Measuring the predictability of life outcomes with a scientific mass collaboration

Matthew J. Salganik[a,1], Ian Lundberg[a], Alexander T. Kindel[a], Caitlin E. Ahearn[b], Khaled Al-Ghoneim[c], Abdullah Almaatouq[d,e], Drew M. Altschul[f], Jennie E. Brand[b,g], Nicole Bohme Carnegie[h], Ryan James Compton[i], Debanjan Datta[j], Thomas Davidson[k], Anna Filippova[l], Connor Gilroy[m], Brian J. Goode[n], Eaman Jahani[o], Ridhi Kashyap[p,q,r], Antje Kirchner[s], Stephen McKay[t], Allison C. Morgan[u], Alex Pentland[e], Kivan Polimis[v], Louis Raes[w], Daniel E. Rigobon[x], Claudia V. Roberts[y], Diana M. Stanescu[z], Yoshihiko Suhara[e], Adaner Usmani[aa], Erik H. Wang[z], Muna Adem[bb], Abdulla Alhajri[cc], Bedoor AlShebli[dd], Redwane Amin[ee], Ryan B. Amos[y], Lisa P. Argyle[ff], Livia Baer-Bositis[gg], Moritz Büchi[hh], Bo-Ryehn Chung[ii], William Eggert[jj], Gregory Faletto[kk], Zhilin Fan[ll], Jeremy Freese[gg], Tejomay Gadgil[mm], Josh Gagné[gg], Yue Gao[nn], Andrew Halpern-Manners[bb], Sonia P. Hashim[y], Sonia Hausen[gg], Guanhua He[oo], Kimberly Higuera[gg], Bernie Hogan[pp], Ilana M. Horwitz[qq], Lisa M. Hummel[gg], Naman Jain[x], Kun Jin[rr], David Jurgens[ss], Patrick Kaminski[bb,tt], Areg Karapetyan[uu,vv], E. H. Kim[gg], Ben Leizman[y], Naijia Liu[z], Malte Möser[y], Andrew E. Mack[z], Mayank Mahajan[y], Noah Mandell[ww], Helge Marahrens[bb], Diana Mercado-Garcia[qq], Viola Mocz[xx], Katariina Mueller-Gastell[gg], Ahmed Musse[yy], Qiankun Niu[ee], William Nowak[zz], Hamidreza Omidvar[aaa], Andrew Or[y], Karen Ouyang[y], Katy M. Pinto[bbb], Ethan Porter[ccc], Kristin E. Porter[ddd], Crystal Qian[y], Tamkinat Rauf[gg], Anahit Sargsyan[eee], Thomas Schaffner[y], Landon Schnabel[gg], Bryan Schonfeld[z], Ben Sender[fff], Jonathan D. Tang[y], Emma Tsurkov[gg], Austin van Loon[gg], Onur Varol[ggg,hhh], Xiafei Wang[iii], Zhi Wang[hhh,jjj], Julia Wang[y], Flora Wang[fff], Samantha Weissman[y], Kirstie Whitaker[kkk,lll], Maria K. Wolters[mmm], Wei Lee Woon[nnn], James Wu[ooo], Catherine Wu[y], Kengran Yang[aaa], Jingwen Yin[ll], Bingyu Zhao[ppp], Chenyun Zhu[ll], Jeanne Brooks-Gunn[qqq,rrr], Barbara E. Engelhardt[y,ii], Moritz Hardt[sss], Dean Knox[z], Karen Levy[ttt], Arvind Narayanan[y], Brandon M. Stewart[a], Duncan J. Watts[uuu,vvv,www], and Sara McLanahan[a,1]

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

How predictable are life trajectories? We investigated this question with 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.

