Review

Biological Psychiatry: GOS

The Transition From Homogeneous to Heterogeneous Machine Learning in Neuropsychiatric Research

Qingyu Zhao, Kate B. Nooner, Susan F. Tapert, Ehsan Adeli, Kilian M. Pohl, Amy Kuceyeski, and Mert R. Sabuncu

ABSTRACT

Despite the advantage of neuroimaging-based machine learning (ML) models as pivotal tools for investigating brain-behavior relationships in neuropsychiatric studies, these data-driven predictive approaches have yet to yield substantial, clinically actionable insights for mental health care. A notable impediment lies in the inadequate accommodation of most ML research to the natural heterogeneity within large samples. Although commonly thought of as individual-level analyses, many ML algorithms are unimodal and homogeneous and thus incapable of capturing the potentially heterogeneous relationships between biology and psychopathology. We review the current landscape of computational research targeting population heterogeneity and argue that there is a need to expand from brain subtyping and behavioral phenotyping to analyses that focus on heterogeneity at the relational level. To this end, we review and suggest several existing ML models with the capacity to discern how external environmental and sociodemographic factors moderate the brain-behavior mapping function in a data-driven fashion. These heterogeneous ML models hold promise for enhancing the discovery of individualized brain-behavior associations and advancing precision psychiatry.

https://doi.org/10.1016/j.bpsgos.2024.100397

Relating individual differences in brain function and structure as recorded by neuroimaging, e.g., anatomical, diffusion, and functional magnetic resonance imaging (fMRI), to phenotypic data such as cognitive performance, behaviors, and psychiatric symptoms is a fundamental pursuit of human neuroscience (1–4). Identifying accurate brain-behavior relationships and generalizable neuroimaging biomarkers can elucidate the pathophysiology underlying psychiatric symptoms (5,6) and, in turn, point to neurobiological targets that inform treatment design, early interventions such as behavioral therapy or neurostimulation (7,8), and risk assessment for future symptom onset and recovery (9–11). An emerging approach to probing brain-behavior mapping is data-driven machine learning (ML) (4,12,13). In particular, ML models are trained to predict the psychiatric phenotypes of an individual from their neuroimaging data (e.g., fMRI). Based on the trained model, a model-explanation procedure generates neuroimaging biomarkers by identifying specific brain circuits, regions, or measurements that drive the model prediction (14).

As brain-based predictive modeling attracts growing attention in various clinical contexts (3,10,15), a critical point often overlooked is how ML research meets the increasingly recognized challenge of population heterogeneity in human neuroimaging studies (16–19). While much research has focused on analyzing heterogeneity in neural and behavioral data (e.g., neural subtyping or behavioral phenotyping), we argue that a key factor limiting the predictive accuracy and clinical impact of current brain-based ML models may be their inability to capture the heterogeneous mapping function between neuroimaging measurements and psychiatric phenotype (i.e., heterogeneity in the relationship). In particular, most brain-based ML models strive to find a single universal pattern that is applicable to most of the data but fail to account for how the brain-behavior mapping itself may vary according to environmental and sociodemographic factors. To overcome this challenge, we outline potential strategies to build ML models that capture such relationship-level heterogeneity and discuss some bottlenecks in model development and deployment. We envision that only by expanding heterogeneity analysis from the data level to the relationship level can ML truly unleash its power to derive neuroimaging biomarkers that can inform customized prevention, intervention, and treatment.

FAILURE OF HOMOGENEOUS AND UNIMODAL ML IN NEUROPSYCHIATRIC RESEARCH

Despite recent advances in ML, a notable limitation is that brain-based predictive modeling often has low accuracy (4,15). When applying a trained model to an independent, out-of-sample population, the predicted phenotypic measure can significantly deviate from the observed value. Reviews on fMRI studies commonly report $<$70% accuracy in predicting diagnosis outcome of patients with various behavioral and mental health disorders (10,20). These accuracy scores, although higher than random chance, are lower than the 91% average

SEE COMMENTARY NO. 100425

ISSN: 2667-1743

accuracy of 503 medical imaging–based artificial intelligence (AI) models (for all types of medical diagnostics) found in a recent meta-analysis (21). In addition to the low accuracy, the compromised interpretability of these models is another critical concern (22). When interpreting a model with a 70% classification accuracy, neuroimaging biomarkers identified from the training samples do not apply to approximately 30% of the unseen population, suggesting a potential room for improvement in the generalizability and translational potential of current brain-based ML models.

There has been a rich discussion on the various reasons for low prediction accuracy (4,15), with increasing attention being paid to the issue of population heterogeneity (17,23–25). Complex demographic, socioenvironmental, physiological, and clinical factors can influence brain development, behaviors, and cognition, such that no single brain-behavior mapping likely fits all (26,27). Thus, population heterogeneity is one of the many reasons why models trained on large, consortia-sized samples (e.g., the Adolescent Brain Cognitive Development [ABCD] Study) (28) achieved notably lower accuracy than those on smaller homogeneous datasets (4,10,29–32). Indeed, existing predictive analyses in neuropsychiatric research are often unimodal and homogeneous: one universal mapping function is learned to predict the target behavioral phenotype solely based on neuroimaging data and, furthermore, is expected to generalize across populations. This conceptual gap between homogeneous and heterogeneous mapping might result in the ML models being accurate only for a substratum of the population. Moreover, the interpretation of homogeneous models tends to identify neuroimaging biomarkers shared across individuals (33,34), which may as well be identified through group-level statistical tests. In fact, in many ML analyses, the prescreening or post hoc validation of neuroimaging biomarkers in turn resorts to group analysis (35,36), thus trapping what seems to be an individual-level ML analysis into a group-level analysis in disguise.

BEHAVIORAL AND NEURAL HETEROGENEITY

The concept of population heterogeneity is not new. Researchers have been mostly studying the heterogeneity issue from two perspectives: precision behavioral phenotyping (Figure 1A) and neural subtyping (Figure 1B). Precision behavioral phenotyping aims to improve the understanding of behavioral heterogeneity underlying traditional clinical diagnosis: individuals meeting the criteria for a particular disorder codified by DSM exhibit significant variations in clinical symptoms, progression, and prognosis (18,24). Thus, researchers have explored alternative frameworks to harness dimensions of cognitive and behavioral constructs that complement existing diagnostic categories, enhance clinical predictions, and clarify etiology. For example, the Research Domain Criteria initiative (37) uses expert-defined dimensionally distributed quantitative traits to replace the traditional taxonomic structure of mental disorders, exploiting symptom variation across the full spectrum of severity. Beyond multidimensionality, the Hierarchical Taxonomy of Psychopathology consortium (38) identifies categorical hierarchically organized levels of diagnosis from the p factor model (39). The hierarchy encompasses both broad transdiagnostic psychopathology and narrow symptom components and thus provides a more precise and granular description of an individual's behavioral and mental health profile. Other than factor models, various clustering and dimension reduction approaches (40–43) are used for analyzing phenotypic data across multiple domains (e.g., psychopathology, personality, cognition, social functioning) to identify stable transdiagnostic clusters that cut across traditional diagnostic labels. These data-driven subtypes may exhibit greater reproducibility than those derived by hypothesis-driven methods, which could be biased by existing theories and prior assumptions.

