Biological Psychiatry

Review

Overview of Challenges in Brain-Based Predictive Modeling: Toward Meaningful Predictive Insights

Vera Komeyer, Nicolás Nieto, Simon B. Eickhoff, Federico Raimondo, and Kaustubh R. Patil

ABSTRACT

Predictive analytics based on machine learning (ML) and artificial intelligence is a powerful tool enabling precision psychiatry and providing insights into brain-behavior relationships. However, given the mixed results observed in the field so far, making meaningful progress requires careful consideration of several key challenges to ensure the validity of models and findings, including overfitting, confounding biases, site effect harmonization, and interpretability, among others. First, we highlight limitations of cross-validation, a ubiquitous ML strategy used to prevent overfitting and obtain generalization estimates, emphasizing the risk of performance inflation and the need for independent validation. Next, we introduce different types of so-called third variables that can influence the examination of a brain-behavioral relationship of interest in different ways, using causal inference principles. We emphasize the biasing impact of confounding variables on ML models and summarize common mitigation strategies. We then discuss site-specific effects in multisite datasets, reviewing different harmonization strategies to reduce unwanted variability and site-specific noise. Finally, we explore post hoc model interpretation methods to enhance model transparency while cautioning against misinterpretation. By integrating rigorous result validation, confounder control, and interpretability techniques, researchers can ensure that ML models produce more reliable and generalizable findings and avoid spurious associations.

https://doi.org/10.1016/j.biopsych.2025.09.003

Predictive machine learning (ML) models using neuroimaging data that reliably forecast clinical outcomes and individual risk profiles can facilitate personalized and effective psychiatric treatments. Such models can assist conventional symptom-based assessments by providing objective, data-driven diagnostic tools, thereby addressing a longstanding need for objective biomarkers in psychiatry (1–5). Improved data acquisition, large-scale data sharing, and advanced analytics together have raised expectations for future advancements (6–9). Recent research has demonstrated the capability of predictive models to uncover complex and subtle patterns in neuroimaging data supporting diverse tasks including diagnosis, prediction of treatment response, and disease subtyping for diverse psychiatric conditions (10–16). Furthermore, neuroimaging-based ML models have helped develop and validate new symptom-based scores that capture individual differences better than traditional diagnoses, supporting more personalized assessment and treatment in mental health (2,17,18).

However, some studies have questioned the robustness and generalizability of ML results (19,20). In addition to addressing data reliability (21,22), it is crucial to tackle key challenges such as data biases including confounding effects, harmonization of multisite datasets, and methodological issues in model evaluation and interpretation. In this perspective article, we aim to deepen the understanding of these challenges and associated methodological caveats and provide guidance where possible.

Supervised ML learns patterns from measured data, where input features (X), such as imaging-derived phenotypes, are used to predict a target variable (Y), such as a clinical diagnosis. Its strength lies in detecting subtle multivariate relationships often missed by conventional analytical methods. Crucially, the main objective of a predictive model is to achieve accurate predictions on new unseen data—referred to as out-of-sample generalization. Cross-validation (CV) is commonly used for estimating generalization performance. However, it can produce biased and unstable results, particularly leading to overoptimistic performance estimates when sample sizes are limited (20,23). Therefore, the output of a CV process must be treated carefully, as we will elaborate in CV Is an Estimation. Biases in data can lead to biased models and misleading insights, compromising both clinical decision making and scientific insights. Understanding Third-Variable Effects in Biomedical Research addresses these biases through the lens of third variables, introducing key causal concepts to identify and manage them. Special emphasis is placed on confounding variables, commonly encountered type of third variables, particularly relevant in observational data (24–27). ML models typically perform better with large datasets (28), but combining data from multiple sites can introduce site effects—biases from systematic differences in

ISSN: 0006-3223

acquisition (29). Data harmonization helps address this (30,31), but Effects of Site and Data Harmonization highlights key caveats when applying commonly used data harmonization methods within ML pipelines. Finally, Post Hoc Model Interpretation details the need to carefully consider model characteristics and the validity of post hoc interpretations to ensure meaningful insights (32)—key to clinical adoption of predictive analytics. The key challenges and potential mitigation strategies are summarized in Table 1.

CV IS AN ESTIMATION

Much like the process of conducting clinical trials in drug development requires extensive testing on independent patient populations—often involving years of research, regulatory approval, and significant financial investment—collecting new datasets to validate ML models can also be lengthy and resource intensive. To circumvent this obstacle, the widely adopted methodology to evaluate a model's out-of-sample generalization is to partition the data into training and testing sets, emulating unseen data. This is known as CV.

CV comes with several challenges and limitations that have been extensively discussed in the literature, including unreliable estimations of variance (33), overoptimistic and unstable performance estimates due to small sample size (23), overfitting (34), concerns related to data averaging (35), and selective reporting of findings (36). A fundamental limitation of CV, however, has not been emphasized—it provides an estimation rather than measuring a model's true performance. While CV is often thought to estimate how a particular model is expected to perform on new unseen samples, it only estimates the average performance of models trained on different but equally sized overlapping subsets of the available data (37).

Drawing on the analogy with clinical trials in drug development, CV estimates can be thought of as results obtained during the early phases of such trials. Like drugs, ML models are intended for real-world deployment once proven effective. However, as with pharmaceuticals—where approximately 90% of candidates ultimately fail, and even 41% fail during phase III despite earlier success (38)—ML models often encounter similar challenges. A model may demonstrate high accuracy under CV but still fail when applied beyond the controlled development setting and data, highlighting the limitations of early-phase evaluation (39).

Building an ML model encompasses selecting from a diverse pool of data processing steps and learning algorithms that can be parametrized and combined in a myriad of ways. This flexibility often leads to iterative refinement to obtain high CV accuracy. This iterative process, even when unintentional or carried out by different research teams, exacerbates performance overestimation: As models are repeatedly tuned and retested on the same data, they become increasingly tailored to idiosyncrasies of the sample rather than learning robust, generalizable patterns (40), which might even lead to false positives (41). Overestimated performance can lead to false confidence in the model's ability to uncover meaningful and general biological or behavioral associations, ultimately skewing conclusions and limiting reproducibility and real-world applicability. Readers should be mindful of the inherent limitations of CV and consider the broader context in which the results were obtained, including sample size, origin of the data, and previous findings in the literature. In addition to every researcher following good practices, the responsibility lies with the reader to critically assess whether the reported findings are robust, generalizable, and meaningful within their specific domain of application.

We recommend using nested CV for unbiased error estimates, with comparisons being restricted to a preselected set of candidate workflows. To avoid data leakage, all preprocessing should be strictly performed within training folds (42). Statistical model comparisons should rely on appropriate paired tests (43), and all tested models and hyperparameters should be reported transparently. Finally, results should be interpreted carefully, keeping in mind that CV estimates reflect the average performance of the modeling procedure rather than the exact error of the final fitted model. The latter should be validated on an independent test dataset to confirm generalizability.

UNDERSTANDING THIRD-VARIABLE EFFECTS IN BIOMEDICAL RESEARCH

Predictive models in psychiatry aim to either elucidate neurobiological mechanisms or support clinical decision making. Both goals require generalizable models. However, third variables (Z) can influence the relationship of interest between features (X, e.g., brain imaging measures) and outcomes (Y, e.g., clinical phenotypes), potentially hindering generalizability. In biomedical and psychological research, where biological, behavioral, and environmental factors are tightly interwoven, third-variable effects are often unavoidable, as has been demonstrated in large-scale observational datasets such as the UK Biobank (44,45).