life course | prediction | machine learning | mass collaboration

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

Author contributions: M.J.S., I.L., A.T.K., J.B.-G., B.E.E., M.H., K.L., A.N., B.M.S., D.J.W., and S. McLanahan designed research; M.J.S., I.L., A.T.K., C.E.A., K.A.-G., A. Almaatouq, D.M.A., J.E.B., N.B.C., R.J.C., D.D., T.D., A.F., C.G., B.J.G., E.J., R.K., A. Kirchner, S. McKay, A.C.M., A.P., K.P., L.R., D.E.R., C.V.R., D.M.S., Y.S., A.U., E.H.W., M.A., A. Alhajri, B.A., R.A., R.B.A., L.P.A., L.B.-B., M.B., B.-R.C., W.E., G.F., Z.F., J.F., T.G., J.G., Y.G., A.H.-M., S.P.H., S.H., G.H., K.H., B.H., I.M.H., L.M.H., N.J., K.J., D.J., P.K., A. Karapetyan, E.H.K., B.L., N.L., M. Möser, A.E.M., M. Mahajan, N.M., H.M., D.M.-G., V.M., K.M.-G., A.M., Q.N., W.N., H.O., A.O., K.O., K.M.P., E.P., K.E.P., C.Q., T.R., A.S., T.S., L.S., B. Schonfeld, B. Sender, J.D.T., E.T., A.v.L., O.V., X.W., Z.W., J. Wang, F.W., S.W., K.W., M.K.W., W.L.W., J. Wu, C.W., K.Y., J.Y., B.Z., and C.Z. analyzed data; and M.J.S., I.L., A.T.K., D.K., and S. McLanahan wrote the paper.

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.

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/.

See online for related content such as Commentaries.

[1] To whom correspondence may be addressed. Email: mjs3@princeton.edu or mclanaha@princeton.edu.

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

First published March 30, 2020.

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.

of any particular researcher or method. An additional benefit of the common task method is that the standardization of the prediction task facilitates comparisons between different methodological and theoretical approaches.

Our mass collaboration builds on a long-running, intensive data collection: the Fragile Families and Child Wellbeing Study (hereafter the Fragile Families study). In contrast to government administrative records and digital trace data that are often used for prediction, these data were created to enable social science research. The ongoing study collects rich longitudinal data about thousands of families, each of whom gave birth to a child in a large US city around the year 2000 (9). The study was designed to understand families formed by unmarried parents and the lives of children born into these families.

The Fragile Families data—which have been used in more than 750 published journal articles (10)—were collected in six waves: child birth and ages 1, 3, 5, 9, and 15. Each wave includes a number of different data collection modules. For example, the first wave (birth) includes survey interviews with the mother and father. Over time, the scope of data collection increased, and the fifth wave (age 9 y) includes survey interviews with the mother, the father, the child's primary caregiver (if not the mother or father), the child's teacher, and the child (Fig. 1).

Each data collection module is made up of ∼10 sections, where each section includes questions about a specific topic. For example, the interview with the mother in wave 1 (birth) has sections about the following topics: child health and development, father–mother relationships, fatherhood, marriage attitudes, relationship with extended kin, environmental factors and government programs, health and health behavior, demographic characteristics, education and employment, and income. The interview with the child in wave 5 (age 9 y) has questions about the following topics: parental supervision and relationship, parental discipline, sibling relationships, routines, school, early delinquency, task completion and behavior, and health and safety.

In addition to the surveys, interviewers traveled to the child's home at waves 3, 4, and 5 (ages 3, 5, and 9 y) to conduct an in-home assessment that included psychometric testing (e.g., Peabody Picture Vocabulary Test, Woodcock–Johnson Passage Comprehension Test, etc.), biometric measurements (e.g., height, weight, etc.), and observations of the neighborhood and home. More information about the Fragile Families data are included in *SI Appendix*, section S1.1.

When we began designing the Fragile Families Challenge, data from waves 1 to 5 (birth to age 9 y) were already available to researchers. However, data from wave 6 (age 15 y) were not yet available to researchers outside of the Fragile Families team. This moment where data have been collected but are not yet available to outside researchers—a moment that exists in all longitudinal surveys—creates an opportunity to run a mass collaboration using the common task method. This setting

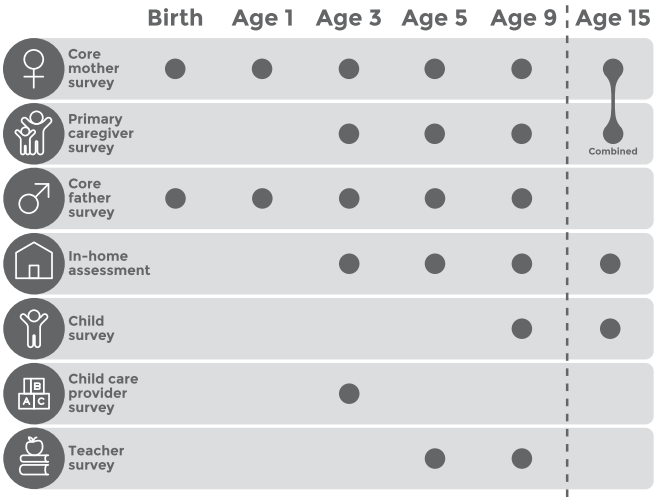

Fig. 1. Data collection modules in the Fragile Families study. Each module is made up of ∼10 sections, where each section includes questions about a specific topic (e.g., marriage attitudes, family characteristics, demographic characteristics). Information about the topics included in each module is 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).

makes it possible to release some cases for building predictive models while withholding others for evaluating the resulting predictions.