Another type of heterogeneity analysis focuses on identifying neurobiological subtypes that support dimensions of psychopathology within a board diagnostic category. As normative modeling studies (44) of diverse brain disorders have repeatedly suggested, the specific location in the brain that indicates differences between cases and controls varies considerably across individuals with the same diagnosis (25,45). Thus, unsupervised or weakly supervised clustering is often used for revealing categorical neural subtypes (46–48), where each person is assigned to one subgroup that has more homogeneous neuropathological patterns than the overall diseased population. Alternative to categorical subtypes, neural heterogeneity can also be formulated as dimensional subtypes (e.g., by canonical correlation analysis) (49–51), where multiple constellations of neuropathological components are linked to continuous scores of specific behavioral dimensions. These biological subtypes (or subdimensions) were shown to have high correlation with within-diagnosis heterogeneity of symptom profiles (52) and have distinct treatment responses (53,54).

RELATIONSHIP-LEVEL HETEROGENEITY IN BRAIN-BASED PREDICTIVE ML

Despite the improved understanding of behavioral (Figure 1A) and neural (Figure 1B) heterogeneity, another type of heterogeneity that is often overlooked in building brain-based predictive ML models is the nonuniform neurobehavioral mapping (Figure 1C). In other words, the heterogeneity of interest lies not only in the neural and behavioral data themselves but also in the mapping function between them. Such mapping can reflect distinct biological mechanisms underlying the same disorder or the natural variation in brain-behavior relationships across individuals. For example, in a post hoc analysis of a single classification model that predicts cognitive test scores from fMRI data, Greene *et al.* (26) found that sociodemographic factors (e.g., race or level of education) are one of the main factors affecting the predictive accuracy of the model across subpopulations, an effect that was replicated in multiple datasets. In another study of 2262 children from the ABCD Study and 752 young adults from the Human Connectome Project (55), researchers show that network features predicting 33 internalizing (e.g., anxiety) and externalizing (e.g., rule breaking) behaviors were distinct across children and adults, suggesting that brain-based predictors of behaviors may change across the life span. Besides demographic constructs, environmental factors can also affect physiological functions of the brain and its ability to counteract pathological changes (56–58). A study analyzing resting-state fMRI of 6839 children

Biological Psychiatry: GOS

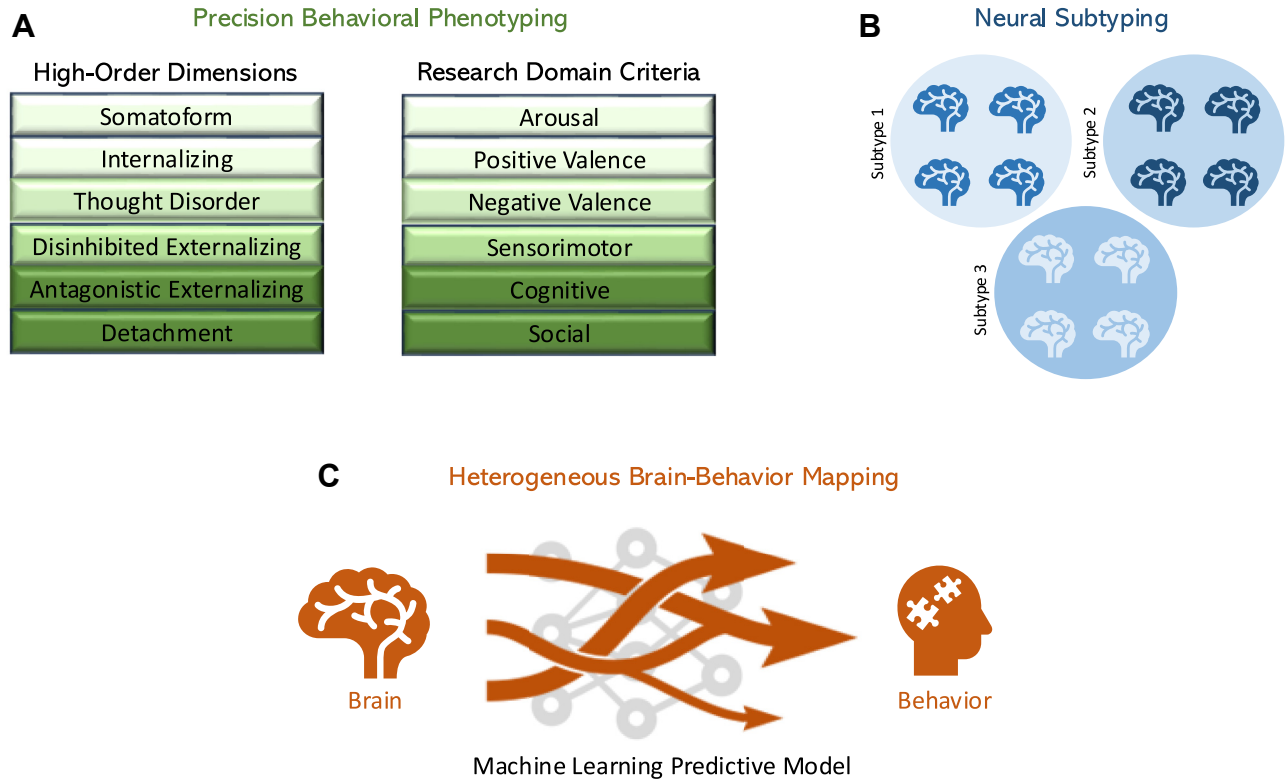

Figure 1. Data heterogeneity refers to categorical or dimensional subtypes of **(A)** psychopathology or **(B)** neural biology, whereas relationship-level heterogeneity examines **(C)** the nonuniversal mapping function between brain and behavioral phenotype.

from the ABCD dataset found that the correlation pattern between selective brain networks and cognitive performance varies as a function of a child's environment (59). While better cognitive performance among children from high-income households correlated with weaker coupling between the lateral frontoparietal network and default mode network, the direction of association was reversed for children from low-income households. Taking all the evidence above, we argue that brain-based ML research may benefit from explicitly incorporating external sociodemographic and environmental factors to gain insights into multiple causal pathways from the brain to behavior.

To capture such relationship-level heterogeneity, some existing solutions rely on training, comparing, and interpreting separate models in each subcohort (27,60). For example, one can train sex-specific models (61–63) to predict disease outcome and examine whether there exist connectome patterns that are uniquely associated with males or females. While being a simple and interpretable approach, training separate models would require subcohort division based on an a priori identified categorical factor. As such, one caveat might be that model training becomes unreliable as the number of subcohorts exponentially grows with the number of considered factors and the number of samples in each subcohort drastically decreases. The separately trained ML models are more likely to overfit these smaller subcohorts (64). Alternative to defining subcohorts a priori is to first use data-driven approaches to subtype the data and then derive subtype-specific predictive models (65–67). For example, Drysdale *et al.* (68) first identified 4 clusters of individuals with depression symptoms and then trained separate ML models to predict treatment response to repetitive transcranial magnetic stimulation for each subtype. Not only were the subtype-specific models significantly more accurate than the universal model trained on the entire population but the neuroimaging biomarkers driving the prediction were also substantially different across subtypes. Similarly, Chen *et al.* (65) clustered 81 infants into 2 subgroups based on brain functional connectivity. Those subgroups had contrasting IQ in a 4-year follow-up and distinct functional connectivity patterns in neonates that predicted 4-year IQ. Although these results highlight the need to consider variability in brain-behavior relationships in brain-based ML analysis, the subtyping and predictive modeling steps are still formulated as 2 separate procedures. The successful training of the predictive models depends on the quality of derived subtypes and may fail to capture common population-level mappings.

Therefore, we believe that the community will benefit from exploring a new ML regime, which we call heterogeneous ML, to enable the discovery of interpretable relationship-level heterogeneity in an end-to-end, data-driven fashion. These models can be regarded as conditional models, wherein the relationship between input and output variables is conditioned on the moderators. Ideally, a heterogeneous ML model should

be able to encode a population-level mapping function just as well as a homogeneous ML model but with the additional capacity to encode how environmental and sociodemographic factors alter that mapping. These models can not only improve understanding of brain-behavior relationship variability across individuals (69) but they also have the potential to increase the generalizability of models across populations that contain sampling bias and/or domain shifts (70,71).