Third variables can act as confounders, colliders, or mediators, each affecting the feature-target relationship differently and therefore requiring distinct handling (Figure 1). Correlation-based criteria alone cannot distinguish between these types as all could produce the same correlation with both X and Y (46). Instead, cause-effect reasoning, often aided by directed acyclic graphs (DAGs), is needed for distinction (47,48).

A confounder is a common cause of both X and Y, biasing their relationship and the respective predictive model if not controlled for [confounder bias, Simpson's paradox (49)]. For example, early childhood trauma may confound the relationship between hippocampal volume (e.g., stress-induced increased glucocorticoid exposure reducing synapto- and neurogenesis) (50) and depression risk (e.g., via higher likelihood of unhealthy lifestyles) (51). In contrast, a collider is a common effect of X and Y. Controlling for a collider induces a spurious X-Y association, biasing the predictive model [collider bias, Berkson's paradox (52)] [e.g., (53,54)]. For example, depression can act as a collider in studies of serotonin receptor function (e.g., using positron emission tomography) and cortisol levels because depressive symptoms can result from both reduced 5-HT_{1A} receptor binding (55) and hypercortisolemia (hypothalamic-pituitary-adrenal axis dysfunction) (56). A mediator lies on the indirect causal path from X to Y, transmitting part of the effect (46). For example,

Table 1. Overview of the Different Challenges, Examples, Suggestions, and Recommended References

Challenge	Example	Recommendation	Key References
Model Evaluation			
Biased and Unstable CV Estimates	Performing k-fold CV in small samples (e.g., $N = 100$) or performing leave-one-out CV can inflate performance.	Use sample sizes that lead to reasonably sized inner CV data splits (e.g., if $N = 100$, in a 10-fold inner and outer CV, there would only be 1 sample left for testing in the inner CV; this is not reasonably sized). Avoid leave-one-out CV.	Varoquaux *et al.* (23)
Cherry-Picking of CV Results	Reporting results from 1) one (the best) CV fold, 2) train performances, or 3) the best-looking error metric.	Report performance mean and SD across folds. Report test set errors. Use multiple error metrics.	Komiyama and Maehara (36) Demsar *et al.* (111)
Confounding/Third Variables			
Biased and Biologically Misleading Models Due to Confounding	A model falsely attributes structural brain changes to schizophrenia when in reality these changes are (partially) driven by aging or long-term medication use.	Use DAGs to systematize variable relationships around the research question of interest to select proper adjustment variables, e.g., through the so-called backdoor criterion. Avoid using default research question agnostic confounders such as age or sex but communicate informed decisions transparently.	Pearl *et al.* (58) Wysocki *et al.* (46) Komeyer *et al.* (27) Rohrer *et al.* (59) Pearl and Mackenzie (61) VanderWeele *et al.* (60) Pearl *et al.* (58)
Incorrect Adjustment for a Collider	A variable is identified as confounder based on correlations but is actually a collider, so its adjustment introduces bias.	Use DAGs to make variable relationships transparent and identify appropriate deconfounding variables.	Wysocki *et al.* (46)
Adjustment for a Default Set of Variables	Default adjustment for demographics, such as age and sex, without further consideration of variable relationships.	Use literature and domain knowledge to arrive at relevant variables and model their relationships using a DAG.	Pearl and Mackenzie (61) Komeyer *et al.* (27)
Data Harmonization			
Separate Train-Test Splits When Harmonizing Data to Avoid Data Leakage	Original proposed ComBat finds its parameters on the whole datasets, which is only compatible with classical statistical analysis, but not with ML studies, where separated train and test sets are needed.	When integrating in ML pipelines, use newer versions of ComBat that allow separation of train-test, such as neuroHarmonize, harmonizer, and ComBat-MEGA.	Fortin *et al.* (76) Marzi *et al.* (90) Radua *et al.* (31) Hu *et al.* (84)
Expected Nonlinear Covariate Effects	Biological information, for example related to age, often presents nonlinear effects that traditional ComBat cannot model.	Allows for estimation of more complex covariate effects. neuroHarmonize allows for this flexible covariate effect estimation.	Fortin *et al.* (76)
Site-Target Relationship	Data acquired at each site may have different proportions of classes, for example patients and control participants. ComBat-based methods may require test labels to correctly harmonize without removing relevant information.	Estimate the degree of site-target dependence and use leakage-free harmonization models such as PrettyHarmonize, which do not need test targets to correctly harmonize.	Nieto *et al.* (92)
Estimating Number of Images per Site to Train the Harmonization Models	There is a minimum number of images for each site that are needed to correctly estimate the parameters of the models.	The required N is a function of the number of sites, number of features, and intrinsic characteristics of the problem. The Mahalanobis distance was proposed to quantify the multivariate site effect and estimate the minimum N. For ComBat, between 20 and 30 samples per site are needed. ComBat-based methods are recommended with a low number of images, in contrast to DL-based models, which require more data.	Parekh *et al.* (112)

Table 1. Continued

Challenge	Example	Recommendation	Key References
Harmonization on Unseen Sites	We aim to harmonize new data that were acquired in a new site that was not used for training the harmonization model.	ComBat is not able to harmonize data from sites that were not included at training time. NeuroHarmony relies on IQMs instead of site ID; thus, it can be applied to any image where the IQM can be extracted, and the obtained IQMs are in the range of the training images. DL methods can also harmonize data from unseen data.	García-Dias *et al.* (113) Abbasi *et al.* (93)
Uncompleted Effect of Site Removal	Effects of site can be due to complex interaction in the data. Some of the methods may partially remove the effects of site and lead to bias estimations.	Evaluate harmonization models to validate its capacity to remove the effect of site. Alternatively, leave-one-site-out CV can help to evaluate robustness and generalization of the models.	Solanes *et al.* (29)
Uneven Classification Prediction Across Sites	When most of the data come from 1 site, ML can underperform in smaller sites.	Report several metrics and perform separate metrics for each site. Calculate multisite-specific metrics.	Solanes *et al.* (74)
Covariance Harmonization	ComBat can only correct mean and variance but cannot correct covariance.	CovBat is recommended in those cases, as it is specifically designed to harmonize mean, variance, and covariance.	Chen *et al.* (91)
Not Possible to Access Raw Data From All the Sites	Privacy-preserving scenarios are common in medical applications. Access to raw data is not always possible.	Distributed ComBat demonstrated similar performance as ComBat without direct access to the raw data.	Chen *et al.* (114)
Repeated Measurements for Each Participant	When monitoring neurodegenerative diseases, such as Parkinson's, several images from the same participant may be acquired.	When repeated measures are available, LongitudinalComBat is recommended.	Beer *et al.* (115)
Result Interpretation			
Feature Importance Misinterpretation	SHAP shows age as the most important feature in a depression classifier—is this meaningful?	Contextualize feature importance (e.g., age may be a confounder). Use domain-specific knowledge to interpret. Do not confuse true to the model with true to the data, and do not confuse feature importance with causal explanations.	Chen *et al.* (116) Molnar (103)
Overstated Importance (Meaningless Explanations)	Gray matter volume is the most important feature. Model accuracy is almost at chance level.	Assess model's performance using multiple complementary error metrics (e.g., AUROC and balanced accuracy). Report model accuracy alongside interpretations. Contextualize interpretation of the model's accuracy.	Molnar (103)

AUROC, area under the receiver operating characteristic curve; CV, cross-validation; DAG, directed acyclic graph; DL, deep learning; IQM, image quality metric; ML, machine learning; SHAP, Shapley Additive exPlanations.

cortisol levels may mediate (indirect path) the direct effect of amygdala hyperactivity (e.g., measured through functional magnetic resonance imaging [MRI]) on depressive symptoms (56,57). Whether to adjust for mediators depends on whether the partial direct effect (amygdala hyperactivity→depressive symptoms, control for mediator) or the full effect including the indirect pathway via cortisol (do not control) is being sought (Figure 1).

