RESEARCH ARTICLE

NEUROSCIENCE

Illusory generalizability of clinical prediction models

Adam M. Chekroud[1,2]*, Matt Hawrilenko[1], Hieronimus Loho[2], Julia Bondar[1], Ralitza Gueorguieva[3], Alkomiet Hasan[4], Joseph Kambeitz[5], Philip R. Corlett[2], Nikolaos Koutsouleris[6], Harlan M. Krumholz[7], John H. Krystal[2], Martin Paulus[8]

It is widely hoped that statistical models can improve decision-making related to medical treatments. Because of the cost and scarcity of medical outcomes data, this hope is typically based on investigators observing a model's success in one or two datasets or clinical contexts. We scrutinized this optimism by examining how well a machine learning model performed across several independent clinical trials of antipsychotic medication for schizophrenia. Models predicted patient outcomes with high accuracy within the trial in which the model was developed but performed no better than chance when applied out-of-sample. Pooling data across trials to predict outcomes in the trial left out did not improve predictions. These results suggest that models predicting treatment outcomes in schizophrenia are highly context-dependent and may have limited generalizability.

One fundamental problem in medicine is that despite similar treatments some patients get better whereas others show no improvement. One goal of precision medicine is to use machine learning to find models that will help predict who will respond to what type of treatment (*1*). For precision medicine to affect clinical practice and improve outcomes, the models that we develop must robustly predict outcomes for unseen, future patients (*2*–*5*).

However, models are not usually tested on new patients in a different context because data—especially data from controlled designs—are scarce and expensive (*6*). Instead, researchers typically split a study's participants into two or more random groups, build a model using the data from one of the groups, and test its predictions on the other group (e.g., k-fold cross-validation) (*3*, *4*). When we use this kind of approximation based on one data set or clinical sample, we have a fundamentally limited insight into the true potential for a model to improve outcomes in the future. Validating clinical prediction models in different clinical samples is an essential step in the model development process. It generally results in predictive performance measures that are lower but allows for a more realistic assessment of the potential for statistical models to improve clinical practice (*7*–*9*).

Open data opens possibilities

As efforts toward mandatory randomized controlled trial (RCT) data deposition, archival data sharing, and open science continue to advance, opportunities arise to more rigorously examine how well treatment prediction models will fare in different contexts. The Yale Open Data Access (YODA) Project is one such effort, which now includes a data archive of over 246 clinical trials from all medical fields.

The YODA project included several RCTs evaluating the comparative efficacy of antipsychotic medications for treating schizophrenia. Predicting treatment outcomes in schizophrenia could be especially advantageous because the clinical response to pharmacological interventions is heterogeneous and depends on many environmental factors such as individual and family-related stress, drug abuse, homelessness, and social isolation. Depending on the clinical outcome definition, up to 20 to 30% of first-episode individuals (*10*) and more than 50% with a relapse do not respond sufficiently to antipsychotic medications (*11*).

We examined the generalizability of clinical prediction models across multiple clinical trials using the case study of antipsychotic treatments for schizophrenia. Critically, this study directly evaluated the performance of a model on its initial training sample as well as how the same model performed on truly independent clinical trial samples. This allowed us to assess two key risks: First, models may "overfit" the data by fitting the random noise of one particular dataset rather than a true signal likely to generalize across samples, leading to good predictions in the training data that do not generalize to the testing data. The second key risk is poor model transportability. Models may lack external validity due to patients, providers, or implementation characteristics varying across trials (*12*).

Data sources

We used treatment data from five international, multisite RCTs (NCT00518323, NCT00334126, NCT00085748, NCT00078039, and NCT00083668) obtained through the YODA Project (https://yoda.yale.edu/). These trials were selected because of their comparability and consistency. All patients had a current DSM-IV diagnosis of schizophrenia at the start of the trial; all trials randomized patients to an antipsychotic medication or placebo; all trials used the same scale to measure treatment outcomes (the Positive and Negative Syndrome Scale, PANSS); all trials included a 4-week timepoint to measure outcomes; and all trials collected similar data about the patients at baseline. Combined, the trials also provide a heterogeneous patient

Table 1. Treatment outcomes across trials.