POTENTIAL APPROACHES FOR HETEROGENEOUS ML

We now turn to the question "What might be an appropriate computational approach for heterogeneous ML?" Traditional approaches for probing nonuniform brain-behavior relationships are largely based on multiple regression analysis (72), which predicts a single brain measurement from diagnosis group, behavioral phenotype, and demographic variables. Known as the moderation effects, the varying brain-behavior relationship is achieved by adding interaction terms between behavioral measurements and covariates of interest (72,73) (Figure 2A). While being simple and interpretable, these models are univariate in nature and typically analyze each voxel or region of interest independently, which can miss important dependencies among multiple brain regions (12).

To capture the multivariate brain patterns, ML predictive models reverted the direction of regression by predicting the diagnosis group or behavioral phenotype from all available brain measurements (4,10,13). However, most predictive models are homogeneous and do not consider important covariates in the training process. The challenge in identifying moderation effects in whole-brain exploratory analysis is that the high dimensional neuroimaging measurements would require a large number of interaction terms as the model input, likely resulting in training instability and overfitting. Thus, researchers need to explore alternative strategies to reliably model moderation effects in brain-based predictive models. A possible existing ML technique that could be used for this purpose is ensemble learning (74,75). For example, the mixture-of-experts algorithm aims to learn an ensemble of expert models, where each expert is specifically tailored for a subset of data (76–78). When used in brain-behavior mapping, mixture-of-experts models can define categories of relationships by learning a finite number of representative brain-behavior mapping functions that exist within the population. The final mapping for each individual is thus a weighted combination of the expert models (79) (Figure 2C) moderated by their external factors. To further interpret the encoding of moderation effects by mixture-of-experts models, future research should focus on understanding how the external factors drive the division of experts (or data subsets) and the weighting scheme for each individual.

The past few years have witnessed the increasing use of deep learning in brain-behavior mapping (10,43,54), with some preliminary evidence suggesting its superior predictive performance compared with traditional ML methods (80). These highly nonlinear models are intrinsically more expressive than traditional ML models and can be potentially better suited for learning heterogeneous mapping functions (81,82). A high-capacity deep learning model coupled with large-scale data, in principle, can implicitly extract the information related to latent factors from the neuroimaging data (83–85) and subsequently use that information to learn nonuniform brain-behavior mapping within the population. To explicitly link those latent factors with interpretable environmental and sociodemographic factors, multimodal predictive models can be used to combine imaging and nonimaging (e.g., tabular) data as the input (i.e., multimodal data fusion) (86–89) to learn factor-modulated mapping functions. However, identifying whether and how external factors interact with neuroimaging measures or whether they are simply learned as final additive effects can be challenging. One potential way to increase the interpretability of relationship-level heterogeneity in deep learning models is to use a hypernetwork (90) mimicking the interaction term in traditional regression analysis. A hypernetwork is a neural network that generates a subset of parameters in a deep learning model. As opposed to the previous categorical setup in mixture-of-experts, this type of relationship-level heterogeneity captures dimensional moderation effect, as the parameters of the core brain-behavior mapping network are continuously moderated by external factors represented within the second hypernetwork (Figure 2B). Given that deep networks are characterized by hierarchical extraction and abstraction of features from input data (91), one has the flexibility to choose which specific layer(s) of the network is moderated, thereby bypassing the need for including the full interaction of external variables with whole-brain connectome measurements.

UPCOMING CHALLENGES AND NEXT STEPS

Despite the vision that heterogeneous ML might help unlock the mystery of nonuniform mapping between biology and psychopathology, some key components require careful consideration.

Related Statistical Concepts

Other than moderation, there exist other types of statistical relationships that are commonly studied in biostatistics to model the causal effects of a third variable on treatment and outcome variables (92). Two closely related concepts are mediation and confounding (Figure 3A). Originating from different research domains, these statistical concepts sometimes can only be distinguished on conceptual grounds (93) while modeled identically in computational analysis (92). Therefore, heterogeneous ML might benefit from the causal ML literature that studies generic probabilistic graphical models based on causal directed acyclic graphs (94–97). In bridging the gap between these statistical methods, we have to always keep in mind that computational models are largely influenced by the context in which they are developed. In most modern medical AI applications, for example, the ultimate goal of modeling variable association or causality is usually to make ML models invariant to external factors rather than dependent on them (98–101). These models, known as fair AI or invariant learning (102–105), hold the belief that the predictive behavior of a model should be identical when applied to different subpopulations and unconfounded by sensitive factors such as sex and race (106). As such, existing learning mechanisms (104,105,107,108) often fall short in capturing the meaningful