To build valid models, researchers must account for third-variable types and mitigate bias accordingly. While simplified DAGs (Figure 1) illustrate basic principles, real-world neurobiological applications often involve complex interdependencies between many variables. To ensure unbiased predictive models, confounders must be controlled, while making sure not to control for colliders. However, in practice, confounder selection often lacks transparency or is based on default variables (e.g., demographics), increasing the risk of inadvertent effects, e.g., through collider adjustment.

A 3-step approach can help identify which variables to correct for. First, using literature-derived and clinical knowledge to build a DAG around the relationship of interest can clarify variable roles (types of third variables) and communicate assumptions transparently. Specifically, this DAG aids in identifying confounding pathways and therefore a correct set of variables to control for in the second step, for example through tools such as the backdoor criterion. The backdoor criterion originates in the causal literature (58) and states that to estimate a causal effect of variable X on outcome Y, it is necessary to block the so-called backdoor paths, i.e., paths

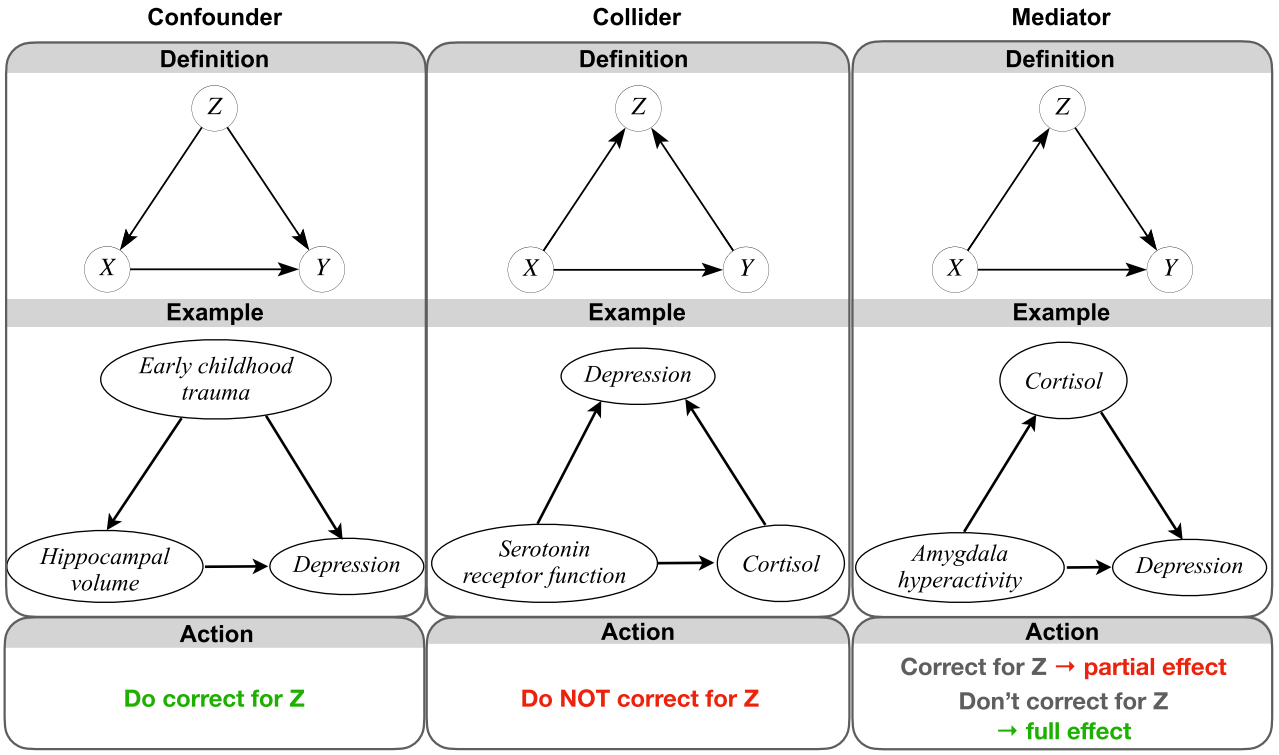

Figure 1. Definition of the different types of third variables, confounder, collider, and mediator using a directed acyclic graph. Each theoretical definition is supported by a simple biomedical example, and a recommendation for handling the respective type of third variable (action) is given.

with noncausal flow of information. Transferred to the field of associative predictive modeling, this translates to adjusting for variables that lie on indirect paths between X and Y with incoming arrows to X in the previously specified DAG (step 1). Online tools such as Dagitty (https://www.dagitty.net/dags.html) can support identifying biasing paths [see e.g., (27,48,58–61) for in-depth information]. Third, the identified variables should be checked for their statistical association with both X and Y, after which the actual model adjustment process can follow standard approaches from the ML literature [e.g., (26,62,63)].

Failing to adjust for confounders can lead ML models to learn spurious associations, capturing dataset-specific artifacts rather than neurobiologically meaningful patterns. This makes proper confounder selection and handling essential. In causal inference and treatment-outcome modeling, confounders are explicitly adjusted for to estimate causal effects, whereas in ML, they need to be considered to avoid unwanted bias or shortcuts that degrade generalizability of models and results. For example, in psychiatric research, age, medication use, and comorbidities can result in misleading associations between imaging-derived features and diagnostic labels. For example, a model may erroneously attribute structural brain changes to, e.g., schizophrenia when, in reality, these changes are (partially) driven by aging or long-term medication use. Likewise, comorbid conditions (e.g., anxiety, substance use disorder) can influence both brain imaging features and psychiatric diagnoses, making it difficult to disentangle disorder-specific neural signatures from overlapping, but distinct, effects.

Once identified, several post hoc confounder mitigation methods exist, each with implications and tradeoffs. Residualization removes confounder influences by often univariate linear regression of the confounder on features or target with subsequent residualization (true minus predicted feature/target) [e.g., (64)] but may leave nonlinear or multivariate confounding unaddressed and can even leak confounding signal into features or target (65). Matching balances distributions of confounding variables across groups or classes, thereby conditioning on the confounders and mitigating bias [e.g., (66)]. However, matching is data inefficient as unmatched samples are discarded and becomes increasingly complex with multiple confounders. Matching should not be confused with stratified CV [e.g., *StratifiedKFold* (67)], which ensures comparable distributions (e.g., of the target) between data splits but does not break confounder-feature/target associations, making it ineffective for mitigating confounding bias [e.g., (68)]. Including confounders as features/covariates can improve model performance but can reduce generalizability if confounder distributions shift across datasets (64) and does not give insights into feature-target mechanisms because these will be distorted by the confounders' influence. Post hoc tests [e.g., partial and full confounder tests (69)] can help assess model reliance on confounders but do not correct for them.

Thoughtful investigation and integration of third-variable structures is essential for unbiased models that allow for valid, generalizable, and interpretable ML research in psychiatry. Unbiased models require deliberate confounder control,

while improper adjustments (e.g., for colliders) must be avoided. Although confounder control may reduce apparent performance, it yields more meaningful and replicable results. Once appropriate confounders are identified, established confounder control strategies can be applied. DAGs provide a transparent framework for highly interwoven biomedical data to identify and justify adjustment variables by clarifying causal roles. Therefore, future work should move beyond default confounders (e.g., age, sex) toward question-specific, DAG-informed adjustments. Making this a standard practice in brain-based association studies will promote more generalizable models with greater psychiatric and clinical relevance.