Outcome definition	Adults first episode (*n* = 321)	Adults - Chronic #1 (*n* = 430)	Adults - Chronic #2 (*n* = 481)	Older adults (*n* = 99)	Teens (*n* = 182)	Total (*n* = 1513)
25% Reduction PANSS	264 (82.2%)	208 (48.4%)	266 (55.3%)	32 (32.3%)	47 (25.8%)	816 (54.0%)
50% Reduction PANSS	119 (37.1%)	85 (19.8%)	82 (17.0%)	7 (7.1%)	12 (6.6%)	306 (20.3%)
RSWG remission criteria	152 (47.4%)	129 (30.0%)	153 (31.8%)	24 (24.2%)	58 (31.9%)	517 (34.2%)
Percentage change in PANSS total score (SD)	-44.1 (23.1)	-26.9 (28.2)	-28.4 (25.3)	-18.0 (21.8)	-13.7 (21.5)	-28.8 (26.7)
Baseline total PANSS (SD)	103.0 (14.3)	92.4 (13.0)	92.9 (10.9)	91.1 (8.8)	90.0 (13.1)	94.4 (13.2)

[1]Spring Health, New York City, NY 10010, USA. [2]Department of Psychiatry, Yale University School of Medicine, New Haven, CT 06520, USA. [3]Department of Biostatistics, Yale University, New Haven, CT 06520, USA. [4]Department of Psychiatry, Psychotherapy and Psychosomatics, University Augsburg, 86159 Augsburg, Germany. [5]Department of Psychiatry and Psychotherapy, University of Cologne, Faculty of Medicine and University Hospital of Cologne, Cologne, Germany. [6]Department of Psychiatry and Psychotherapy, Ludwig-Maximilians-University, Munich, Germany. [7]Center for Outcomes Research and Evaluation, Yale New Haven Hospital, New Haven, CT 06520, USA. [8]Laureate Institute for Brain Research, Tulsa, OK 74136, USA.
*Corresponding author. Email: adam.chekroud@yale.edu

population, with patients recruited from 194 sites across 4 continents, a pediatric trial, an older adult trial, and a trial of individuals with a first episode (see SM for more details). The study design, outcome measure, and cross-validation approach were preregistered on 2 August 2016 (YODA 2016-1005). Minor updates to the preregistration were submitted on 2 May 2023 (included in the SM).