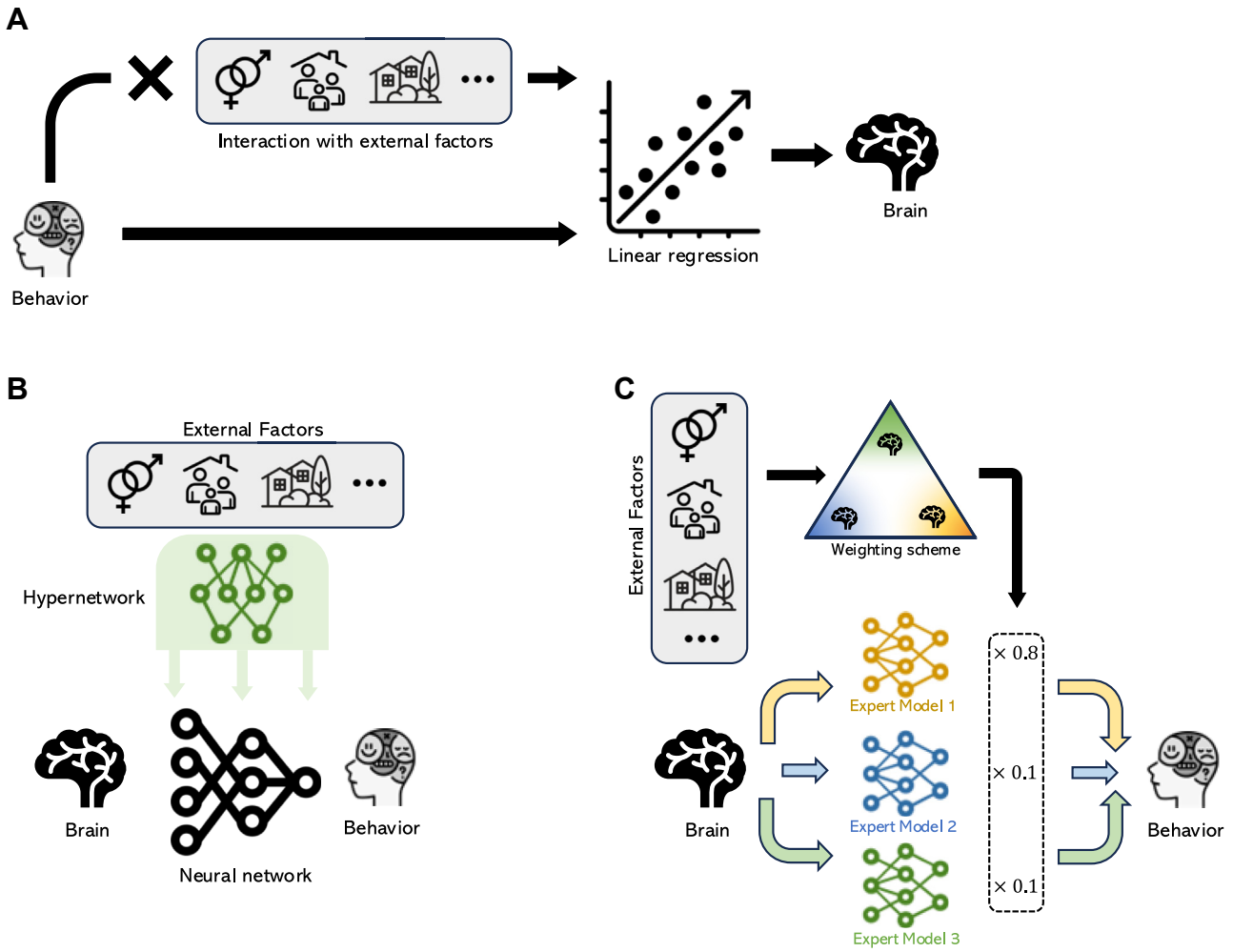

Figure 2. **(A)** A multiple linear regression uses interaction terms to model how external factors moderate the univariate association between a neuroimaging measurement and a behavior variable. **(B)** In a nonlinear machine learning model (e.g., a deep neural network), the dimensional moderation effect can be characterized by using a second hypernetwork framework to generate the parameters of a neural network mapping brain to behavioral phenotype. **(C)** A mixture-of-experts model can learn categorical subtypes of brain-behavior mapping by constructing a set of expert models, each encoding a mapping function tailored for a subpopulation. The specific mapping associated with an individual is a weighted average of the expert models.

moderation effects by those factors. While the idea of fair AI seems to contradict heterogeneous ML, we emphasize that bias, confounding effects, and moderation effects are distinct concepts that do not conflict with each other. Mediators and confounders influence the distribution of input and output variables, whereas moderators influence the mapping function between them (Figure 3A). In other words, it is theoretically possible to design a model in which the model parameters are factor dependent, but the prediction outcome is factor invariant (Figure 3B). Thus, to be inclusive, future brain-based ML research should embrace the concept of factor-dependent models but needs to find new ways to disentangle different statistical relationships.

Training Large-Scale Models

As mentioned, a universal high-capacity model can theoretically capture multiple latent brain-behavior pathways when trained on sufficient data. This argument aligns with the current trend in ML to train a single large-scale foundation model on extensive datasets that is generalizable to various tasks and populations (109–111). This approach has shown promise in advancing clinical, research, and educational workflows in many health care applications (112–114), including neuroscience (115). In the context of brain-behavior mapping, the existence of a shared neural basis underlying psychiatric comorbidity (116–120) suggests the feasibility of developing a unified brain foundation model that can be integrated with neurobiological, psychological, environmental, and physical measurements for predicting symptom outcomes. However, the practical question is whether existing neuroimaging studies contain enough samples for training large-scale models. To date, even the largest neuroimaging datasets (e.g., UK Biobank and the ABCD Study) (28) and meta-analytic approaches (121) contain sample size magnitudes smaller than those for training text- and natural image–based deep learning models.

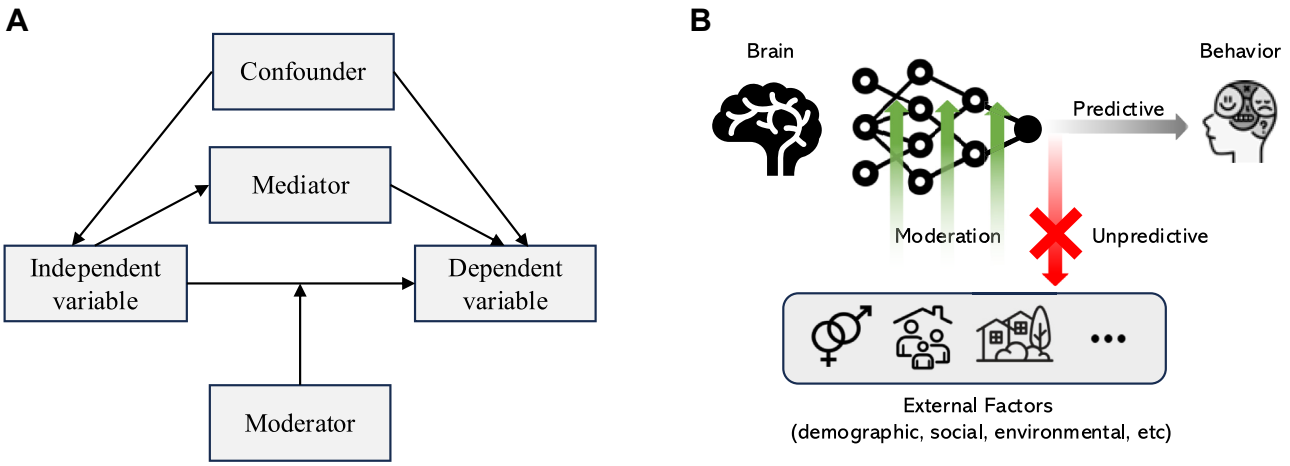

Figure 3. **(A)** Although heterogeneous machine learning aims to model moderation effects in machine learning prediction, different types of statistical relationships exist among independent, dependent, and third variables. However, moderation effects characterize the statistical influence only on variable relationships, not on the variables themselves; **(B)** brain-behavior mapping research should embrace the concept of factor-dependent models and correctly disentangle moderation effects from confounding effects. External factors can moderate the mapping function without biasing the prediction outcome.

Moreover, the increased capacity of large-scale models also heightens the risk of low interpretability, overfitting, and biasing predictions toward confounding effects (122). Thus, until multimodal foundation models are proven feasible in neuropsychiatric prediction tasks, in-depth and reproducible studies will likely rely on smaller-scale models with explicit modeling of relationship-level heterogeneity.

Reproducibility of Brain-Behavior Subtypes

Despite the growing interest in data-driven heterogeneity analysis, efforts in reproducing revealed subtypes are generally lacking (51,123,124). The challenges in replication stem from several key issues. First, there is an increasing concern about misusing ML tools and validation procedures to obtain findings that are not reproducible when subjected to more rigorous statistical analysis (29,125,126). ML recipes discussed in this review need to be combined with stringent data processing, confounding-effect removal, hyperparameter tuning, and preregistration of validation data and experimental setups to generate unbiased subtypes. Proper dissemination of experimental data, code, and trained models is also critical for reproducibility analysis. Next, the complexity of the heterogeneity may arise from the interaction among multiple biological, psychological, and environmental factors (127). As a result, replicating subtypes on independent studies is often challenging, given that the strength of specific moderation effects can be further modulated by other factors across diverse populations and contexts (23). Thus, integrating the proposed methodology (Figure 2B) with large-scale multisite data to model the interaction of multiple moderators can potentially capture the full complexity of the disorder and bridge the discrepancy across studies. Finally, reproducibility analysis should also focus on long-term longitudinal validation (128,129) to determine whether proposed subtypes represent stable, clinically relevant distinctions in disease trajectories or merely transient states or different stages of a single developmental trajectory (130).