EFFECTS OF SITE AND DATA HARMONIZATION

The acquisition of brain imaging data has expanded, greatly driven by advances in neuroimaging technologies. Detecting brainwide associations requires large samples for statistical and predictive analyses (28,70). Open science initiatives have facilitated this by making numerous datasets publicly available (71). The use of multisite data may also improve generalization by capturing biologically and demographically diverse samples (72). However, because collecting datasets is time- and resource intensive, pooling data from multiple sites has become common practice. While this approach offers great potential for advancing empirical neuroscience and predictive modeling, it also introduces new challenges, because systematic differences across sites can bias the resulting models.

Site-related data variability can stem from 2 main sources: acquisition differences and population differences. Variability due to acquisition is primarily driven by factors such as differences in scanners, imaging protocols, acquisition parameters, target definitions, or measurement procedures, none of which are related to biological signals (73). When this unwanted variability only affects the features (X) it is referred to as effects of site (Figure 2A) and can introduce bias into research outcomes if not properly identified or inadequately addressed (29,74). For example, MR images acquired from 2 scanners from the same manufacturer with the same parameters can differ (75), which extends to imaging-derived features (76) and nuances in image processing pipelines (77). Additionally, target can be also influenced by site due to sampling bias, differences in acquisition instruments, or different target definitions at each site/study (20). For example, there are different criteria in the Alzheimer's disease stages in the available datasets (78–80). These cases of different target definitions are also present in schizophrenia (20) and depression (81).

Beyond measurement-related variance, each site may recruit diverse populations, introducing sampling bias tied to the site's location and demographic reach. This may result in differences in diets, genetics, environmental factors, or socioeconomic factors, which are correlated with brain characteristics (82). Different sites may also recruit different target distributions, e.g., healthy control participants recruited at one site and patients at another.

If site affects the features, its role as a confounder depends on whether it also influences the target (Figure 2). When site does not affect the target, it acts as systematic noise, masking the biological signal (Figure 2A). In this case, removing site effects from features can improve the signal-to-noise ratio, thereby aiding meaningful learning and robustness (83). However, if site also influences the target, it becomes a confounder, enabling models to achieve high accuracy by exploiting nonbiological site information (Figure 2B) (see previous section).

Harmonization methods aim to eliminate site-specific variability while preserving signals of interest. When properly applied, they enhance statistical power, generalizability, and interpretability (73,84–87). These methods can be broadly classified into statistical approaches, primarily based on ComBat, and deep learning (DL) methods. It should be noted that data harmonization has different meanings across fields. For example, in psychology it refers to aligning different textual expressions to a common, semantically equivalent form (88). These fall outside the scope of this discussion. Finally, although data harmonization can provide appealing advantages, it may not be beneficial or even detrimental for some tasks (89).

ComBat is a commonly used harmonization method in statistical analyses and is the core of other proposed methods. ComBat was developed for genomics and later adapted for neuroimaging (76). Initially, ComBat estimates its parameters using the entire dataset, which is appropriate for statistical analysis but conflicts with ML principles—specifically, it violates the separation between training and test data, leading to data leakage (42,90). Extensions of ComBat allow for train-test separation (31,76,90). Some of the most prominent of ComBat's limitations are that it assumes that all features are in the same range, sites have similar numbers of images (at least 20), and the variance is equally distributed across sites. Additionally, it cannot correct features covariance (91), and it cannot be applied to data from an unseen site. Fortunately, several methods have been proposed to overcome these limitations (see Table 1). Additionally, the method struggles if site is a confounder (Figure 2B), as it assumes that any nonshared variance across sites is undesirable and could remove target variance, unless the target is preserved by specifying it as a covariate. However, this requires knowing the target value at test time, introducing leakage and precluding real-world application (92). Alternatively, normative modeling has also been proposed for harmonization, with the main difference being that the site effects are not estimated and removed, but the data are normalized instead (73).

DL-based harmonization methods offer a more flexible data-driven approach (93) by leveraging different ML architectures (94–97). DL approaches do not explicitly make assumptions about the nature of the site effects and can be applied at the image or feature level and can harmonize data from unseen sites. However, they require a substantial amount of training data. Finally, phantoms or traveling subjects allow training harmonization models without mixing biological and site variance (98), but it is inefficient and costly (99,100).

In summary, data harmonization has become a fundamental step in large-scale neuroimaging analysis. While acquisition sites mainly affect the features through instrumental factors, it can also affect the targets. It is essential to adapt the harmonization approach to the specific context of the study; in classical statistical studies, harmonization should

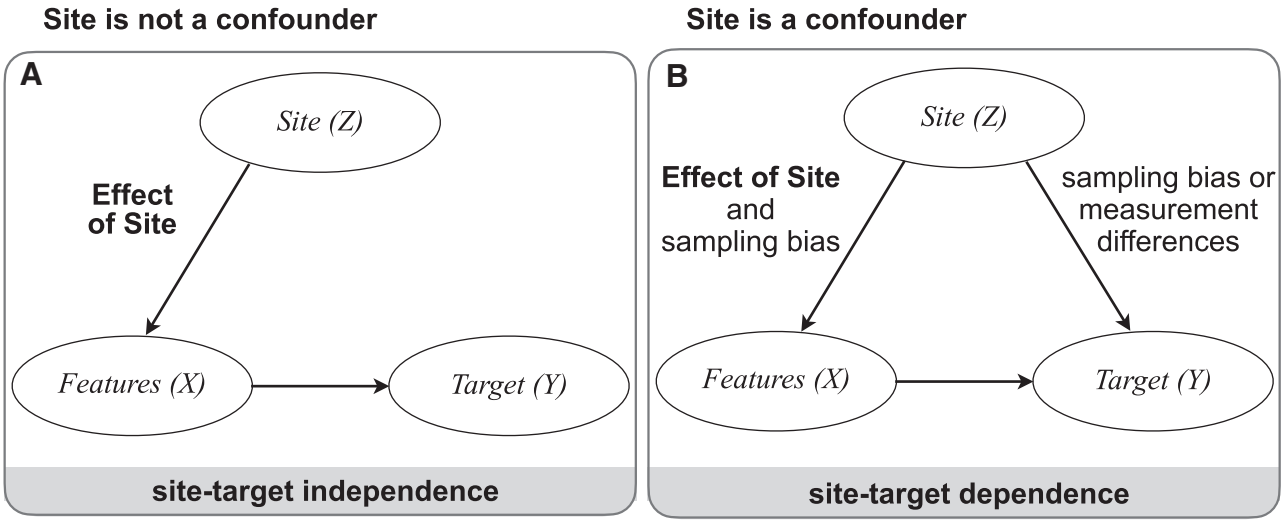

Figure 2. Different possible scenarios of the impact of site on features and target. Depending on the influence of site on the target, site does not act as a confounder **(A)** or does act as a confounder **(B)**.

prioritize the removal of site-specific biases to increase statistical power. In ML applications, extra considerations must be taken to correctly integrate harmonization methods in ML pipelines to avoid data leakage, unintentional removal of relevant signal, and ungeneralizable results. Choosing an appropriate harmonization method critically depends on the problem at hand, the data type (structural, functional or diffusion MRI), and the overarching research question. Decisions such as harmonizing on voxel or feature level, availability of traveling subjects or longitudinal data, and application on unseen sites are just a few examples of considerations important to identify a suitable harmonization method. While basic recommendations are presented in Table 1, please see (73,84,93) for a detailed discussion.