Patients and outcomes

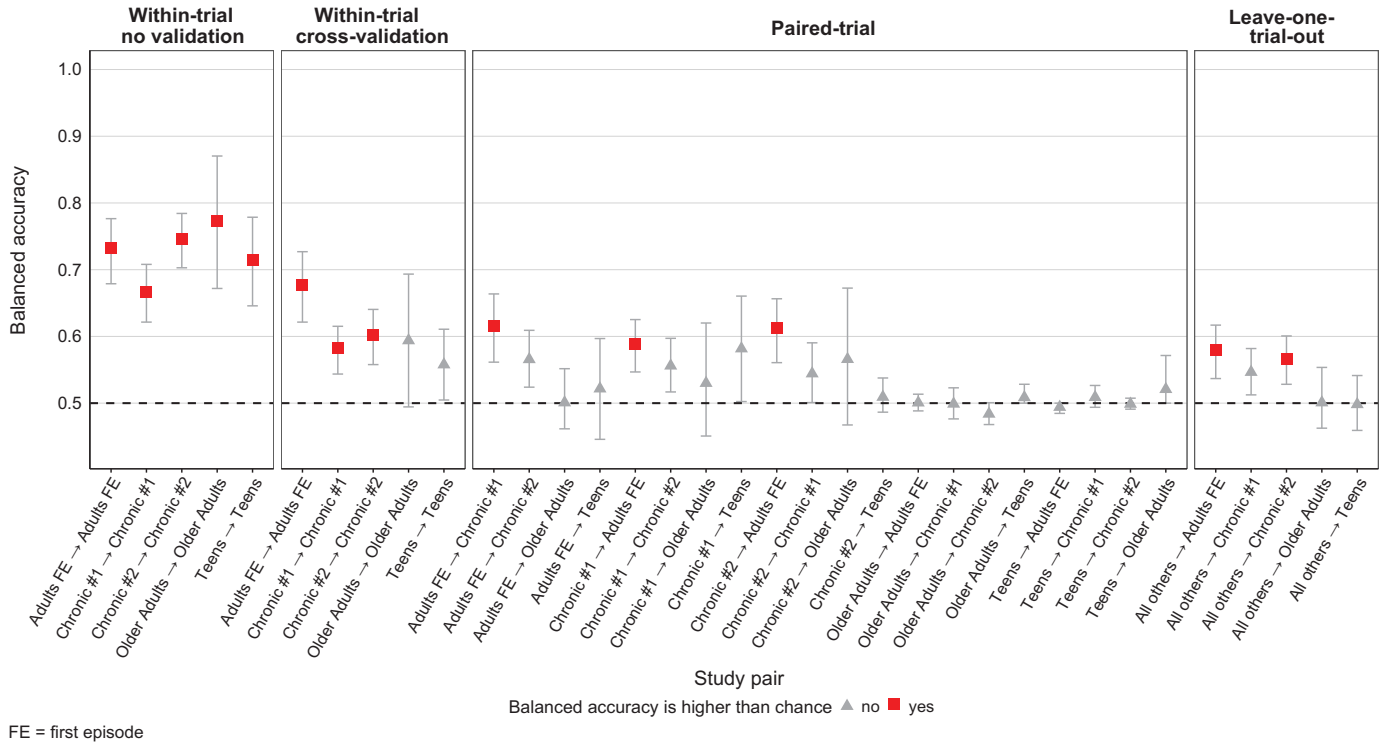
From 29 March 2004 to 30 March 2009, 1962 total patients aged 12 to 81 years were enrolled across five randomized controlled trials at 194 sites in North America, Asia, Europe, and Africa. We assessed symptomatic outcomes based on the PANSS (*13*) at week 4 for the 1513 participants with baseline and 4-week follow-up data. Different definitions of response, remission, and recovery are used in schizophrenia research, which makes comparing and applying results in clinical practice difficult (*14*–*16*). The primary outcome reported here is the Remission in Schizophrenia Working Group criteria (RSWG) (*17*). To ensure that our findings were not driven by idiosyncrasies in how we defined treatment response, we included three other definitions commonly used in the field, including percentage change with baseline correction (*15*, *16*), and two binary definitions of 25 and 50% symptom reduction. Table 1 reports treatment outcomes for all definitions across the five trials.

We extracted all information available at baseline across all trials and retained it as a predictor variable if it was available for more than 80% of patients. We also computed condition (control versus treatment) X predictor interaction terms. Drug dose was standardized to paliperidone dose equivalents using the defined daily dose method (*18*). Together, this yielded 217 predictor variables that included basic demographic features, psychiatric history (DSM-IV diagnosis category, age of diagnosis, psychiatric hospitalizations), clinical data (PANSS, Clinical Global Impression) (*17*), extrapyramidal symptom scales (Abnormal Involuntary Movement Scale) (*19*) and Simpson Angus Scale (*20*), biometric data (blood chemistry panel, hematology, urinalysis), and treatment randomization. The detailed list of predictors, selection criteria, and missing data approach is provided in the SM.

Machine learning approach

We applied machine learning methods using baseline data to predict whether a patient would achieve clinically significant improvements in symptoms over four weeks of antipsychotic treatment. We used the elastic net algorithm (*21*, *22*), a penalized regression method that is appropriate when covariates are correlated with one another and predictors may only be sparsely endorsed. It has been successful in research predicting psychiatric treatment outcomes (*5*, *23*–*25*).