Evaluation and Interpretation Procedures

In addition to developing novel algorithms and architectures for training heterogeneous ML, new evaluation procedures need to be introduced to define the success criteria of heterogeneous ML. Commonly used metrics defined on the overall population (e.g., prediction accuracy, area under the curve, R^2) need to be placed into the context of heterogeneous ML to quantify variation in predictive power across individuals (i.e., stratified performance evaluations) (131,132). Therefore, more informative might be metrics such as the correlation between an individual's age and prediction error or an analysis of variance test of area under the curve for different racial/ethnic groups. These metrics should be subject to convergent validity based on external dataset validation, longitudinal assessment, and different types of data sources (e.g., structural or functional neuroimaging, inflammatory biomarkers). To accelerate the translation of the heterogeneous predictive modeling into clinical practice, evaluation should focus on examining the association between subtypes and differences in trajectories of symptom progression (e.g., motor, cognitive, life quality, social functional) and treatment responses (e.g., remission rate, relapse rate, functional improvement) (53,54,133). Finally, the development of heterogeneous ML needs to be accompanied by new compatible model-interpretation procedures. The universal neuroimaging biomarkers need to be refined to ones that are more representative of specific subpopulations. In particular, the interpretation should stringently specify which specific brain circuits are most significantly moderated by external factors and which subset of external factors are significant moderators. Based on the results, researchers should be able to draw conclusions in the following format, e.g., "Lower functional activation in the frontoparietal network is a more salient risk factor for substance use onset in males than in females" or "The association between dorsal attention network functional connectivity and anxiety symptoms is more pronounced with more adverse childhood experiences." Having these

subcohort-specific neuroimaging biomarkers is fundamental to designing tailored prevention and treatment strategies to advance precision psychiatry.

From Subtyping to Individualized Analysis

As discussed, the interpretation of a heterogeneous ML model will produce subtypes of neuroimaging biomarkers linked to external factors. As a more comprehensive set of factors is introduced to the model, the number of subtypes will drastically increase, such that each individual will be characterized by a distinct descriptor. The subtyping analysis is then transformed into a truly personalized predictive analysis. Such fine-grained analysis holds the promise for understanding the variability across individuals (69) and increasing the generalizability of models across populations due to sampling bias and/or domain shifts (70,71). Meanwhile, we also acknowledge that ML is a data-centric approach at its core. The quality of model personalization is tightly tied to the size and quality of training data (88). Data diversity is highly desired to fully stratify the effect of certain factors on brain-behavior mapping (134,135). Thus, common issues in data collection, including missing values, subject selection bias, underrepresentation of specific groups, and skewed distribution in factors, can all lead to model underspecification (34,88,131).

CONCLUSIONS

Many challenges need to be addressed before brain-based modeling can substantially advance our understanding of complex brain-behavior relationships and subsequently translate to improved health care. Here, we reviewed current research analyzing population heterogeneity and identified one underexplored problem: the current design of brain-based ML models fails to examine the heterogeneous nature of brain-behavior mappings. Given that converging evidence points to the benefit of modeling relationship-level heterogeneity, we acknowledge that a more comprehensive study design is required to test the efficacy of heterogeneous ML and evaluate its improvement over unimodal and homogeneous ML. In pursuing this path, understanding how to accurately model moderation and confounding effects of environmental and sociodemographic factors, along with optimally evaluating and interpreting factor-moderated ML models, remain key challenges. Once addressed, ML can unleash its unique power to quantify and disentangle brain-behavior-factor relationships to truly personalize medicine.

ACKNOWLEDGMENTS AND DISCLOSURES

The work was partly supported by the National Institutes of Health (Grant Nos. AA028840 [to QZ], R61AG084471 [to EA], and DA057567 and AA021697 [to KMP]), BBRF Young Investigator Grant (to QZ), the Stanford University Jaswa Innovator Award (to EA), the 2024 Stanford HAI Hoffman-Yee Grant (to EA), the Stanford HAI Google Cloud Credit (to KMP), the DGIST Joint Research Project (to KMP), and University of North Carolina Wilmington Trauma & Resilience Lab (to KBN).

All authors report no biomedical financial interests or potential conflicts of interest.

ARTICLE INFORMATION

From the Department of Radiology, Weill Cornell Medicine, New York, New York (QZ, AK, MRS); Department of Psychology, University of North Carolina Wilmington, Wilmington, North Carolina (KBN); Department of Psychiatry, University of California San Diego, La Jolla, California (SFT); Department of Psychiatry & Behavioral Sciences, Stanford University, Stanford, California (EA, KMP); Department of Computer Science, Stanford University, Stanford, California (EA); Department of Electrical Engineering, Stanford University, Stanford, California (KMP); and School of Electrical and Computer Engineering, Cornell University and Cornell Tech, New York, New York (MRS).

Address correspondence to Qingyu Zhao, Ph.D., at qingyuz@med.cornell.edu.

Received Jul 1, 2024; revised Sep 17, 2024; accepted Sep 18, 2024.

REFERENCES

1. Sui J, Jiang R, Bustillo J, Calhoun V (2020): Neuroimaging-based individualized prediction of cognition and behavior for mental disorders and health: Methods and promises. Biol Psychiatry 88: 818–828.
2. Beam E, Potts C, Poldrack RA, Etkin A (2021): A data-driven framework for mapping domains of human neurobiology. Nat Neurosci 24:1733–1744.
3. Shen X, Finn ES, Scheinost D, Rosenberg MD, Chun MM, Papademetris X, Constable RT (2017): Using connectome-based predictive modeling to predict individual behavior from brain connectivity. Nat Protoc 12:506–518.
4. Woo CW, Chang LJ, Lindquist MA, Wager TD (2017): Building better biomarkers: Brain models in translational neuroimaging. Nat Neurosci 20:365–377.
5. (2013): The benefits of brain mapping. Nature 499:253.
6. Joyce DW, Kormilitzin A, Smith KA, Cipriani A (2023): Explainable artificial intelligence for mental health through transparency and interpretability for understandability. NPJ Digit Med 6:6.
7. Porto PR, Oliveira L, Mari J, Volchan E, Figueira I, Ventura P (2009): Does cognitive behavioral therapy change the brain? a systematic review of neuroimaging in anxiety disorders. J Neuropsychiatry Clin Neurosci 21:114–125.
8. Marzbani H, Marateb HR, Mansourian M (2016): Neurofeedback: A comprehensive review on system design, methodology and clinical applications. Basic Clin Neurosci 7:143–158.
9. Gabrieli JDE, Ghosh SS, Whitfield-Gabrieli S (2015): Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. Neuron 85:11–26.
10. Rashid B, Calhoun V (2020): Towards a brain-based predictome of mental illness. Hum Brain Mapp 41:3468–3535.
11. Ju Y, Horien C, Chen W, Guo W, Lu X, Sun J, *et al.* (2020): Connectome-based models can predict early symptom improvement in major depressive disorder. J Affect Disord 273:442–452.
12. Davatzikos C (2019): Machine learning in neuroimaging: Progress and challenges. Neuroimage 197:652–656.
13. Yarkoni T, Westfall J (2017): Choosing prediction over explanation in psychology: Lessons from machine learning. Perspect Psychol Sci 12:1100–1122.
14. Kohoutová L, Heo J, Cha S, Lee S, Moon T, Wager TD, Woo C-W (2020): Toward a unified framework for interpreting machine-learning models in neuroimaging. Nat Protoc 15:1399–1435.
15. Wu J, Li J, Eickhoff SB, Scheinost D, Genon S (2023): The challenges and prospects of brain-based prediction of behaviour. Nat Hum Behav 7:1255–1264.
16. Price RB, Lane S, Gates K, Kraynak TE, Horner MS, Thase ME, Siegle GJ (2017): Parsing Heterogeneity in the Brain Connectivity of Depressed and Healthy Adults During Positive Mood. Biol Psychiatry 81:347–357.
17. Feczko E, Fair DA (2020): Methods and challenges for assessing heterogeneity. Biol Psychiatry 88:9–17.
18. Tiego J, Martin EA, Deyoung CG, Hagan K, Cooper SE, Pasion R, *et al.* (2023): Precision behavioral phenotyping as a strategy for uncovering the biological correlates of psychopathology. Nat Ment Health 1:304–315.
19. Klooster DCW, Siddiqi SH (2023): Embracing the heterogeneity in depression neuroimaging. Nat Mental Health 1:243–244.