Wave 6 (age 15 y) of the Fragile Families study includes 1,617 variables. From these variables, we selected six outcomes to be the focus of the Fragile Families Challenge: 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 outcomes for many reasons, three of which were to include different types of variables (e.g., binary and continuous), to include a variety of substantive topics (e.g., academics, housing, employment), and to include a variety of units of analysis (e.g., child, household, primary caregiver). All outcomes are based on self-reported data. *SI Appendix*, section S1.1 describes how each outcome was measured in the Fragile Families study.

In order to predict these outcomes, participants had access to a background dataset, a version of the wave 1 to 5 (birth to age 9 y) data that we compiled for the Fragile Families Challenge. For privacy reasons, the background data excluded genetic and geographic information (11). The background data included 4,242 families and 12,942 variables about each family. The large number of predictor variables is the result of the intensive and long-term data collection involved in the Fragile Families study. In addition to the background data, participants in the Fragile Families Challenge also had access to training data that included the six outcomes for half of the families (Fig. 2). Similar to other projects using the common task method, the task was to use data collected in waves 1 to 5 (birth to age 9 y) and some data from wave 6 (age 15 y) to build a model that could then be used to predict the wave 6 (age 15 y) outcomes for other families. The prediction task was not to forecast outcomes in wave 6 (age 15 y) using only data collected in waves 1 to 5 (birth to age 9 y), which would be more difficult.

The half of the outcome data that was not available for training was used for evaluation. These data were split into two sets: leaderboard and holdout. During the Fragile Families Challenge, participants could assess their predictive accuracy in the leaderboard set. However, predictive accuracy in the holdout set was unknown to participants—and organizers—until the end

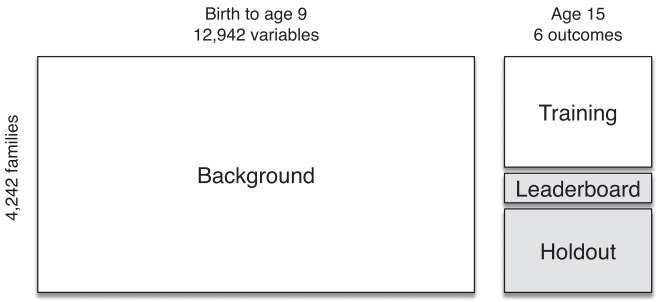

Fig. 2. Datasets in the Fragile Families Challenge. During the Fragile Families Challenge, participants used the background data (measured from child's birth to age 9 y) and the training data (measured at child age 15 y) to predict the holdout data as accurately as possible. While the Fragile Families 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.

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). Participants often worked in teams. We ultimately received valid submissions from 160 teams. Many of these teams used machine-learning methods that are not typically used in social science research and that explicitly seek to maximize predictive accuracy (12, 13).

While the Fragile Families Challenge was underway (March 5, 2017 to August 1, 2017), participants could upload their submissions to the Fragile Families Challenge website. Each submission included predictions, the code that generated those predictions, and a narrative explanation of the approach. After the submission was uploaded, participants could see their score on a leaderboard, which ranked the accuracy of all uploaded predictions in the leaderboard data (14). In order to take part in the mass collaboration, all participants provided informed consent to the procedures of the Fragile Families Challenge, including agreeing to open-source their final submissions (*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_{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).