The elastic net model uses two penalty parameters, lambda and alpha, which balance stability with parsimony. We examined 400 combinations of alpha and lambda penalties (see supplement) and selected the optimal penalties using repeated 10-fold cross-validation. The cross-validation part of this procedure separates the data set into 10 random folds and uses 9 of the subsets for training, repeating the process such that each subset is left out once for testing. The repeated part of this procedure re-splits the data ten times to reduce the impact of the random data split; in aggregate, 100 total models were fit to the 10 folds by 10 repeats. Model performance was calculated by averaging the performance metric across all 100 models. This entire procedure was run for each of the 400 combinations of alpha and lambda values, and the final values were chosen as the combination of alpha and lambda values that optimized the model performance metric. We used the metrics of area under the receiver operating curve for binary outcomes and root mean square error for continuous outcomes. The final alpha and lambda values were applied to the aggregate sample to estimate the prediction



FE = first episode

Fig. 1. Balanced accuracy for models predicting treatment outcome (Remission in Schizophrenia Working Group criteria) across all modeling scenarios. Gray intervals represent 95% confidence intervals, not adjusted for multiple comparisons. Red markers denote statistical significance after applying the Benjamini-Hochberg adjustment with the false discovery rate set to 5%. Repeated 10-fold cross-validation; FE, first episode.

model coefficients. To interpret differential performance across samples, we report a metric known as balanced accuracy [(sensitivity + specificity) / 2] whose null distribution is centered on 50% (*26*, *27*). To determine whether balanced accuracy was statistically significantly above chance, we bootstrapped confidence intervals and adjusted for multiple comparisons across all 35 comparisons using the Benjamini-Hochberg adjustment with the false discovery rate set to 5% (*28*). All analysis was conducted using R version 4.1 (*29*), with machine learning models fit using the caret package (*30*).

Exploring the generalizability of machine learning models

We evaluated the applicability of machine learning models across four distinct scenarios to gain insights into their generalizability: First, we assessed the predictive accuracy of the model within the trial, without any external validation beyond the training data. Second, we also focused on within-trial prediction accuracy but this time estimated using the data excluded from the training set in a repeated tenfold cross-validation process. Third, we conducted a paired-across-trial prediction accuracy assessment. In this case, models trained on one trial were applied to all other trials to evaluate their performance. Finally, in the fourth scenario, we implemented a leave-one-trial-out prediction accuracy assessment. Models were trained using data aggregated from four trials and their predictive accuracy was tested on the fifth trial (Fig. 1). Balanced accuracy for the RSWG criteria are shown in Fig. 1, and data for alternative outcome definitions and additional outcome metrics are shown in the supplement.

No validation

In the scenario where we assessed within-trial performance without any external validation, the final prediction model created for a specific trial was applied to the entire sample from that same trial. The balanced accuracy was high and significantly above random chance for all models, with an average of 0.72 (range: 0.66 to 0.77) across all five prediction models. However, because the model was evaluated on the same sample used to develop it, there is a risk of overfitting, making these results less likely to generalize.

Cross-validation

To estimate more generalizable prediction accuracy, we employed within-trial cross-validation. Performance characteristics of the optimal alpha and lambda values were averaged across the 100 left out folds (10 folds * 10 repeats) from the repeated cross-validation procedure. Each trial's data were divided into 10 subsets, with coefficients developed on 9 subsets and then tested on the remaining subset. In this scenario, balanced accuracy was lower in each dataset, averaging 0.60 (range: 0.56 to 0.67) across all five prediction models. Only three out of five models performed above chance.

Paired-trial validation

Next, we directly assessed out-of-sample performance in the paired-trial validation (*16*). We applied the prediction models developed using within-trial models across each of the other trials, for a total of 20 trial pairs. Model performance was low (mean across all trial pairs was 0.54, range 0.48 to 0.61) with only three trial pairs performing above chance.

