Measuring the predictability of life outcomes with a scientific mass collaboration


How predictable are life trajectories? We investigated this question using a scientific mass collaboration using the common task method: 160 teams built predictive models for six life outcomes using data from the Fragile Families and Child Wellbeing Study, a high-quality birth cohort study. Despite using a rich dataset and applying machine-learning methods optimized for prediction, the best predictions were not very accurate and were only slightly better than those from a simple benchmark model. Within each outcome, prediction error was strongly associated with the family being predicted and weakly associated with the technique used to generate the prediction. Overall, these results suggest practical limits to the predictability of life outcomes in some settings and illustrate the value of mass collaborations in the social sciences.

Social scientists studying the life course have described social patterns, theorized factors that shape outcomes, and estimated causal effects. Although this research has advanced scientific understanding and informed policy interventions, it is unclear how much it translates into an ability to predict individual life outcomes. Assessing predictability is important for three reasons. First, accurate predictions can be used to target assistance to children and families at risk (1, 2). Second, predictability of a life outcome from a person’s life trajectory can indicate social rigidity (3), and efforts to understand differences in predictability across social contexts can stimulate scientific discovery and improve policy-making (4). Finally, efforts to improve predictive performance can spark developments in theory and methods (5).

In order to measure the predictability of life outcomes for children, parents, and households, we created a scientific mass collaboration. Our mass collaboration—the Fragile Families Challenge—used a research design common in machine learning but not yet common in the social sciences: the common task method (6). To create a project using the common task method, an organizer designs a prediction task and then recruits a large, diverse group of researchers who complete the task by predicting the exact same outcomes using the exact same data. These predictions are then evaluated with the exact same error metric that exclusively assesses their ability to predict held-out data: data that are held by the organizer and not available to participants. Although the structure of the prediction task is completely standardized, participants are free to use any technique to generate predictions.

The common task method produces credible estimates of predictability because of its design. If predictability is higher than expected, the results cannot be dismissed because of concerns about overfitting (7) or researcher degrees of freedom (8). Alternatively, if predictability is lower than expected, the results cannot be dismissed because of concerns about the limitations to the Fragile Families Challenge are currently available to approved researchers and C.Z. analyzed data; and M.J.S., I.L., A.T.K., D.K., and S. McLanahan wrote the paper.

Contributed by Sara McLanahan, January 24, 2020 (sent for review October 1, 2019; reviewed by Sendhil Mullainathan and Brian Uzzi)


Reviewers: S.M., University of Chicago; and B.U., Northwestern University.

Competing interest statement: B.E.E. is on the scientific advisory boards of Celsius Therapeutics and Freenome, is currently employed by Genomics plc and Freenome, and is on a year leave-of-absence from Princeton University.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

Data deposition: Predictions, code, and narrative explanations for valid submissions to the Fragile Families Challenge and code to reproduce the results of this paper are available from Dataverse at https://doi.org/10.7910/DVN/CXSECU. Data used in the Fragile Families Challenge are currently available to approved researchers from the Princeton University Office of Population Research Data Archive at https://opr.princeton.edu/archive.

1To whom correspondence may be addressed. Email: mj33@princeton.edu or mclanahan@princeton.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1915006117/-/DCSupplemental.
Significance

Hundreds of researchers attempted to predict six life outcomes, such as a child’s grade point average and whether a family would be evicted from their home. These researchers used machine-learning methods optimized for prediction, and they drew on a vast dataset that was painstakingly collected by social scientists over 15 y. However, no one made very accurate predictions. For policymakers considering using predictive models in settings such as criminal justice and child-protective services, these results raise a number of concerns. Additionally, researchers must reconcile the idea that they understand life trajectories with the fact that none of the predictions were very accurate.

Each data collection module is made up of ∼10 sections, where each section includes questions about a specific topic (e.g., marriage attitudes, family characteristics, demographic characteristics). Information about the topics included in each module is presented in SI Appendix, section S1.1. During the Fragile Families Challenge, data from waves 1 to 5 (birth to age 9 y) were used to predict outcomes in wave 6 (age 15 y).

Fig. 1. Data collection modules in the Fragile Families study. Each module is made up of ∼10 sections, where each section includes questions about a specific topic (e.g., marriage attitudes, family characteristics, demographic characteristics). Information about the topics included in each module is presented in SI Appendix, section S1.1. During the Fragile Families Challenge, data from waves 1 to 5 (birth to age 9 y) were used to predict outcomes in wave 6 (age 15 y).
of the Fragile Families Challenge. All predictions were evaluated based on a common error metric: mean squared error (SI Appendix, section S1.2). SI Appendix, section S1.1 includes more information about the background, training, leaderboard, and holdout data.