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

R^2_{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_{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

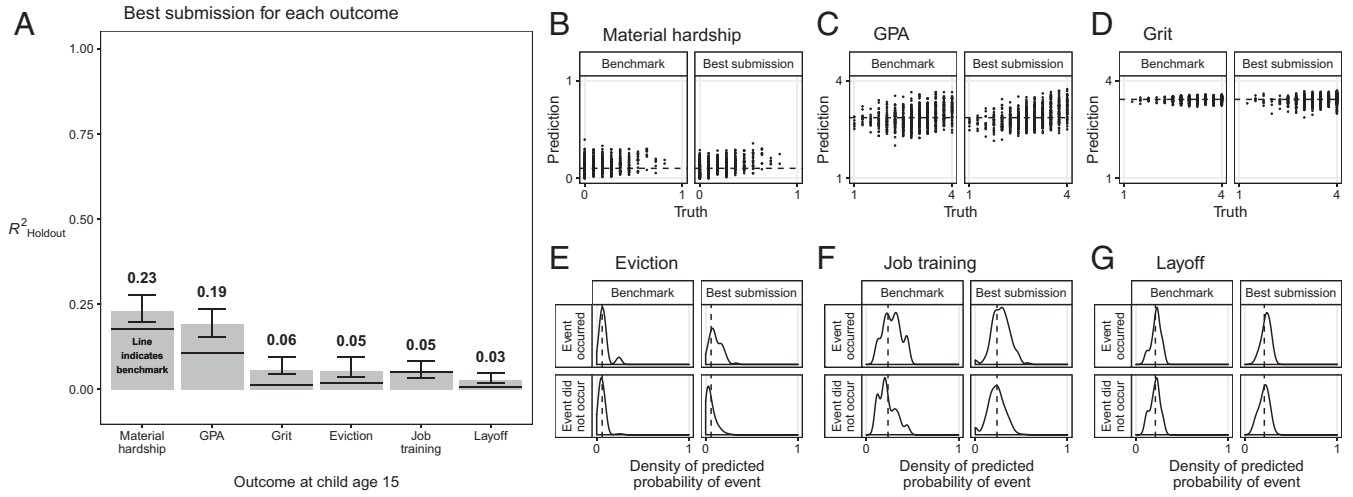

Fig. 3. Performance in the holdout data of the best submissions and a four variable benchmark model (*SI Appendix*, section S2.2). *A* shows the best performance (bars) and a benchmark model (lines). Error bars are 95% confidence intervals (*SI Appendix*, section S2.1). *B–D* compare the predictions 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.