Leave-one-trial-out validation

Given the availability of multiple archival trials for developing a prediction model, a natural extension of the paired-trial validation would be a leave-one-trial-out approach. This approach might enhance generalizability by allowing the algorithm to be exposed to more information through between-trial variability in baseline phenotypes. We aggregated data across four trials, leaving the fifth out for testing, and repeated the process 5 times so that each trial was left out once. Performance was once again poor with low balanced accuracy in all conditions (mean across all left out trials was 0.54 with range 0.50 to 0.58) and performance was significantly above chance in only two of the five testing sets.

Sensitivity analyses

The pattern of results observed was not due to idiosyncrasies of how we measured treatment response. We found the same pattern of results when we reproduced all four modeling scenarios using other binary and continuous definitions of treatment response (see SM).

This lack of model generalizability to unseen patients was also observed for another machine learning algorithm. When we applied random forest models, which can detect complex patterns of interactions amongst predictor variables, we observed the same pattern of results except that excessive overfitting occurred for no-validation conditions (see SM).

Discussion

Machine learning prediction of treatment outcomes in medicine is exciting but challenging. Our modeling scenarios using antipsychotic treatment outcome prediction in schizophrenia suggest that predictive models are fragile and that excellent performance in one clinical context is not a strong indicator of performance on future patients. This is highly concerning as most predictive studies today rely on internal samples for testing and validation. When models were tested on the same sample on which they were developed, models routinely produced strong predictions. Cross-validation tempered these performance estimates but even the models that performed well in cross-validation were little better than chance when predicting outside of the sample in which they were developed—even when the unseen samples were well-phenotyped. In a world where we hope that predictive models might eventually improve clinical practice, the ability to generalize to other carefully controlled clinical contexts is only the first step to generalize to settings with more heterogeneity in patient presentations and methods of care delivery.

Why model generalizability is challenging

There are three key reasons why predictive models might not generalize across trials. First, patient groups may be too different across trials. The umbrella category of schizophrenia is useful for clinical practice but also means that patients with different disease stages are coerced into the same diagnostic category in clinical trials. If key information that differentiates patients is not captured in the data or if the range of that information is more restricted in the dataset used to develop the model compared with the target trial, predictions may be inaccurate. Thus, patient populations may differ considerably between trials within the same diagnostic category. However, the current study found little evidence that results would generalize across even the most similar trials. The three cross-trial pairs with predictions slightly greater than chance were amongst the three studies of adults aged 18 and over but this pattern of results did not consistently replicate across other outcome definitions.

Second, these trials may not have collected the type or volume of data needed to make good predictions. This study used clinical, sociodemographic, and simple biomarker data based on almost 2000 patients. However, additional data types may have been more relevant to treatment outcomes. Psychosocial information and social determinants of health were not included in this study but have previously been found to predict treatment outcomes in first episode psychosis (*27*). Preliminary research suggests that longitudinal patterns of symptom co-occurrence—either before or during treatment—can be specifically relevant to how a patient will respond to treatment although it may delay care to collect this data (*31*–*34*). Some have suggested the use of neuroimaging and genetic data but there is currently little evidence to suggest that such data would improve predictions; further, collecting these data would pose additional barriers for routine implementation (*35*–*37*). Finally, having data from more participants may allow for more nuanced modeling of individual differences.

A third reason why predictive models may not generalize is that patient outcomes may be

too context-dependent. Trials may have subtly important differences in recruiting procedures, inclusion criteria, or treatment protocols. Because these characteristics do not vary across patients within a trial, they cannot be modeled as predictors within a single trial. However, this study used multinational RCTs conducted by large pharmaceutical companies and contract research organizations, minimizing nonspecific concerns especially in comparison to the variability we would expect in real clinical practice going from one site or provider to the next. Of course, different antipsychotic drugs may differ from one another in ways that affect outcome prediction, and the D2 dopamine receptor blockade intended to correct overstimulation of D2 receptors by endogenous dopamine may be too far downstream from the primary pathology of schizophrenia or the symptom severity criteria used to measure it (*38*).