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 visiting universities, courses, and scientific conferences to host workshops to help participants get started. Ultimately, we received 457 applications to participate from researchers in a variety of fields and career stages (SI Appendix, section S1.3). All procedures for the Fragile Families Challenge were approved by the Princeton University Institutional Review Board (no. 8061).

As noted above, participants in the Fragile Families Challenge attempted to minimize the mean squared error of their predictions on the holdout data. To aid interpretation and facilitate comparison across the six outcomes, we present results in terms of $R^2_{\text{Holdout}}$, which rescales the mean squared error of a prediction by the mean squared error when predicting the mean of the training data (SI Appendix, section S1.2).

$$R^2_{\text{Holdout}} = 1 - \frac{\sum_{i \in \text{Holdout}} (y_i - \hat{y}_i)^2}{\sum_{i \in \text{Holdout}} (y_i - \bar{y}_{\text{Training}})^2}.$$  [1]

$R^2_{\text{Holdout}}$ is bounded above by 1 and has no lower bound. It provides a measure of predictive performance relative to two reference points. A submission with $R^2_{\text{Holdout}} = 0$ is no more accurate than predicting the mean of the training data, and a submission with $R^2_{\text{Holdout}} = 1$ is perfectly accurate.

Results

Once the Fragile Families Challenge was complete, we scored all 160 submissions using the holdout data. We discovered that even the best predictions were not very accurate: $R^2_{\text{Holdout}}$ of about 0.2 for material hardship and GPA and about 0.05 for the other four outcomes (Fig. 3). In other words, even though the Fragile Families data included thousands of variables collected to help scientists understand the lives of these families, participants were not able to make accurate predictions for the holdout cases. Further, the best submissions, which often used complex machine-learning methods and had access to thousands of predictor variables, were only somewhat better than the results from a simple benchmark model that used linear regression (continuous outcomes) or logistic regression (binary outcomes) with four predictor variables selected by a domain expert: three variables about the mother measured at the child’s birth (race/ethnicity, marital status, and education level) and a measure of the outcome—or

![Fig. 2. Datasets in the Fragile Families Challenge. During the Fragile Families Challenge, participants used the background data (measured from child’s birth to age 9 y) and the training data (measured at child age 15 y) to predict the holdout data as accurately as possible. While the Fragile Families Challenge was underway, participants could assess the accuracy of their predictions in the leaderboard data. At the end of the Fragile Families Challenge, we assessed the accuracy of the predictions in the holdout data.](image)

![Fig. 3. Performance in the holdout data of the best submissions and a four variable benchmark model (SI Appendix, section S2.2). A shows the best performance (bars) and a benchmark model (lines). Error bars are 95% confidence intervals (SI Appendix, section S2.1). B–D compare the predictions and the truth; perfect predictions would lie along the diagonal. E–G show the predicted probabilities for cases where the event happened and where the event did not happen. In B–G, the dashed line is the mean of the training data for that outcome.](image)
Fig. 4. Heatmaps of the squared prediction error for each observation in the holdout data. Within each heatmap, each row represents a team that made a qualifying submission (sorted by predictive accuracy), and each column represents a family (sorted by predictive difficulty). Darker colors indicate higher squared error; scales are different across subfigures; order of rows and columns are different across subfigures. The hardest-to-predict observations tend to be those that are very different from the mean of the training data, such as children with unusually high or low GPAs (SI Appendix, section S3). This pattern is particularly clear for the three binary outcomes—eviction, job training, layoff—where the errors are large for families where the event occurred and small for the families where it did not.

a closely related proxy—measured when the child was 9 y (Fig. 3) (SI Appendix, section S2.2). Finally, we note that our procedure—using the holdout data to select the best of the 160 submissions and then using the same holdout data to evaluate that selected submission—will produce slightly optimistic estimates of the performance of the selected submission in new holdout data, but this optimistic bias is likely small in our setting (SI Appendix, section S2.4).