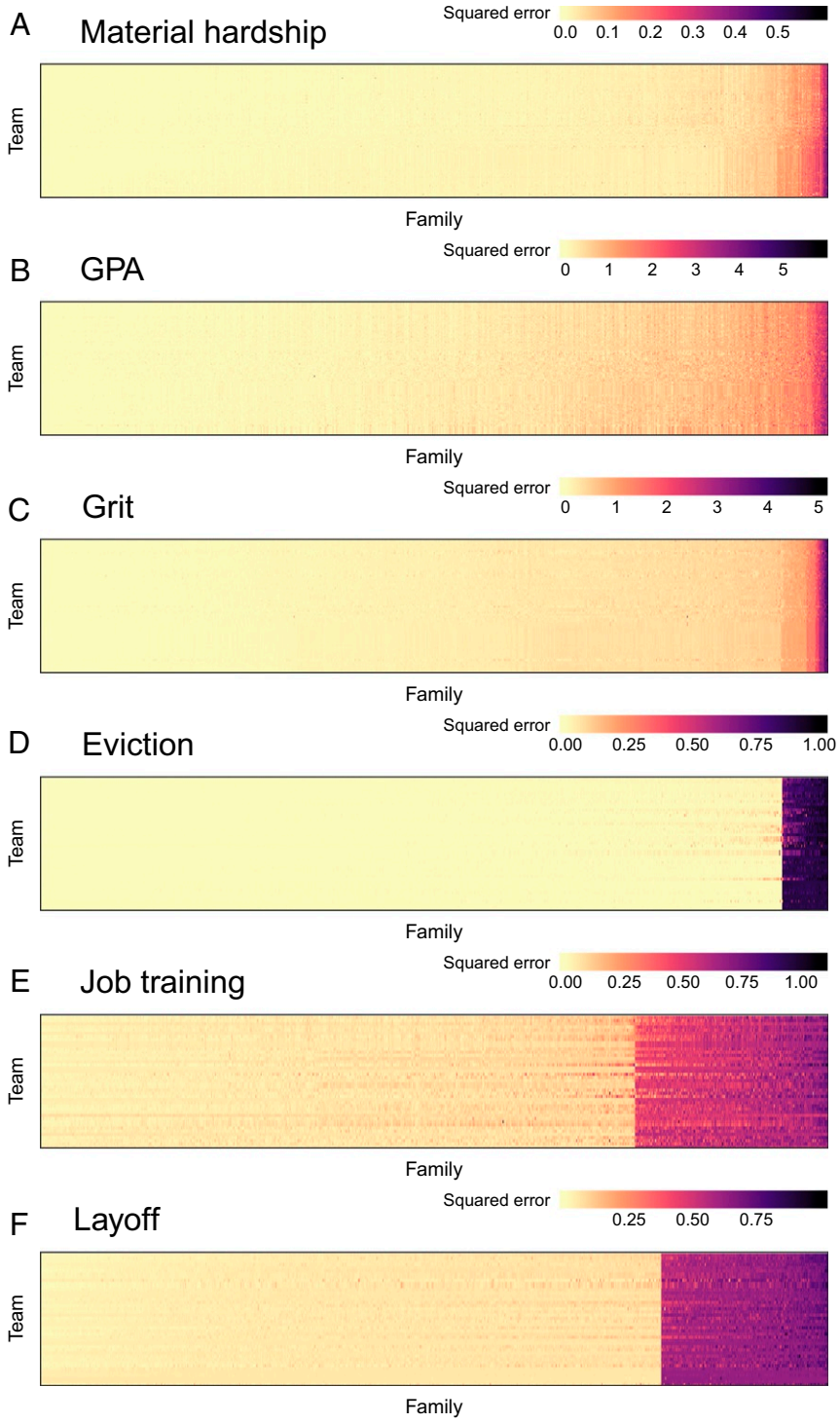

Fig. 4. Heatmaps of the squared prediction error for each observation in the holdout data. Within each heatmap, each row represents a team that made a qualifying submission (sorted by predictive accuracy), and each column represents a family (sorted by predictive difficulty). Darker colors indicate higher 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 Challenge speaks directly to the predictability of life outcomes in only one setting: six specific outcomes, as predicted by a particular set of variables measured by a single study for a particular group of people. Predictability is likely to vary across settings, such as for different outcomes, over different time gaps between the predictors and outcomes, using different 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 (26). 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 P2-CHD047879. Funding for the Fragile Families and Child Wellbeing Study was provided by NICHD Grants R01-HD36916, R01-HD39135, and R01-HD40421; and a consortium of private foundations, including the Robert Wood Johnson Foundation (see also *SI Appendix*, section S8.2).

[a]Department of Sociology, Princeton University, Princeton, NJ 08544; [b]Department of Sociology, University of California, Los Angeles, CA 90095; [c]Hawaz, Riyadh 12363, Saudi Arabia; [d]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02142; [e]Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139; [f]Mental Health Data Science Scotland, Department of Psychology, The University of Edinburgh, Edinburgh EH8 9JZ, United Kingdom; [g]Department of Statistics, University of California, Los Angeles, CA 90095; [h]Department of Mathematical Sciences, Montana State University, Bozeman, MT 59717; [i]Human Computer Interaction Lab, University of California, Santa Cruz, CA 95064; [j]Discovery Analytics Center, Virginia Polytechnic Institute and State University, Arlington, VA 22203; [k]Department of Sociology, Cornell University, Ithaca, NY 14853; [l]GitHub, San Francisco, CA 94107; [m]Department of Sociology, University of Washington, Seattle, WA 98105; [n]Social and Decision Analytics Laboratory, Fralin Life Sciences Institute, Virginia Polytechnic Institute and State University, Arlington, VA 22203; [o]Institute for Data, Systems and Society, Massachusetts Institute of Technology, Cambridge, MA 02139; [p]Department of Sociology, University of Oxford, Oxford OX1 1JD, United Kingdom; [q]Nuffield College, University of Oxford, Oxford OX1 1NF, United Kingdom; [r]School of Anthropology and Museum Ethnography, University of Oxford, Oxford OX2 6PE, United Kingdom; [s]Program for Research in Survey Methodology, Survey Research Division, RTI International, Research Triangle Park, NC 27709; [t]School of Social and Political Sciences, University of Lincoln, Brayford Pool, Lincoln LN6 7TS, United Kingdom; [u]Department of Computer Science, University of Colorado, Boulder, CO 80309; [v]Center for the Study of Demography and Ecology, University of Washington, Seattle, WA 98105; [w]Department of Economics, Tilburg School of Economics and Management, Tilburg University, 5037 AB Tilburg, The Netherlands; [x]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544; [y]Department of Computer Science, Princeton University, Princeton, NJ 08544; [z]Department of Politics, Princeton University,