Improving model generalizability

It is worth considering how we might improve the situation in the future. From a statistical modeling perspective, capturing important heterogeneity through phenotyping or stratification procedures might help improve the generalizability of models. Identifying trial-level characteristics that relate to patient outcomes may provide information to better equip prediction models to generalize across settings. Such trial-level variation can be studied using Bayesian approaches or recent techniques that incorporate replicability across contexts or populations into the algorithm training process (*39*). From a population perspective, there may be some patients for whom the choice of treatment has no impact on their clinical course, which represents an inherent limitation of predicting treatment outcomes. However, this could also be an opportunity for further improvement in identifying which patients have a wider range of potential outcomes and for whom selecting the optimal treatment would provide clinical benefit (*40*).

Longitudinal validation methods, in which a validation sample is drawn from the same population at a later point in time, may provide a limited but pragmatic path to avoid generalizing from one clinical setting to another. The growth of large mental health care delivery systems provides the opportunity to collect large amounts of data and deploy prediction models in the same setting in which they were developed (*41*). This strategy can reduce challenges associated with patient heterogeneity and context-dependence, and also help identify temporal or geographic trends that affect a model's predictions. However, when a model is trained and validated on samples from the same population, it may perform well in that specific context but fail when applied to a different population with different characteristics.

Conclusions

The present study offers an underwhelming but realistic picture of our current ability to develop truly useful predictive models for schizophrenia treatment outcomes. Models that performed with excellent accuracy in one sample routinely failed to generalize to unseen patients. These findings suggest that approximations based on a single data set are a fundamentally limited insight into future performance and represent a potential concern for prediction models throughout medicine. The field as a whole—present authors included—hope that machine learning approaches can eventually improve the allocation of treatments in medicine; however, we should a priori remain skeptical of any predictive model findings that lack an independent sample for validation.