Beyond identifying the best submissions, we observed three important patterns in the set of submissions. First, teams used
a variety of different data processing and statistical learning techniques to generate predictions (*SI Appendix*, section S4). Second, despite diversity in techniques, the resulting predictions were quite similar. For all outcomes, the distance between the most divergent submissions was less than the distance between the best submission and the truth (*SI Appendix*, section S3). In other words, the submissions were much better at predicting each other than at predicting the truth. The similarities across submissions meant that our attempts to create an ensemble of predictions did not yield a substantial improvement in predictive accuracy (*SI Appendix*, section S2.5). Third, many observations (e.g., the GPA of a specific child) were accurately predicted by all teams, whereas a few observations were poorly predicted by all teams (Fig. 4). Thus, within each outcome, squared prediction error was strongly associated with the family being predicted and weakly associated with the technique used to generate the prediction (*SI Appendix*, section S3).

**Discussion**

The Fragile Families Challenge provides a credible estimate of the predictability of life outcomes in this setting. Because of the design of the Fragile Families Challenge, low predictive accuracy cannot easily be attributed to the limitations of any particular researcher or approach; hundreds of researchers attempted the task, and none could predict accurately. However, the Fragile Families data sources, and for other social groups (*SI Appendix*, section S5). Nonetheless, the results in this specific setting have implications for scientists and policymakers, and they suggest directions for future research.

Social scientists studying the life course must find a way to reconcile a widespread belief that understanding has been generated by these data—as demonstrated by more than 750 published journal articles using the Fragile Families data (10)—with the fact that the very same data could not yield accurate predictions of these important outcomes. Reconciling this understanding/prediction paradox is possible in at least three ways. First, if one measures our degree of understanding by our ability to predict (8, 15), then the results of the Fragile Families Challenge suggest that our understanding of child development and the life course is actually quite poor. Second, one can argue that prediction is not a good measure of understanding and that understanding can come from description or causal inference. For example, in the study of racial disparities, researchers may build understanding by carefully describing the black–white wealth gap (16), even if they are not able to accurately predict the wealth of any individual. Third, one can conclude that the prior understanding is correct but incomplete because it lacks theories that explain why we should expect outcomes to be difficult to predict even with high-quality data. Insights for how to construct such theories may come from research on the weather (17), the stock market (18), and other phenomena (19–22) where unpredictability is an object of study.

Policymakers using predictive models in settings such as criminal justice (23) and child-protective services (24) should be concerned by these results. In addition to the many serious legal and ethical questions raised by using predictive models for decision-making (23–26), the results of the Fragile Families Challenge raise questions about the absolute level of predictive performance that is possible for some life outcomes, even with a rich dataset. Further, the results raise questions about the relative performance of complex machine-learning models compared with simple benchmark models (26, 27). In the Fragile Families Challenge, the simple benchmark model with only a few predictors was only slightly worse than the most accurate submission, and it actually outperformed many of the submissions (*SI Appendix*, section S2.2). Therefore, before using complex predictive models, we recommend that policymakers determine whether the achievable level of predictive accuracy is appropriate for the setting where the predictions will be used, whether complex models are more accurate than simple models or domain experts in their setting (26–28), and whether possible improvement in predictive performance is worth the additional costs to create, test, and understand the more complex model (29). Ideally, these assessments would be carried out with government administrative data used in policy settings because the properties of these data likely differ from the properties of the Fragile Families data, but legal and privacy issues make it difficult for researchers to access many types of administrative data (29).

In addition to providing estimates of predictability in a single setting, the Fragile Families Challenge also provides the building blocks for future research about the predictability of life outcomes more generally. The predictions and open-sourced submissions from participants provide a data source for future study with the Fragile Families sample (*SI Appendix*, section S6). The Fragile Families Challenge also provides a template for one type of mass collaboration in the social sciences (30, 31). There are currently many longitudinal studies happening around the world, all with different study populations, measurement characteristics, and research goals. Each of these studies could serve as the basis for a mass collaboration similar to the Fragile Families Challenge. Progress made in these future mass collaborations might also reveal other social research problems that we can solve better collectively than individually.

**ACKNOWLEDGMENTS.** We thank the Fragile Families Challenge Board of Advisors for guidance and T. Hartshorne for research assistance. This study was supported by the Russell Sage Foundation, NSF Grant 1760052, and Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) Grant P2CHD047879. Funding for the Fragile Families and Child Wellbeing Study was provided by NICHD Grants R01-HD36916, R01-HD29113, and R01-HD040421; and a consortium of private foundations, including the Robert Wood Johnson Foundation (see also *SI Appendix*, section S8.2).