20. Feng W, Liu G, Zeng K, Zeng M, Liu Y (2022): A review of methods for classification and recognition of ASD using fMRI data. J Neurosci Methods 368:109456.
21. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, *et al.* (2021): Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis. NPJ Digit Med 4:65.
22. Liu B, Udell M (2020): Impact of accuracy on model interpretations. arXiv https://doi.org/10.48550/arXiv.2011.09903.
23. Benkarim O, Paquola C, Park B-Y, Kebets V, Hong S-J, Vos de Wael R, *et al.* (2022): Population heterogeneity in clinical cohorts affects the predictive accuracy of brain imaging. PLoS Biol 20: e3001627.
24. Feczko E, Miranda-Dominguez O, Marr M, Graham AM, Nigg JT, Fair DA (2019): The heterogeneity problem: Approaches to identify psychiatric subtypes. Trends Cogn Sci 23:584–601.
25. Segal A, Parkes L, Aquino K, Kia SM, Wolfers T, Franke B, *et al.* (2023): Regional, circuit and network heterogeneity of brain abnormalities in psychiatric disorders. Nat Neurosci 26:1613–1629.
26. Greene AS, Shen X, Noble S, Horien C, Hahn CA, Arora J, *et al.* (2022): Brain-phenotype models fail for individuals who defy sample stereotypes. Nature 609:109–118.
27. Dhamala E, Yeo BT, Holmes AJ (2022): One size does not fit all: Methodological considerations for brain-based predictive modeling in psychiatry. Biol Psychiatry 93:717–728.
28. Casey BJ, Cannonier T, Conley M, Cohen AO, Barch D, Heitzeg M, *et al.* (2018): The adolescent brain cognitive development (ABCD) study: Imaging acquisition across 21 sites. Dev Cogn Neurosci 32:43–54.
29. Marek S, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, Hatoum AS, *et al.* (2022): Reproducible brain-wide association studies require thousands of individuals. Nature 603:654–660.
30. Gaus R, Pölsterl S, Greimel E, Schulte-Körne G, Wachinger C (2023): Can we diagnose mental disorders in children? a large-scale assessment of machine learning on structural neuroimaging of 6916 children in the adolescent brain cognitive development study. JCPP Adv 3:e12184.
31. Xu M, Calhoun V, Jiang R, Yan W, Sui J (2021): Brain imaging-based machine learning in autism spectrum disorder: Methods and applications. J Neurosci Methods 361:109271.
32. Yeung AWK, More S, Wu J, Eickhoff SB (2022): Reporting details of neuroimaging studies on individual traits prediction: A literature survey. Neuroimage 256:119275.
33. Verdi S, Marquand AF, Schott JM, Cole JH (2021): Beyond the average patient: How neuroimaging models can address heterogeneity in dementia. Brain 144:2946–2953.
34. Tejavibulya L, Rolison M, Gao S, Liang Q, Peterson H, Dadashkarimi J, *et al.* (2022): Predicting the future of neuroimaging predictive models in mental health. Mol Psychiatry 27:3129–3137.
35. Kira K, Rendell LA (1992): A practical approach to feature selection. In: Machine Learning Proceedings. Amsterdam: Elsevier, 249–256.
36. Hall M (2000): Correlation-based feature selection for machine learning. New Zealand: Doctoral Thesis, University of Waikato.
37. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, *et al.* (2010): Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. Am J Psychiatry 167:748–751.
38. Kotov R, Krueger RF, Watson D, Cicero DC, Conway CC, Deyoung CG, *et al.* (2021): The hierarchical taxonomy of psychopathology (HiTOP): A quantitative nosology based on consensus of evidence. Annu Rev Clin Psychol 17:83–108.
39. Caspi A, Houts R, Belsky D, Goldman-Mellow S, Harrington H, Israel S, *et al.* (2014): The p factor: One general psychopathology factor in the structure of psychiatric disorders? Clin Psychol Sci 2:119–137.
40. Fleming LM, Lemonde AC, Benrimoh D, Gold JM, Taylor JR, Malla A, *et al.* (2023): Using dimensionality-reduction techniques to understand the organization of psychotic symptoms in persistent psychotic illness and first episode psychosis. Sci Rep 13:4841.
41. Beijers L, Loo H, Romeijn JW, Lamers F, Schoevers R, Wardenaar K (2022): Investigating data-driven biological subtypes of psychiatric disorders using specification-curve analysis. Psychol Med 52: 1089–1100.
42. Pelin H, Ising M, Stein F, Meinert S, Meller T, Brosch K, *et al.* (2021): Identification of transdiagnostic psychiatric disorder subtypes using unsupervised learning. Neuropsychopharmacology 46:1895–1905.
43. Chang M, Womer FY, Gong X, Chen X, Tang L, Feng R, *et al.* (2021): Identifying and validating subtypes within major psychiatric disorders based on frontal-posterior functional imbalance via deep learning. Mol Psychiatry 26:2991–3002.
44. Rutherford S, Kia SM, Wolfers T, Fraza C, Zabihi M, Dinga R, *et al.* (2022): The normative modeling framework for computational psychiatry. Nat Protoc 17:1711–1734.
45. Marquand A, Rezek I, Buitelaar J, Beckmann C (2016): Understanding heterogeneity in clinical cohorts using normative models: Beyond case-control studies. Biol Psychiatry 80:552–561.
46. Kaczkurkin AN, Moore TM, Sotiras A, Xia CH, Shinohara RT, Satterthwaite TD (2020): Approaches to defining common and dissociable neurobiological deficits associated with psychopathology in youth. Biol Psychiatry 88:51–62.
47. Yang Z, Wen J, Abdulkadir A, Cui Y, Erus G, Mamourian E, *et al.* (2024): Gene-SGAN: discovering disease subtypes with imaging and genetic signatures via multi-view weakly-supervised deep clustering. Nat Commun 15:354.
48. Young A, Marinescu R, Oxtoby N, Bocchetta M, Yong K, Firth N, *et al.* (2018): Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. Nat Commun 9:4273.
49. Xia C, Ma Z, Ciric R, Gu S, Betzel R, Kaczkurkin A, *et al.* (2018): Linked dimensions of psychopathology and connectivity in functional brain networks. Nat Commun 9:3003.
50. Mihalik A, Ferreira FS, Rosa MJ, Moutoussis M, Ziegler G, Monteiro JM, *et al.* (2019): Brain-behaviour modes of covariation in healthy and clinically depressed young people. Sci Rep 9:11536.
51. Helmer M, Warrington S, Mohammadi-Nejad AR, Ji JL, Howell A, Rosand B, *et al.* (2024): On the stability of canonical correlation analysis and partial least squares with application to brain-behavior associations. Commun Biol 7:217.
52. Dunlop K, Grosenick L, Downar J, Vila-Rodriguez F, Gunning FM, Daskalakis ZJ, *et al.* (2024): Dimensional and categorical solutions to parsing depression heterogeneity in a large single-site sample. Biol Psychiatry 96:422–434.
53. Zhao K, Fonzo GA, Xie H, Oathes DJ, Keller CJ, Carlisle NB, *et al.* (2024): Discriminative functional connectivity signature of cocaine use disorder links to rtms treatment response. Nat Ment Health 2:388–400.
54. Jiao Y, Fonzo G, Zhang Y (2024): Deep learning of multimodal brain connectome signatures for predicting treatment response in major depressive disorder (MDD). Biol Psychiatry 95:S183.
55. Qu Y, Chen J, Tam A, Ooi L, Dhamala E, Cocuzza C, *et al.* (2023): Distinct brain network features predict internalizing and externalizing traits in children and adults. bioRxiv https://doi.org/10.1101/2023.05.20.541490.
56. Schmitt A, Malchow B, Hasan A, Falkai P (2014): The impact of environmental factors in severe psychiatric disorders. Front Neurosci 8:19.
57. Mandolesi L, Gelfo F, Serra L, Montuori S, Polverino A, Curcio G, Sorrentino G (2017): Environmental factors promoting neural plasticity: Insights from animal and human studies. Neural Plast 2017: 7219461.
58. Tooley UA, Bassett DS, Mackey AP (2021): Environmental influences on the pace of brain development. Nat Rev Neurosci 22:372–384.
59. Ellwood-Lowe ME, Whitfield-Gabrieli S, Bunge SA (2021): Brain network coupling associated with cognitive performance varies as a function of a child's environment in the ABCD study. Nat Commun 12:7183.
60. Schinkel M, Bennis FC, Boerman AW, Wiersinga WJ, Nanayakkara PWB (2023): Embracing cohort heterogeneity in clinical machine learning development: A step toward generalizable models. Sci Rep 13:8363.