POST HOC MODEL INTERPRETATION

In clinical and research applications, high predictive accuracy alone is insufficient; understanding how models arrive at their predictions is equally important. A model must be evaluated beyond predictive accuracy (101), because it may incorrectly rely on site-specific factors or spurious third-variable associations, such as predicting improved myocardial infarction recovery in smokers due to age-related confounding (i.e., smokers are younger with a higher chance of recovery) (102). In psychiatry, an example of a common pitfall is the misattribution of brain-related differences to a disorder rather than medication effects. Model interpretability can help detect such wrong associations. Beyond determining what a model predicts, model interpretation can clarify why it reached that conclusion, ensuring meaningful and clinically sound findings (103). By examining the model's decision process, researchers can check whether it is consistent with biological and clinical knowledge before interpreting them as novel biomarkers.

Interpretability can be achieved either by model-specific (by design) or model-agnostic methods (103). By-design interpretability uses inherently interpretable algorithms (e.g., decision trees, linear models) as they rely on internal representations learned by the selected algorithm. For example, decision trees reveal exactly how decisions were made, or linear regression weights, particularly when Haufe-transformed (104,105), provide reliable feature importances. While informative, model-specific approaches restrict algorithm choice and may be too simple for capturing complex data patterns, potentially compromising predictive performance. Nonetheless, they can be a suitable choice if they align with the assumed variable relationships in the data. Model-agnostic interpretation, in contrast, can be applied to any model and must always be performed post hoc. In practice, psychiatric research questions likely will benefit from more complex models; thus, we will focus on challenges related to model-agnostic interpretation.

Consider a study yielding a successful model that accurately predicts unseen data. The next step is to determine how the model used the features to predict the target variable, i.e., to uncover the internal rules governing the model's decision-making process. Model-agnostic approaches quantify the contribution of each feature to the model's prediction by performing predictions on perturbed input data and comparing the outcomes (106). For example, permutation feature importance (107) randomly permutes a feature (or set of features) to break potential relationships with the target, compares performances between original (unpermuted) and permuted features, and the diverging loss in performance is considered the feature's contribution to the model's prediction. A widely adopted alternative is Shapley Additive exPlanations (SHAP) (108), based on Shapley values (109), which decomposes the predictions into additive feature contributions. SHAP captures the relative importance of a feature in driving a model's decisions and is applicable to a variety of model types by providing respectively suitable explainers [e.g., Molnar (103)].

Nonetheless, 2 caveats are critical when interpreting insights obtained through post hoc model interpretation. First,

feature importance is always assessed in a multivariate context, i.e., the interpretation depends on the relationships and interactions among all variables. A feature may appear important because it provides information not captured by any other variable. However, this does not necessarily imply that the feature is informative about the target variable in isolation. Conversely, a feature may appear unimportant because its information is shared with other variables (multicollinearity). In brain-related and psychiatric research, multicollinearities are common (e.g., among neighboring brain areas or between age, medication use, and illness duration). In such cases, interpreting feature groups rather than individual features is more meaningful. Usage of Owen values instead of SHAP values can be a suitable solution (110). Second, and critically, these methods estimate the contribution of features to a model's decision but do not assess whether the decision itself is correct. A feature may strongly influence a prediction, yet the prediction could still be incorrect. The insights from explanation methods are inherently tied to the model's predictive performance, reflecting only the aspects of the domain that contribute beyond chance-level predictions. As feature importance reliability is correlated with prediction accuracy (104), feature importance results should always be presented alongside performance metrics and not be interpreted as clinically meaningful when performance is near chance level.

Taken together, model interpretation is as essential as achieving high accuracy when applying ML approaches. Importantly, interpretation methods should be applied only after confirming that the model performs above chance level. Unlike hypothesis testing with defined significance thresholds, no standard defines how much above chance is enough to be considered successful. This ambiguity, in combination with the multivariate nature of predictive models, means that as readers, we must interpret feature importances while carefully considering the model's limitations (also see CV Is an Estimation). Generally, model-derived feature importances should be related to existing domain knowledge to avoid overinterpretating spurious results.

CONCLUSIONS

While ML provides powerful tools for neuroimaging-based decision making in psychiatry, evaluating the generalizability and reliability of such models demands rigorous scrutiny. CV can inflate performance estimates, and findings may be distorted by confounding variables, site-specific biases, or spurious correlations that misrepresent true brain-behavior associations. Importantly, the patterns uncovered by ML models reflect the statistical structure of the data, not necessarily causal or biologically meaningful mechanisms. To draw valid inferences, researchers must adopt robust methodological practices such as identifying and controlling for key confounders, applying proper harmonization techniques, using independent validation datasets, and predefining analysis pipelines.

Tools such as SHAP can support interpretation by highlighting which features contribute to predictions, but these outputs must be considered in light of the model's overall performance—evaluated using clinically relevant metrics such as accuracy, sensitivity, specificity, area under the receiver operating characteristic curve, and precision—and the context in which the data were collected. Where appropriate, decision-curve analysis and confusion matrices can help assess the practical utility of model-guided decisions in psychiatric settings.

Ultimately, we underline that ML should not be viewed as a shortcut to understanding complex brain-behavior associations but rather as an approach that, when used carefully, can generate testable hypotheses and clinically relevant insights. We encourage both researchers and clinicians to interpret ML findings with a critical eye, weighing methodological transparency, validation rigor, and clinical plausibility to ensure that conclusions reflect meaningful and generalizable relationships, not statistical artifacts.

ACKNOWLEDGMENTS AND DISCLOSURES

This work was supported by the Helmholtz Imagining grant BrainShapes (Grant No. ZT-I-PF-4-062 [to KRP]); the Multi-Omics Data Science project was funded from the program Profilbildung 2020 (Grant No. PROFILNRW-2020-107-A [to SBE]), an initiative of the Ministry of Culture and Science of the State of North Rhine-Westphalia; the H2020 Research Infrastructures (Grant No. EBRAIN-Health 101058516 [to SBE]); the Deutsche Forschungsgemeinschaft Collaborative Research Centre CRC1451 (Project No. 431549029 [to SBE]) on motor performance project B05; and the Universitätsklinikum Düsseldorf, Forschungskommission funded project VoxNorm [to KRP].

We thank Dr. Athena Demertzi for their insightful feedback and assistance in improving the clarity and relevance of this article to better align with the target readership of *Biological Psychiatry*.

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

ARTICLE INFORMATION

From the Institute of Neuroscience and Medicine, Brain and Behavior, Forschungszentrum Jülich, Jülich, Germany (VK, NN, SBE, FR, KRP); Institute for Systems Neuroscience, Medical Faculty, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany (VK, NN, SBE, FR, KRP); Department of Biology, Faculty of Mathematics and Natural Sciences, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany (VK); and Institute of Diagnostic and Interventional Radiology, University Hospital Düsseldorf, Düsseldorf, Germany (VK).

FR and KRP contributed equally to this work.

Address correspondence to Federico Raimondo, Ph.D., at f.raimondo@fz-juelich.de.

Received Apr 1, 2025; revised Aug 29, 2025; accepted Sep 7, 2025.