Princeton, NJ, 08544; [aa]Department of Sociology, Harvard University, Cambridge, MA 02138; [bb]Department of Sociology, Indiana University, Bloomington, IN 47405; [cc]Department of Nuclear Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; [dd]Computational Social Science Lab, Social Science Division, New York University Abu Dhabi, 129188 Abu Dhabi, United Arab Emirates; [ee]Bendheim Center for Finance, Princeton University, Princeton, NJ 08544; [ff]Department of Political Science, Brigham Young University, Provo, UT 84602; [gg]Department of Sociology, Stanford University, Stanford, CA 94305; [hh]Department of Communication and Media Research, University of Zurich, Zurich, Switzerland, ZH-8050; [ii]Center for Statistics & Machine Learning, Princeton University, Princeton, NJ 08544; [jj]Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544; [kk]Statistics Group, Department of Data Sciences and Operations, Marshall School of Business, University of Southern California, Los Angeles, CA 90089; [ll]Department of Statistics, Columbia University, New York, NY 10027; [mm]Center for Data Science, New York University, New York, NY 10011; [nn]Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027; [oo]Department of Molecular Biology, Princeton University, Princeton, NJ 08544; [pp]Oxford Internet Institute, University of Oxford, Oxford OX1 3JS, United Kingdom; [qq]Graduate School of Education, Stanford University, Stanford, CA, 94305; [rr]Department of Computer Science, Ohio State University, Columbus, OH 43210; [ss]School of Information, University of Michigan, Ann Arbor, MI 48104; [tt]Center for Complex Networks and Systems Research, Indiana University, Bloomington, IN 47405; [uu]Department of Computer Science, Masdar Institute, Khalifa University, 127788 Abu Dhabi, United Arab Emirates; [vv]Research Institute for Mathematical Sciences, Kyoto University, Kyoto 606-8502, Japan; [ww]Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544; [xx]Department of Neuroscience, Princeton University, Princeton, NJ 08544; [yy]Department of Electrical Engineering, Princeton University, Princeton, NJ, 08544; [zz]Dataiku, New York, NY 10010; [aaa]Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ 08544; [bbb]Department of Sociology, California State University, Dominguez Hills, Carson, CA 90747; [ccc]School of Media and Public Affairs, George Washington University, Washington, DC 20052; [ddd]Center for Data Insights, MDRC, Oakland, CA 94612; [eee]Social Science Division, New York University Abu Dhabi, 129188 Abu Dhabi, United Arab Emirates; [fff]Department of Economics, Princeton University, Princeton, NJ 08544; [ggg]Center for Complex Network Research, Northeastern University Networks Science Institute, Boston, MA 02115; [hhh]Luddy School of Informatics, Computing, & Engineering, Indiana University, Bloomington, IN 47408; [iii]School of Social Work, David B. Falk College of Sport and Human Dynamics, Syracuse University, NY 13244; [jjj]School of Public Health, Indiana University, Bloomington, IN 47408; [kkk]The Alan Turing Institute, London NW1 2DB, United Kingdom; [lll]Department of Psychiatry, University of Cambridge, Cambridge CB2 0SZ, United Kingdom; [mmm]School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, United Kingdom; [nnn]Department of Marketplaces & Yield Data Science, Expedia Group, Seattle, WA 98119; [ooo]Department of the Applied Statistics, Social Science, and Humanities, New York University, New York, NY 10003; [ppp]Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, United Kingdom; [qqq]Department of Human Development, Teachers College, Columbia University, New York, NY 10027; [rrr]Department of Pediatrics, Vagelos College of Physicians and Surgeons, Columbia University, New York, NY 10032; [sss]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720; [ttt]Department of Information Science, Cornell University, Ithaca, NY 14853; [uuu]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104; [vvv]Annenberg School of Communication, University of Pennsylvania, Philadelphia, PA 19104; and [www]Operations, Information and Decisions Department, University of Pennsylvania, Philadelphia, PA 19104