REFERENCES AND NOTES

1. F. S. Collins, H. Varmus, *N. Engl. J. Med.* **372**, 793–795 (2015).
2. D. G. Altman, P. Royston, *Stat. Med.* **19**, 453–473 (2000).
3. D. G. Altman, Y. Vergouwe, P. Royston, K. G. M. Moons, *BMJ* **338**, 1432 (2009).
4. E. W. Steyerberg, in *Clinical Prediction Models* (Springer, 2009), pp. 11–31.
5. A. M. Chekroud *et al*., *Lancet Psychiatry* **3**, 243–250 (2016).
6. G. C. M. Siontis, I. Tzoulaki, P. J. Castaldi, J. P. A. Ioannidis, *J. Clin. Epidemiol.* **68**, 25–34 (2015).
7. E. W. Steyerberg, Y. Vergouwe, *Eur. Heart J.* **35**, 1925–1931 (2014).
8. R. D. Riley *et al*., *Stat. Med.* **40**, 4230–4251 (2021).
9. M. Pavlou *et al*., *Stat. Methods Med. Res.* **30**, 2187–2206 (2021).
10. Y. Zhu *et al*., *Eur. Neuropsychopharmacol.* **27**, 835–844 (2017).
11. S. Leucht *et al*., *Am. J. Psychiatry* **174**, 927–942 (2017).
12. B. Li, C. Gatsonis, I. J. Dahabreh, J. A. Steingrimsson, *Biometrics* **79**, 2382–2393 (2023).
13. S. R. Kay, A. Fiszbein, L. A. Opler, *Schizophr. Bull.* **13**, 261–276 (1987).
14. M. Lambert, A. Karow, S. Leucht, B. G. Schimmelmann, D. Naber, *Dialogues Clin. Neurosci.* **12**, 393–407 (2010).
15. S. Leucht, *J. Clin. Psychiatry* **75**, 8–14 (2014).
16. M. Obermeier *et al*., *BMC Psychiatry* **11**, 113 (2011).
17. J. Busner, S. D. Targum, *Psychiatry* **4**, 28–37 (2007).
18. S. Leucht, M. Samara, S. Heres, J. M. Davis, *Schizophr. Bull.* **42**, S90–S94 (2016).
19. W. Guy, *Ecdeu Assessment Manual for Psychopharmacology* (The George Washington University, 1976).
20. G. M. Simpson, J. W. Angus, *Acta Psychiatr Scand Suppl.* **45**, 11–19 (1970).
21. H. Zou, T. Hastie, *J. R. Stat. Soc. Series B Stat. Methodol.* **67**, 301–320 (2005).
22. J. Friedman, T. Hastie, R. Tibshirani, *J. Stat. Softw.* **33**, 1–22 (2010).
23. A. M. Chekroud *et al*., *Psychiatr. Serv.* **69**, 927–934 (2018).
24. A. M. Chekroud *et al*., *JAMA Psychiatry* **74**, 370–378 (2017).
25. Z. D. Cohen, R. J. DeRubeis, *Annu Rev Clin Psychol.* **14**, 209–236 (2018).
26. K. H. Brodersen, C. S. Ong, K. E. Stephan, J. M. Buhmann, in *2010 20th International Conference on Pattern Recognition* (2010), pp. 3121–3124.
27. N. Koutsouleris *et al*., *Lancet Psychiatry* **3**, 935–946 (2016).
28. Y. Benjamini, Y. Hochberg, *J. R. Stat. Soc. B* **57**, 289–300 (1995).
29. R Core Team, R: A language and environment for statistical computing (R Foundation for Statistical Computing, 2021); www.R-project.org/.
30. M. Kuhn, *J. Stat. Softw.* **28**, 1 (2008).
31. A. J. Fisher, J. D. Medaglia, B. F. Jeronimus, *Proc. Natl. Acad. Sci. U.S.A.* **115**, E6106–E6115 (2018).
32. A. J. Fisher, P. Soyster, Generating Accurate Personalized Predictions of Future Behavior: A Smoking Exemplar. PsyArXiv e24v6 [Preprint] (2019); doi:10.31234/osf.io/e24v6
33. A. J. Fisher, J. F. Boswell, *Assessment* **23**, 496–506 (2016).
34. A. J. Fisher *et al*., *Behav. Res. Ther.* **116**, 69–79 (2019).
35. A. M. Chekroud, N. Koutsouleris, *Mol. Psychiatry* **23**, 24–25 (2018).
36. A. M. Chekroud, *JAMA Psychiatry* **74**, 1183–1184 (2017).
37. M. P. Paulus, *JAMA Psychiatry* **74**, 1185–1186 (2017).
38. R. A. McCutcheon *et al*., *Biol. Psychiatry* **94**, 561–568 (2023).
39. P. Patil, G. Parmigiani, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2578–2583 (2018).
40. R. J. DeRubeis *et al*., *PLOS ONE* **9**, e83875 (2014).
41. J. Bondar *et al*., *JAMA Netw. Open* **5**, e2216349 (2022).