REFERENCES

1. Wolfers T, Buitelaar JK, Beckmann CF, Franke B, Marquand AF (2015): From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. Neurosci Biobehav Rev 57:328–349.
2. 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.
3. Rutledge RB, Chekroud AM, Huys QJ (2019): Machine learning and big data in psychiatry: Toward clinical applications. Curr Opin Neurobiol 55:152–159.
4. Lucasius C, Ali M, Patel T, Kundur D, Szatmari P, Strauss J, Battaglia M (2025): A procedural overview of why, when and how to use machine learning for psychiatry. Nat Mental Health 3:8–18.
5. Bzdok D, Meyer-Lindenberg A (2018): Machine learning for precision psychiatry: Opportunities and challenges. Biol Psychiatry Cogn Neurosci Neuroimaging 3:223–230.

6. Milham MP, Craddock RC, Son JJ, Fleischmann M, Clucas J, Xu H, *et al.* (2018): Assessment of the impact of shared brain imaging data on the scientific literature. Nat Commun 9:2818.
7. Singh NM, Harrod JB, Subramanian S, Robinson M, Chang K, Cetin-Karayumak S, *et al.* (2022): How machine learning is powering neuroimaging to improve brain health. Neuroinformatics 20:943–964.
8. Giehl K, Mutsaerts H-J, Aarts K, Barkhof F, Caspers S, Chetelat G, *et al.* (2024): Sharing brain imaging data in the Open Science era: How and why? Lancet Digit Health 6:e526–e535.
9. Chen J, Patil KR, Yeo BTT, Eickhoff SB (2023): Leveraging machine learning for gaining neurobiological and nosological insights in psychiatric research. Biol Psychiatry 93:18–28.
10. Abraham A, Milham MP, Di Martino A, Craddock RC, Samaras D, Thirion B, Varoquaux G (2017): Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. Neuroimage 147:736–745.
11. Wilkinson J, Arnold KF, Murray EJ, van Smeden M, Carr K, Sippy R, *et al.* (2020): Time to reality check the promises of machine learning-powered precision medicine. Lancet Digit Health 2:e677–e680.
12. Chen J, Patil KR, Weis S, Sim K, Nickl-Jockschat T, Zhou J, *et al.* (2020): Neurobiological divergence of the positive and negative schizophrenia subtypes identified on a new factor structure of psychopathology using non-negative factorization: An international machine learning study. Biol Psychiatry 87:282–293.
13. Chen J, Müller VI, Dukart J, Hoffstaedter F, Baker JT, Holmes AJ, *et al.* (2021): Intrinsic connectivity patterns of task-defined brain networks allow individual prediction of cognitive symptom dimension of schizophrenia and are linked to molecular architecture. Biol Psychiatry 89:308–319.
14. Gallo S, El-Gazzar A, Zhutovsky P, Thomas RM, Javaheripour N, Li M, *et al.* (2023): Functional connectivity signatures of major depressive disorder: Machine learning analysis of two multicenter neuroimaging studies. Mol Psychiatry 28:3013–3022.
15. Winter NR, Blanke J, Leenings R, Ernsting J, Fisch L, Sarink K, *et al.* (2024): A systematic evaluation of machine learning-based biomarkers for major depressive disorder. JAMA Psychiatry 81:386–395.
16. Van Essen DC, Ugurbil K, Auerbach E, Barch D, Behrens TEJ, Bucholz R, *et al.* (2012): The Human connectome Project: A data acquisition perspective. Neuroimage 62:2222–2231.
17. Chen J, Müller V, Hoffstaedter F, Nickl-Jockschat T, Derntl B, Kogler L, *et al.* (2020): Linking schizophrenia symptom dimensions to neuro-cognitive processes by multivariate pattern prediction. Biol Psychiatry 87:S408–S409.
18. Wen J, Antoniades M, Yang Z, Hwang G, Skampardoni I, Wang R, Davatzikos C (2024): Dimensional neuroimaging endophenotypes: Neurobiological representations of disease heterogeneity through machine learning. Biol Psychiatry 96:564–584.
19. Omidvarnia A, Sasse L, Larabi DI, Raimondo F, Hoffstaedter F, Kasper J, *et al.* (2024): Individual characteristics outperform resting-state fMRI for the prediction of behavioral phenotypes. Commun Biol 7:771.
20. Chekroud AM, Hawrilenko M, Loho H, Bondar J, Gueorguieva R, Hasan A, *et al.* (2024): Illusory generalizability of clinical prediction models. Science 383:164–167.
21. Milham MP, Vogelstein J, Xu T (2021): Removing the reliability bottleneck in functional magnetic resonance imaging research to achieve clinical utility. JAMA Psychiatry 78:587–588.
22. Gell M, Eickhoff SB, Omidvarnia A, Küppers V, Patil KR, Satterthwaite TD, *et al.* (2024): How measurement noise limits the accuracy of brain-behaviour predictions. Nat Commun 15:10678.
23. Varoquaux G (2018): Cross-validation failure: Small sample sizes lead to large error bars. Neuroimage 180:68–77.
24. Li J, Bzdok D, Chen J, Tam A, Ooi LQR, Holmes AJ, *et al.* (2022): Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. Sci Adv 8:eabj1812.
25. Görgen K, Hebart MN, Allefeld C, Haynes J-D (2018): The Same Analysis Approach: Practical protection against the pitfalls of novel neuroimaging analysis methods. Neuroimage 180:19–30.
26. Rao A, Monteiro JM, Mourao-Miranda J, Alzheimer's Disease Initiative (2017): Predictive modelling using neuroimaging data in the presence of confounds. Neuroimage 150:23–49.
27. Komeyer V, Eickhoff SB, Rathkopf C, Grefkes C, Patil KR, Raimondo F (2024): Correct deconfounding enables causal machine learning for precision medicine and beyond. medRxiv https://doi.org/10.1101/2024.09.20.24314055.
28. Schulz M-A, Bzdok D, Haufe S, Haynes J-D, Ritter K (2024): Performance reserves in brain-imaging-based phenotype prediction. Cell Rep 43:113597.
29. Solanes A, Gosling CJ, Fortea L, Ortuño M, Lopez-Soley E, Llufriu S, *et al.* (2023): Removing the effects of the site in brain imaging machine-learning—Measurement and extendable benchmark. Neuroimage 265:119800.
30. Saponaro S, Giuliano A, Bellotti R, Lombardi A, Tangaro S, Oliva P, *et al.* (2022): Multi-site harmonization of MRI data uncovers machine-learning discrimination capability in barely separable populations: An example from the ABIDE dataset. Neuroimage Clin 35:103082.
31. Radua J, Vieta E, Shinohara R, Kochunov P, Quidé Y, Green MJ, *et al.* (2020): Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. Neuroimage 218:116956.
32. Ehsan U, Passi S, Liao QV, Chan L, Lee I-H, Muller M, Riedl MO (2024): The who in XAI: How AI background shapes perceptions of AI explanations. In: Proceedings of the Chi Conference on Human Factors in Computing Systems. New York, NY: ACM, 1–32.
33. Bengio Y, Grandvalet Y (2004): No unbiased estimator of the variance of K-fold cross-validation. J Mach Learn Res 5:1089–1105.
34. Ng AY (1997): Preventing "overfitting" of cross-validation data. In: Proceedings of the Fourteenth International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann, 245–253.
35. Giles CL, Lawrence S (1997): Presenting and analyzing the results of AI experiments: Data averaging and data snooping. In: Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence. Providence, RI: AAAI Press, 362–367.
36. Komiyama J, Maehara T (2018): A simple way to deal with Cherry-picking. arXiv https://doi.org/10.48550/arXiv.1810.04996.
37. Bates S, Hastie T, Tibshirani R (2024): Cross-validation: What does it estimate and how well does it do it? J Am Stat Assoc 119:1434–1445.
38. Sun D, Gao W, Hu H, Zhou S (2022): Why 90% of clinical drug development fails and how to improve it? Acta Pharm Sin B 12:3049–3062.
39. Paleyes A, Urma R-G, Lawrence ND (2023): Challenges in deploying machine learning: A survey of case studies. ACM Comput Surv 55:1–29.
40. Beyer L, Hénaff OJ, Kolesnikov A, Zhai X, van den Oord A (2020): Are we done with ImageNet? arXiv https://doi.org/10.48550/arXiv.2006.07159.
41. Thompson WH, Wright J, Bissett PG, Poldrack RA (2020): Dataset decay and the problem of sequential analyses on open datasets. eLife 9:e53498.
42. Sasse L, Nicolaisen-Sobesky E, Dukart J, Eickhoff SB, Götz M, Hamdan S, *et al.* (2024): On leakage in machine learning pipelines. arXiv https://doi.org/10.48550/arXiv.2311.04179.
43. Nadeau C, Bengio Y (2003): Inference for the generalization error. Mach Learn 52:239–281.
44. Miller KL, Alfaro-Almagro F, Bangerter NK, Thomas DL, Yacoub E, Xu J, *et al.* (2016): Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat Neurosci 19:1523–1536.
45. Carey CE, Shafee R, Wedow R, Elliott A, Palmer DS, Compitello J, *et al.* (2024): Principled distillation of UK Biobank phenotype data reveals underlying structure in human variation. Nat Hum Behav 8:1599–1615.
46. Wysocki AC, Lawson KM, Rhemtulla M (2022): Statistical control requires causal justification. Adv Methods Pract Psychol Sci 5:25152459221095823.