1. J. Kleinberg, J. Ludwig, S. Mullainathan, Z. Obermeyer, Prediction policy problems. *Am. Econ. Rev.* **105**, 491–495 (2015).
2. S. Athey, Beyond prediction: Using big data for policy problems. *Science* **355**, 483–485 (2017).
3. P. M. Blau, O. D. Duncan, *The American Occupational Structure* (John Wiley and Sons, 1967).
4. R. Chetty, N. Hendren, P. Kline, E. Saez, Where is the land of opportunity? The geography of intergenerational mobility in the United States. *Q. J. Econ.* **129**, 1553–1623 (2014).
5. A. Feuerverger, Y. He, S. Khatri, Statistical significance of the Netflix challenge. *Stat. Sci.*, 202–231 (2012).
6. D. Donoho, 50 years of data science. *J. Comput. Graph Stat.* **26**, 745–766 (2017).
7. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Science & Business Media, 2009).
8. J. M. Hofman, A. Sharma, D. J. Watts, Prediction and explanation in social systems. *Science* **355**, 486–488 (2017).
9. N. E. Reichman, J. O. Teitler, I. Garfinkel, S. S. McLanahan, Fragile Families: Sample and design. *Child. Youth Serv. Rev.* **23**, 303–326 (2001).
10. Fragile Families & Child Wellbeing Study, Fragile Families publication search. https://ffpubs.princeton.edu/. Accessed 29 February 2020.
11. I. Lundberg, A. Narayanan, K. Levy, M. J. Salganik, Privacy, ethics, and data access: A case study of the Fragile Families Challenge. arXiv:1809.00103 (1 September 2018).
12. S. Mullainathan, J. Spiess, Machine learning: An applied econometric approach. *J. Econ. Perspect.* **31**, 87–106 (2017).
13. M. Molina, F. Garip, Machine learning for sociology. *Annu. Rev. Sociol.* **45**, 27–45 (2019).
14. A. Blum, M. Hardt, "The ladder: A reliable leaderboard for machine learning competitions" in *Proceedings of the 32nd International Conference on Machine Learning* F. Bach, D. Blei, Eds. (Proceedings of Machine Learning Research, 2015), Vol. 37, pp. 1006–1014.
15. D. J. Watts, Common sense and sociological explanations. *Am. J. Sociol.* **120**, 313–351 (2014).
16. M. Oliver, T. Shapiro, *Black Wealth/White Wealth: A New Perspective on Racial Inequality* (Routledge, 2013).
17. R. B. Alley, K. A. Emanuel, F. Zhang, Advances in weather prediction. *Science* **363**, 342–344 (2019).
18. B. G. Malkiel, The efficient market hypothesis and its critics. *J. Econ. Perspect.* **17**, 59–82 (2003).
19. M. J. Salganik, P. S. Dodds, D. J. Watts, Experimental study of inequality and unpredictability in an artificial cultural market. *Science* **311**, 854–856 (2006).
20. T. Martin, J. M. Hofman, A. Sharma, A. Anderson, D. J. Watts, "Exploring limits to prediction in complex social systems" in *Proceedings of the 25th International Conference on World Wide Web* (International World Wide Web Conferences Steering Committee, 2016), pp. 683–694.
21. L. E. Cederman, N. B. Weidmann, Predicting armed conflict: Time to adjust our expectations? *Science* **355**, 474–476 (2017).
22. J. Risi, A. Sharma, R. Shah, M. Connelly, D. J. Watts, Predicting history. *Nat. Hum. Behav.* **3**, 906–912 (2019).
23. J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, Human decisions and machine predictions. *Q. J. Econ.* **133**, 237–293 (2017).
24. A. Chouldechova, D. Benavides-Prado, O. Fialko, R. Vaithianathan, "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions" in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, S. A. Friedler, C. Wilson, Eds. (Proceedings of Machine Learning Research, 2018), vol. 81, pp. 134–148.
25. S. Barocas, A. D. Selbst, Big data's disparate impact. *Calif. Law Rev.* **104**, 671–732 (2016).
26. C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
27. D. J. Hand, Classifier technology and the illusion of progress. *Stat. Sci.* **21**, 1–14 (2006).
28. R. M. Dawes, D. Faust, P. E. Meehl, Clinical versus actuarial judgment. *Science* **243**, 1668–1674 (1989).
29. M. Salganik, *Bit by Bit: Social Research in the Digital Age* (Princeton University Press, 2018).
30. OS Collaboration, Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
31. P. E. Tetlock, B. A. Mellers, J. P. Scoblic, Bringing probability judgments into policy debates via forecasting tournaments. *Science* **355**, 481–483 (2017).