61. Nostro AD, Müller VI, Varikuti DP, Pläschke RN, Hoffstaedter F, Langner R, *et al.* (2018): Predicting personality from network-based resting-state functional connectivity. Brain Struct Funct 223:2699–2719.
62. Jiang R, Calhoun V, Fan L, Zuo N, Jung R, Qi S, *et al.* (2020): Gender differences in connectome-based predictions of individualized intelligence quotient and sub- domain scores. Cereb Cortex 30:888–900.
63. Dhamala E, Jamison KW, Jaywant A, Kuceyeski A (2022): Shared func- tional connections within and between cortical networks predict cognitive abilities in adult males and females. Hum Brain Mapp 43:1087–1102.
64. Jollans L, Boyle R, Artiges E, Banaschewski T, Desrivières S, Grigis A, *et al.* (2019): Quantifying performance of machine learning methods for neuroimaging data. Neuroimage 199:351–365.
65. Chen Y, Liu S, Salzwedel A, Stephens R, Cornea E, Goldman B, *et al.* (2021): The subgrouping structure of newborns with heterogenous brain-behavior relationships. Cereb Cortex 31:301–311.
66. Leroy A, Latouche P, Guedj B, Gey S (2023): Cluster-specific predictions with multi-task Gaussian processes. JMLR 24:1–49.
67. Kam TE, Suk H-I, Lee SW (2017): Multiple functional networks modeling for autism spectrum disorder diagnosis. Hum Brain Mapp 38:5804–5821.
68. Drysdale AT, Grosenick L, Downar J, Dunlop K, Mansouri F, Meng Y, *et al.* (2017): Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat Med 23:28–38.
69. Falk E, Hyde L, Mitchell C, Faul J, Gonzalez R, Heitzeg M, *et al.* (2013): What is a representative brain? neuroscience meets population science. Proc Natl Acad Sci U S A 110:17615–17622.
70. Wachinger C, Rieckmann A, Pölsterl S, Alzheimer's Disease Neuroimaging Initiative and the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (2021): Detect and correct bias in multi-site neuroimaging datasets. Med Image Anal 67:101879.
71. Chen Z, Liu X, Yang Q, Wang YJ, Miao K, Gong Z, *et al.* (2023): Evaluation of risk of bias in neuroimaging-based artificial intelligence models for psychiatric diagnosis: A systematic review. JAMA Netw Open 6:e231671.
72. Cohen J, Cohen P, West S, Aiken L (2003): Applied multiple regression/correlation analysis for the behavioral sciences. J R Stat Soc 52.
73. Fairchild A, MacKinnon D (2008): A general model for testing mediation and moder- ation effects. Prev Sci Off J Soc Prev Res 10:87–99.
74. Opitz D, Maclin R (1999): Popular ensemble methods: An empirical study. J Artif Intell Res 11:169–198.
75. Oota SR, Avvaru A, Manwani N, Bapi RS (2018): Mixture of regression experts in fMRI encoding. arXiv https://doi.org/10.48550/arXiv.1811.10740.
76. Baldacchino T, Cross E, Worden K, Rowson J (2016): Variational Bayesian mixture of experts models and sensitivity analysis for nonlinear dynamical systems. Mechanical Systems and Signal Processing 66–67:178–200.
77. Chen Z, Deng Y, Wu Y, Gu Q, Li Y (2022): Towards understanding mixture of experts in deep learning. arXiv https://doi.org/10.48550/arXiv.2208.02813.
78. Xie Y, Huang S, Chen T, Wei F (2022): MoEC: Mixture of expert clusters. AAAI 37:13807–13815.
79. Hampshire JB, Waibel AH (1990): The meta-pi network: Connectionist rapid adaptation for high-performance multi-speaker phoneme recognition. ICASSP 1:165–168.
80. Abrol A, Fu Z, Salman M, Silva R, Du Y, Plis S, Calhoun V (2021): Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. Nat Commun 12:353.
81. Raghu M, Poole B, Kleinberg J, Ganguli S, Sohl-Dickstein J (2017): On the expressive power of deep neural networks. arXiv https://doi.org/10.48550/arXiv.1606.05336.
82. Poole B, Lahiri S, Raghu M, Sohl-Dickstein J, Ganguli S (2016): Exponential expressivity in deep neural networks through transient chaos. arXiv https://doi.org/10.48550/arXiv.1606.05340.
83. Jonsson BA, Bjornsdottir G, Thorgeirsson TE, Ellingsen LM, Walters GB, Gudbjartsson DF, *et al.* (2019): Brain age prediction using deep learning uncovers associated sequence variants. Nat Commun 10:5409.
84. Ryali S, Zhang Y, de Los Angeles C, Supekar K, Menon V (2024): Deep learning models reveal replicable, generalizable, and behaviorally relevant sex differences in human functional brain organization. Proc Natl Acad Sci U S A 121:e2310012121.
85. Gichoya J, Banerjee I, Bhimireddy A, Burns J, Celi L, Chen LC, *et al.* (2022): AI recognition of patient race in medical imaging: A modelling study. Lancet Digit Health 4:e406–e414.
86. Kline A, Wang H, Li Y, Dennis S, Hutch M, Xu Z, *et al.* (2022): Multimodal machine learning in precision health: A scoping review. NPJ Digit Med 5:171.
87. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ (2022): Multimodal biomedical AI. Nat Med 28:1773–1784.
88. Chen ZS, Kulkarni PP, Galatzer-Levy IR, Bigio B, Nasca C, Zhang Y (2022): Modern views of machine learning for precision psychiatry. Patterns (N Y) 3:100602.
89. Zhang YD, Dong Z, Wang SH, Yu X, Yao X, Zhou Q, *et al.* (2020): Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation. Inf Fusion 64:149–187.
90. Ha D, Dai A, Le QV (2016): Hypernetworks. arXiv https://doi.org/10.48550/arXiv.1609.09106.
91. Alzubaidi L, Zhang J, Humaidi A, Al-Dujaili A, Duan Y, Al-Shamma O, *et al.* (2021): Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. J Big Data 8:53.
92. MacKinnon DP, Krull JL, Lockwood CM (2000): Equivalence of the mediation, confounding and suppression effect. Prev Sci 1:173–181.
93. Kraemer HC (2016): Messages for clinicians: Moderators and mediators of treatment outcome in randomized clinical trials. Am J Psychiatry 173:672–679.
94. Nath T, Caffo B, Wager T, Lindquist MA (2023): A machine learning based approach towards high-dimensional mediation analysis. Neuroimage 268:119843.
95. Coelho de Castro D, Walker I, Glocker B (2020): Causality matters in medical imaging. Nat Commun 11:3673.
96. Kaddour J, Lynch A, Liu Q, Kusner MJ, Silva R (2022): Causal machine learning: a survey and open problems. arXiv https://doi.org/10.48550/arXiv.2206.15475.
97. Shen X, Ma S, Vemuri P, Simon G, Weiner M, Aisen P, *et al.* (2020): Challenges and opportunities with causal discovery algorithms: Application to Alzheimer's pathophysiology. Sci Rep 10:2975.
98. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K (2019): The practical implementation of artificial intelligence technologies in medicine. Nat Med 25:30–36.
99. Topol EJ (2019): High-performance medicine: The convergence of human and artificial intelligence. Nat Med 25:44–56.
100. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M (2021): Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 27:2176–2182.
101. Mittermaier M, Raza MM, Kvedar JC (2023): Bias in AI-based models for medical applications: Challenges and mitigation strategies. NPJ Digit Med 6:113.
102. Zou J, Schiebinger L (2018): AI can be sexist and racist — It's time to make it fair. Nature 559:324–326.
103. Lu C, Lemay A, Chang K, Höbel K, Kalpathy-Cramer J (2022): Fair conformal predictors for applications in medical imaging. AAAI 36:12008–12016.
104. Moyer D, Gao S, Brekelmans R, Galstyan A, Ver Steeg G (2018): Invariant representations without adversarial training. arXiv https://doi.org/10.48550/arXiv.1805.09458.
105. Creager E, Madras D, Jacobsen JH, Weis M, Swersky K, Pitassi T, Zemel R (2019): Flexibly fair representation learning by disentanglement. arXiv https://doi.org/10.48550/arXiv.1906.02589.
106. Liu M, Ning Y, Teixayavong S, Mertens M, Xu J, Ting DSW, *et al.* (2023): A translational perspective towards clinical AI fairness. NPJ Digit Med 6:172.
107. Xie Q, Dai Z, Du Y, Hovy EH, Neubig G (2017): Controllable invariance through adversarial feature learning. arXiv https://doi.org/10.48550/arXiv.1705.11122.