47. Maxwell SE, Cole DA (2007): Bias in cross-sectional analyses of longitudinal mediation. Psychol Methods 12:23–44.
48. Pearl J (1995): Causal diagrams for empirical research. Biometrika 82:669–688.
49. Sprenger J, Weinberger N (2021): Simpson's paradox. Stanf Encycl Philos. Available at: https://plato.stanford.edu/entries/paradox-simpson/?ref=praxarchy.com. Accessed March 14, 2025.
50. Humphreys KL, King LS, Sacchet MD, Camacho MC, Colich NL, Ordaz SJ, *et al.* (2019): Evidence for a sensitive period in the effects of early life stress on hippocampal volume. Dev Sci 22:e12775.
51. Wang X, Cao Z, Yin S, Duan T, Sun T, Xu C (2025): Childhood maltreatment and depression: Mediating role of lifestyle factors, personality traits, adult traumas, and social connections among middle-aged and elderly participants. BMC Med 23:319.
52. Berkson J (1946): Limitations of the application of 4-fold table analysis to hospital data. Biometrics 2:47–53.
53. Elwert F, Winship C (2014): Endogenous selection bias: The problem of conditioning on a collider variable. Annu Rev Sociol 40:31–53.
54. Tönnies T, Kahl S, Kuss O (2022): Collider bias in observational studies. Dtsch Ärztebl Int 119:107–122.
55. Drevets WC, Thase ME, Moses-Kolko EL, Price J, Frank E, Kupfer DJ, Mathis C (2007): Serotonin-1A receptor imaging in recurrent depression: Replication and literature review. Nucl Med Biol 34:865–877.
56. Pariante CM, Lightman SL (2008): The HPA axis in major depression: Classical theories and new developments. Trends Neurosci 31:464–468.
57. Ulrich-Lai YM, Herman JP (2009): Neural regulation of endocrine and autonomic stress responses. Nat Rev Neurosci 10:397–409.
58. Pearl J (2009): Causal inference in statistics: An overview. Statist Surv 3:96–146.
59. Rohrer JM (2018): Thinking clearly about correlations and causation: Graphical causal models for observational data. Adv Methods Pract Psychol Sci 1:27–42.
60. VanderWeele TJ (2019): Principles of confounder selection. Eur J Epidemiol 34:211–219.
61. Pearl J, Mackenzie D (2018): The New Science of Cause and Effect. New York, NY: Basic Books.
62. Dinga R, Schmaal L, Penninx BWJH, Veltman DJ, Marquand AF (2020): Controlling for effects of confounding variables on machine learning predictions. bioRxiv https://doi.org/10.1101/2020.08.17.255034.
63. Snoek L, Miletić S, Scholte HS (2019): How to control for confounds in decoding analyses of neuroimaging data. Neuroimage 184:741–760.
64. Chyzhyk D, Varoquaux G, Milham M, Thirion B (2022): How to remove or control confounds in predictive models, with applications to brain biomarkers. GigaScience 11:giac014.
65. Hamdan S, Love BC, von Polier GG, Weis S, Schwender H, Eickhoff SB, Patil KR (2022): Confound-leakage: Confound removal in machine learning leads to leakage. arXiv https://doi.org/10.48550/arXiv.2210.09232.
66. Wiersch L, Hamdan S, Hoffstaedter F, Votinov M, Habel U, Clemens B, *et al.* (2023): Accurate sex prediction of cisgender and transgender individuals without brain size bias. Sci Rep 13:13868.
67. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al.* (2012): Scikit-learn: Machine learning in python. J Mach Learn Res 12:2825–2830.
68. Hernan MA, Robins JM (2024): Causal Inference: What If, 1st ed. Boca Raton: Taylor & Francis.
69. Spisak T (2022): Statistical quantification of confounding bias in machine learning models. GigaScience 11:giac082.
70. 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.
71. Poldrack RA, Gorgolewski KJ (2014): Making big data open: Data sharing in neuroimaging. Nat Neurosci 17:1510–1517.
72. Ma Q, Zhang T, Zanetti MV, Shen H, Satterthwaite TD, Wolf DH, *et al.* (2018): Classification of multi-site MR images in the presence of heterogeneity using multi-task learning. Neuroimage Clin 19:476–486.
73. Bayer JMM, Thompson PM, Ching CRK, Liu M, Chen A, Panzenhagen AC, *et al.* (2022): Site effects how-to and when: An overview of retrospective techniques to accommodate site effects in multi-site neuroimaging analyses. Front Neurol 13:923988.
74. Solanes A, Palau P, Fortea L, Salvador R, González-Navarro L, Llach CD, *et al.* (2021): Biased accuracy in multisite machine-learning studies due to incomplete removal of the effects of the site. Psychiatry Res Neuroimaging 314:111313.
75. Li H, Smith SM, Gruber S, Lukas SE, Silveri MM, Hill KP, *et al.* (2020): Denoising scanner effects from multimodal MRI data using linked independent component analysis. Neuroimage 208:116388.
76. Fortin J-P, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA, *et al.* (2018): Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 167:104–120.
77. Antonopoulos G, More S, Raimondo F, Eickhoff SB, Hoffstaedter F, Patil KR (2023): A systematic comparison of VBM pipelines and their application to age prediction. Neuroimage 279:120292.
78. Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL (2007): Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J Cogn Neurosci 19:1498–1507.
79. Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack CR, Jagust W, *et al.* (2005): Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimers Dement 1:55–66.
80. Fowler C, Rainey-Smith SR, Bird S, Bomke J, Bourgeat P, Brown BM, *et al.* (2021): Fifteen years of the Australian imaging, biomarkers and lifestyle (AIBL) study: Progress and observations from 2,359 older adults spanning the spectrum from cognitive normality to Alzheimer's disease. J Alzheimers Dis Rep 5:443–468.
81. National Collaborating Centre for Mental Health (2010): The Classification of Depression and Depression Rating Scales/Questionnaires. Depression in Adults with a Chronic Physical Health Problem: Treatment and Management. Leicester: British Psychological Society. Available at: https://www.ncbi.nlm.nih.gov/books/NBK82926/. Accessed June 16, 2025.
82. Institute of Medicine (US) Committee on Assessing Interactions Among Social, Behavioral, and Genetic Factors in Health (2006): Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate. Washington, DC: National Academies Press.
83. Andrade C (2013): Signal-to-noise ratio, variability, and their relevance in clinical trials. J Clin Psychiatry 74:479–481.
84. Hu F, Chen AA, Horng H, Bashyam V, Davatzikos C, Alexander-Bloch A, *et al.* (2023): Image harmonization: A review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization. Neuroimage 274:120125.
85. Da-Ano R, Visvikis D, Hatt M (2020): Harmonization strategies for multicenter radiomics investigations. Phys Med Biol 65:24TR02.
86. Wang Y-W, Chen X, Yan C-G (2023): Comprehensive evaluation of harmonization on functional brain imaging for multisite data-fusion. Neuroimage 274:120089.
87. Nan Y, Ser JD, Walsh S, Schönlieb C, Roberts M, Selby I, *et al.* (2022): Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic review, meta-analysis and future research directions. Inf Fusion 82:99–122.
88. McElroy E, Wood T, Bond R, Mulvenna M, Shevlin M, Ploubidis GB, *et al.* (2024): Using natural language processing to facilitate the harmonisation of mental health questionnaires: A validation study using real-world data. BMC Psychiatry 24:530.
89. Yu Y, Cui HQ, Haas SS, New F, Sanford N, Yu K, *et al.* (2024): Brain-age prediction: Systematic evaluation of site effects, and sample age range and size. Hum Brain Mapp 45:e26768.
90. Marzi C, Giannelli M, Barucci A, Tessa C, Mascalchi M, Diciotti S (2024): Efficacy of MRI data harmonization in the age of machine learning: A multicenter study across 36 datasets. Sci Data 11:115.
91. Chen AA, Beer JC, Tustison NJ, Cook PA, Shinohara RT, Shou H, Alzheimer's Disease Neuroimaging Initiative (2022): Mitigating site

