhao: Turing Research Fellowship under EPSRC grant EP/N510129/1 (The Alan Turing Institute) and the Cambridge Trust.

S9 Acknowledgements

The Fragile Families Challenge was overseen by a Board of Advisors consisting of Jeanne Brooks-Gunn, Kathryn Edin, Barbara Engelhardt, Irwin Garfinkel, Moritz Hardt, Dean Knox, Nicholas Lemann, Karen Levy, Sara McLanahan, Arvind Narayanan, Timothy Nelson, Matthew Salganik, Brandon Stewart, and Duncan Watts. We would also like to thank the Fragile Families study data team—Kristin Catena, Tom Hartshorne, Kate Jaeger, Dawn Koffman, and Shiva Rouhani—for their assistance during the Challenge.

References

[1] Revolution Analytics and Steve Weston. *foreach: Provides Foreach Looping Construct for R*, 2015. R package version 1.4.3.

[2] Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2016. R package version 2.2.1.

[3] Stefan Milton Bache and Hadley Wickham. *magrittr: A Forward-Pipe Operator for R*, 2014. R package version 1.5.

[4] Leo Breiman. Stacked regressions. *Machine Learning*, 24(1):49–64, 1996.

[5] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[6] Matthew Desmond, Ashley Gromis, Lavar Edmonds, James Hendrickson, Katie Krywokulski, Lillian Leung, and Adam Porton. *Eviction Lab National Database: Version 1.0.* Princeton University, `www.evictionlab.org`, 2018.

[7] Matthew Desmond and Rachel Tolbert Kimbro. Eviction's fallout: Housing, hardship, and health. *Social Forces*, 94(1):295–324, 2015.

[8] Angela L. Duckworth, Christopher Peterson, Michael D. Matthews, and Dennis R. Kelly. Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, 92(6):1087, 2007.

[9] Jan Marvin Garbuszus and Sebastian Jeworutzki. *readstata13: Import 'Stata' Data Files*, 2016. R package version 0.8.5.

[10] Irwin Garfinkel, Sara S McLanahan, and Christopher Wimer. *Children of the Great Recession*. Russell Sage Foundation, 2016.

[11] Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, and Torsten Hothorn. *mvtnorm: Multivariate Normal and t Distributions*, 2017. R package version 1.0-6.

[12] Marek Hlavac. *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Harvard University, Cambridge, USA, 2015. R package version 5.2.

[13] Jake M. Hofman, Amit Sharma, and Duncan J. Watts. Prediction and explanation in social systems. *Science*, 355(6324):486–488, 2017.

[14] James Honaker, Gary King, and Matthew Blackwell. Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7):1–47, 2011.

[15] Alexander Kindel, Vineet Bansal, Kristin Catena, Thomas Hartshorne, Kate Jaeger, Dawn Koffman, Sara McLanahan, Maya Phillips, Shiva Rouhani, Ryan Vinh, and Matthew J. Salganik. Improving metadata infrastructure for complex surveys: Insights from the Fragile Families Challenge. *SocArXiv*, 2018.

[16] Tarald O. Kvålseth. Cautionary note about R^2. *The American Statistician*, 39(4):279–285, 1985.

[17] Ian Lundberg and Louis Donnelly. A research note on the prevalence of housing eviction among children born in US cities. *Demography*, 56(1):391–404, 2019.

[18] Ian Lundberg, Arvind Narayanan, Karen Levy, and Matthew J Salganik. Privacy, ethics, and data access: A case study of the Fragile Families Challenge. *arXiv preprint arXiv:1809.00103*, 2018.

[19] Kirill Müller. *here: A Simpler Way to Find Your Files*, 2017. R package version 0.1.

[20] R Core Team. *foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ...*, 2016. R package version 0.8-67.

[21] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[22] Nancy E. Reichman, Julien O. Teitler, Irwin Garfinkel, and Sara S. McLanahan. Fragile Families: Sample and design. *Children and Youth Services Review*, 23(4-5):303–326, 2001.

[23] David Robinson. *broom: Convert Statistical Analysis Objects into Tidy Data Frames*, 2017. R package version 0.4.2.

[24] Edgar Ruiz, Simon Jackson, and Jorge Cimentada. *corrr: Correlations in R*, 2019. R package version 0.4.0.

[25] Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model Assisted Aurvey Sampling*. Springer Science & Business Media, 2003.

[26] Jason Seawright. The case for selecting cases that are deviant or extreme on the independent variable. *Sociological Methods & Research*, 45(3):493–525, 2016.

[27] Julia Silge and David Robinson. tidytext: Text mining and analysis using tidy data principles in R. *JOSS*, 1(3), 2016.

[28] Berwin A. Turlach and Andreas Weingessel. *quadprog: Functions to Solve Quadratic Programming Problems*, 2013. R package version 1.5-5.

[29] Hadley Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20, 2007.

[30] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[31] Hadley Wickham. *scales: Scale Functions for Visualization*, 2017. R package version 0.5.0.

[32] Hadley Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*, 2017. R package version 1.2.0.

[33] Hadley Wickham. *tidyverse: Easily Install and Load the 'Tidyverse'*, 2017. R package version 1.2.1.

[34] Hadley Wickham. *forcats: Tools for Working with Categorical Variables (Factors)*, 2019. R package version 0.4.0.

[35] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2019. R package version 0.8.3.

[36] Hadley Wickham and Lionel Henry. *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*, 2017. R package version 0.7.2.

[37] Hadley Wickham, Jim Hester, and Romain Francois. *readr: Read Rectangular Text Data*, 2017. R package version 1.1.1.

[38] Hadley Wickham and Evan Miller. *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*, 2017. R package version 1.1.0.

[39] Claus O. Wilke. *ggridges: Ridgeline Plots in 'ggplot2'*, 2017. R package version 0.4.1.

[40] David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

[41] Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017.