108. Roy PC, Boddeti V (2019): Mitigating information leakage in image representations: A maximum entropy approach. arXiv https://doi.org/10.48550/arXiv.1904.05514.
109. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, *et al.* (2022): On the opportunities and risks of foundation models. arXiv https://doi.org/10.48550/arXiv.2108.07258.
110. Azad B, Azad R, Eskandari S, Bozorgpour A, Kazerouni A, Rekik I, Merhof D (2023): Foundational models in medical imaging: A comprehensive survey and future vision. arXiv https://doi.org/10.48550/arXiv.2310.18689.
111. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, Rajpurkar P (2023): Foundation models for generalist medical artificial intelligence. Nature 616:259–265.
112. Singhal K, Azizi S, Tu T, Mahdavi S, Wei J, Chung H, *et al.* (2023): Large language models encode clinical knowledge. Nature 620:172–180.
113. Huang Z, Bianchi F, Yuksekgonul M, Montine TJ, Zou J (2023): A visual-language foundation model for pathology image analysis using medical twitter. Nat Med 29:2307–2316.
114. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, *et al.* (2023): The Shaky Foundations of clinical foundation models: A survey of large language models and foundation models for EMRs. arXiv https://doi.org/10.48550/arXiv.2303.12961.
115. Caro JO, Fonseca AHde O, Averill C, Rizvi SA, Rosati M, Cross JL, *et al.* (2023): BrainLM: A foundation model for brain activity recordings. bioRxiv https://doi.org/10.1101/2023.09.12.557460.
116. Xie C, Xiang S, Shen C, Peng XR, Kang J, Li Y, *et al.* (2023): A shared neural basis underlying psychiatric comorbidity. Nat Med 29:1232–1242.
117. Huang ZA, Liu R, Zhu Z, Tan K (2022): Multitask learning for joint diagnosis of multiple mental disorders in resting-state fMRI. IEEE Trans Neural Netw Learn Syst 35:8161–8175.
118. Adeli E, Kwon D, Pohl K (2018): Multi-label transduction for identifying disease comorbidity Patterns. Med Image Comput Assist Interv 11072:575–583.
119. Barch DM (2017): The neural correlates of transdiagnostic dimensions of psychopathology. Am J Psychiatry 174:613–615.
120. McTeague LM, Rosenberg BM, Lopez JW, Carreon DM, Huemer J, Jiang Y, *et al.* (2020): Identification of common neural circuit disruptions in emotional processing across psychiatric disorders. Am J Psychiatry 177:411–421.
121. Thompson P, Jahanshad N, Ching C, Salminen LE, Thomopoulos SI, Bright J, *et al.* (2020): ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries. Transl Psychiatry 10:100.
122. Mahfuzur Rahman Md, Calhoun VD, Plis SM (2023): Looking deeper into interpretable deep learning in neuroimaging: a comprehensive survey. arXiv https://doi.org/10.48550/arXiv.2307.09615.
123. Baker M (2016): 1,500 scientists lift the lid on reproducibility. Nature 533:452–454.
124. Mohanty R, Mårtensson G, Poulakis K, Muehlboeck JS, Rodriguez-Vieitez E, Chiotis K, *et al.* (2020): Comparison of subtyping methods for neuroimaging studies in alzheimer's disease: a call for harmonization. Brain Commun 2:fcaa192.
125. Argamon SE (2019): People cause replication problems, not machine learning. Available at: https://www.americanscientist.org/blog/macroscope/people-cause-replication-problems-not-machine-learning#:~:text=There%20is%20simply%20no%20substitute,and%20not%20with%20anything%20else. Accessed June 15, 2024.
126. Beam AL, Manrai AK, Ghassemi M (2020): Challenges to the reproducibility of machine learning models in health care. JAMA 323:305–306.
127. Fishbein D (2000): The importance of neurobiological research to the prevention of psychopathology. Prev Sci 1:89–106.
128. Besiroglu L, Uguz F, Ozbebit O, Guler O, Cilli AS, Askin R (2007): Longitudinal assessment of symptom and subtype categories in obsessive-compulsive disorder (2007). Depress Anxiety 24:461–466.
129. Pourzinal D, Yang J, Sivakumaran K, McMahon KL, Mitchell L, O'Sullivan JD, *et al.* (2023): Longitudinal follow up of data-driven cognitive subtypes in Parkinson's disease. Brain Behav 13:e3218.
130. Poulakis K, Pereira J, Muehlboeck JS, Wahlund LO, Smedby Ö, Volpe G, *et al.* (2022): Multi-cohort and longitudinal bayesian clustering study of stage and subtype in alzheimer's disease. Nat Commun 13:4566.
131. D'Amour A, Heller K, Moldovan D, Adlam B, Alipanahi B, Beutel A, *et al.* (2022): Underspecification presents challenges for credibility in modern machine learning. J Mach Learn Res 23:1–61.
132. Miller AC, Gatys LA, Futoma J, Fox EB (2021): Model-Based Metrics: Sample-Efficient Estimates of Predictive Model Subpopulation Performance. arXiv https://doi.org/10.48550/arXiv.2104.12231.
133. Maletic V, Robinson M, Oakes T, Iyengar S, Ball SG, Russell J (2007): Neurobiology of depression: An integrated view of key findings. Int J Clin Pract 61:2030–2040.
134. Kopal J, Uddin LQ, Bzdok D (2023): The end game: Respecting major sources of population diversity. Nat Methods 20:1122–1128.
135. Gong Z, Zhong P, Hu W (2019): Diversity in machine learning. IEEE Access 7:64323–64350.