ACKNOWLEDGMENTS

Funding: No funding source had any role in the study design, data collection, data analysis, data interpretation, writing, or submission of this report. All trials were originally funded by Janssen Research and Development. The study design, outcome measure, and cross-validation approach were preregistered on 1 Aug 2016 (YODA #2016-1005). The study was approved on 15 December 2016 (IRB/HSC# 1610018521) by the Yale University Institutional Review Board. Access was granted on 5 January 2017, and data were analyzed until October, 2023. **Author contributions:** Conceptualization: A.C. Methodology: A.C., M.H., R.G., H.L., J.B., A.H., N.K., J.H.K., H.K. Data Acquisition: H.K., on behalf of the Yale Open Data Access Initiative. Data Analysis: A.C., M.H., H.L., J.B. Visualization: M.H., H.L., R.G. Supervision: A.C., M.P., P.C., and J.H.K. Writing – original draft: A.C. Writing – substantial review and editing: A.C., M.H., A.H., N.K., P.C., H.K., J.H.K., and M.P. **Competing interests:** A.C. holds equity in Spring Care and is the lead inventor on 3 patent submissions relating to treatment for major depressive disorder (US Patent and Trademark Office number Y0087.70116US00 and provisional application numbers 62/491 660 and 62/629 041). M.H. and J.B. are employed by and hold equity in Spring Care. RG received royalties from the book *Statistical Methods in Psychiatry and Related Fields* published by CRC Press and is an inventor on US patent application 20200143922. A.H. was a member of advisory boards and received paid speakership by Boehringer-Ingelheim, Lundbeck, Otsuka, Rovi, and Recordati. He received paid speakership by AbbVie and Advanz. He is editor of the AWMF German guidelines for schizophrenia. J.H.K. has been a consultant and/or advisor to or has received honoraria from Janssen/J&J, Lundbeck and Boehringer Ingelheim. P.C. is co-founder and Board Member of Tetricus Labs and reported holding stock and stock options in Tetricus Labs Inc. H.K. received funding from Johnson and Johnson through Yale University. J.Kr. reported holding patents licensed to Johnson and Johnson and Freedom Biosciences; co-founder of Freedom Biosciences, stock in Spring Health, Biohaven Pharmaceuticals, Neumora Pharmaceuticals; Consultant to Biogen, Bionomics Limited, Boehringer Ingelheim International, Cerevel Therapeutics, Jazz Pharmaceuticals, Otsuka American Pharmaceutical Inc., Perception Neuroscience, Sumitomo America, Taishisho, Takeda, BioXcel, Psychogenics Inc. **Data and materials availability:** The study design, outcome measure, and cross-validation approach was preregistered on 1 August 2016 (YODA #2016-1005). All data are available via the Yale University Open Data Access (YODA) platform (https://yoda.yale.edu/request). Data accession numbers for the five trials analyzed here are: R076477-PSZ-3001 (Teens trial), R076477-SCH-3015 (Adults First Episode), R076477-SCH-302 (Older Adults), R076477-SCH-305 (Adults Chronic #1), R076477-SCH-303 (Adults Chronic #2). The code to reproduce all results in this manuscript and the supplement is available at Zenodo (*42*). This study, carried out under YODA Project #2016-1005, used data obtained from the Yale University Open Data Access Project, which has an agreement with JANSSEN RESEARCH & DEVELOPMENT, L.L.C. The interpretation and reporting of research using this data are solely the responsibility of the authors and does not necessarily represent the official views of the Yale University Open Data Access Project or JANSSEN RESEARCH & DEVELOPMENT, L.L.C.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adg8538
Methods and Results
Figs. S1 to S15
Tables S1 to S20
References (*42*)
MDAR Reproducibility Checklist

Submitted 27 January 2023; resubmitted 20 July 2023
Accepted 10 November 2023
10.1126/science.adg8538

Illusory generalizability of clinical prediction models

Adam M. Chekroud, Matt Hawrilenko, Hieronimus Loho, Julia Bondar, Ralitza Gueorguieva, Alkomiet Hasan, Joseph Kambeitz, Philip R. Corlett, Nikolaos Koutsouleris, Harlan M. Krumholz, John H. Krystal, and Martin Paulus

Science **383** (6679), . DOI: 10.1126/science.adg8538

Editor's summary
A central promise of artificial intelligence (AI) in healthcare is that large datasets can be mined to predict and identify the best course of care for future patients. Unfortunately, we do not know how these models would perform on new patients because they are rarely tested prospectively on truly independent patient samples. Chekroud *et al*. showed that machine learning models routinely achieve perfect performance in one dataset even when that dataset is a large international multisite clinical trial (see the Perspective by Petzschner). However, when that exact model was tested in truly independent clinical trials, performance fell to chance levels. Even when building what should be a more robust model by aggregating across a group of similar multisite trials, subsequent predictive performance remained poor. — Peter Stern