effects in covariance for machine learning in neuroimaging data. Hum Brain Mapp 43:1179–1195.
92. Nieto N, Eickhoff SB, Jung C, Reuter M, Diers K, Kelm M, *et al.* (2024): Impact of leakage on data harmonization in machine learning pipelines in class imbalance across sites. arXiv https://doi.org/10.48550/arXiv.2410.19643.
93. Abbasi S, Lan H, Choupan J, Sheikh-Bahaei N, Pandey G, Varghese B (2024): Deep learning for the harmonization of structural MRI scans: A survey. Biomed Eng OnLine 23:90.
94. Dewey BE, Zhao C, Reinhold JC, Carass A, Fitzgerald KC, Sotirchos ES, *et al.* (2019): DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. Magn Reson Imaging 64:160–170.
95. Torbati ME, Tudorascu DL, Minhas DS, Maillard P, DeCarli CS, Hwang SJ (2021): Multi-scanner harmonization of paired neuroimaging data via structure preserving embedding learning. IEEE Int Conf Comput Vis Workshops 2021 3277–3286.
96. Cackowski S, Barbier EL, Dojat M, Christen T (2023): ImUnity: A generalizable VAE-GAN solution for multicenter MR image harmonization. Med Image Anal 88:102799.
97. Komandur D, Gupta U, Chattopadhyay T, Dhinagar NJ, Thomopoulos SI, Chen J-C, *et al.* (2023): Unsupervised harmonization of brain MRI using 3D CycleGANs and its effect on brain age prediction. In: 2023 19th International Symposium on Medical Information Processing and Analysis (SIPAIM), 1–5.
98. Tian D, Zeng Z, Sun X, Tong Q, Li H, He H, *et al.* (2022): A deep learning-based multisite neuroimage harmonization framework established with a traveling-subject dataset. Neuroimage 257:119297.
99. Maikusa N, Zhu Y, Uematsu A, Yamashita A, Saotome K, Okada N, *et al.* (2021): Comparison of traveling-subject and ComBat harmonization methods for assessing structural brain characteristics. Hum Brain Mapp 42:5278–5287.
100. Ibrahim A, Primakov S, Beuque M, Woodruff HC, Halilaj I, Wu G, *et al.* (2021): Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. Methods 188:20–29.
101. Doshi-Velez F, Kim B (2017): Towards a rigorous science of interpretable machine learning. arXiv https://doi.org/10.48550/arXiv.1702.08608.
102. Chen K-Y, Rha S-W, Li Y-J, Jin Z, Minami Y, Park JY, *et al.* (2012): 'Smoker's paradox' in young patients with acute myocardial infarction. Clin Exp Pharmacol Physiol 39:630–635.
103. Molnar C (2018): Interpretable Machine Learning, 3rd ed. Leanpub. Available at: https://leanpub.next/interpretable-machine-learning. Accessed March 26, 2025.
104. Chen J, Ooi LQR, Tan TWK, Zhang S, Li J, Asplund CL, *et al.* (2023): Relationship between prediction accuracy and feature importance reliability: An empirical and theoretical study. Neuroimage 274:120115.
105. Haufe S, Meinecke F, Görgen K, Dähne S, Haynes J-D, Blankertz B, Bießmann F (2014): On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87:96–110.
106. Scholbeck CA, Molnar C, Heumann C, Bischl B, Casalicchio G (2020): Sampling, intervention, prediction, aggregation: A generalized framework for model-agnostic interpretations. In: Cellier P, Driessens K, editors. Machine Learning and Knowledge Discovery in Databases. Cham: Springer International Publishing, 205–216.
107. Breiman L (2001): Random forests. Mach Learn 45:5–32.
108. Lundberg SM, Lee S-I (2017): A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Advances in Neural Information Processing Systems, vol. 30. Red Hook, NY: Curran Associates, Inc. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.
109. Shapley LS (1953): 17. A value for n-person games. In: Kuhn HW, Tucker AW, editors. (1953), Contributions to the Theory of Games, II: Princeton: Princeton University Press, 307–318.
110. López S, Saboya M (2009): On the relationship between Shapley and Owen values. Cent Eur J Oper Res 17:415–423.
111. Demšar J, Zupan B (2021): Hands-on training about overfitting. PLoS Comput Biol 17:e1008671.
112. Parekh P, Vivek Bhalerao G, ADBS consortium, John JP, Venkatasubramanian G (2022): Sample size requirement for achieving multisite harmonization using structural brain MRI features. Neuroimage 264:119768.
113. Garcia-Dias R, Scarpazza C, Baecker L, Vieira S, Pinaya WHL, Corvin A, *et al.* (2020): Neuroharmony: A new tool for harmonizing volumetric MRI data from unseen scanners. Neuroimage 220:117127.
114. Chen AA, Luo C, Chen Y, Shinohara RT, Shou H, Alzheimer's Disease Neuroimaging Initiative (2022): Privacy-preserving harmonization via distributed ComBat. Neuroimage 248:118822.
115. Beer JC, Tustison NJ, Cook PA, Davatzikos C, Sheline YI, Shinohara RT, *et al.* (2020): Longitudinal ComBat: A method for harmonizing longitudinal multi-scanner imaging data. Neuroimage 220:117129.
116. Chen H, Janizek JD, Lundberg S, Lee S-I (2020): True to the model or true to the data? arXiv https://doi.org/10.48550/arXiv.2006.16234.