Identifiability in Functional Connectivity May Unintentionally Inflate Prediction Results

Anton Orlichenko[a], Gang Qu[a], Kuan-Jui Su[b], Anqi Liu[b], Hui Shen[b], Hong-Wen Deng[b], and Yu-Ping Wang[a]

[a]Department of Biomedical Engineering, Tulane University, New Orleans, LA, USA
[b]School of Medicine, Tulane University, New Orleans, LA, USA

ABSTRACT

Functional magnetic resonance (fMRI) is an invaluable tool in studying cognitive processes in vivo. Many recent studies use functional connectivity (FC), partial correlation connectivity (PC), or fMRI-derived brain networks to predict phenotypes with results that sometimes cannot be replicated. At the same time, FC can be used to identify the same subject from different scans with great accuracy. In this paper, we show a method by which one can unknowingly inflate classification results from 61% accuracy to 86% accuracy by treating longitudinal or contemporaneous scans of the same subject as independent data points. Using the UK Biobank dataset, we find one can achieve the same level of variance explained with 50 training subjects by exploiting identifiability as with 10,000 training subjects without double-dipping. We replicate this effect in four different datasets: the UK Biobank (UKB), the Philadelphia Neurodevelopmental Cohort (PNC), the Bipolar and Schizophrenia Network for Intermediate Phenotypes (BSNIP), and an OpenNeuro Fibromyalgia dataset (Fibro). The unintentional improvement ranges between 7% and 25% in the four datasets. Additionally, we find that by using dynamic functional connectivity (dFC), one can apply this method even when one is limited to a single scan per subject. One major problem is that features such as ROIs or connectivities that are reported alongside inflated results may confuse future work. This article hopes to shed light on how even minor pipeline anomalies may lead to unexpectedly superb results.

Keywords: fMRI, functional connectivity, identifiability, fingerprinting, replicability, UKB, PNC, BSNIP, OpenNeuro

1. INTRODUCTION

Functional magnetic resonance is a non-invasive imaging modality that uses the blood oxygen level dependent (BOLD) signal to infer the level of neural activity in different regions of the brain.[1] fMRI has been used to localize visual processing,[2] attention,[3,4] emotional processing,[5,6,7] and language[8] to specific locations in the cortex. It has also been used to identify hemispheric dominance for, e.g., language.[9] Functional connectivity is the Pearson correlation between the time-varying BOLD signal of different regions of the brain.[10] It has recently been used to predict age,[11,12] sex,[13,14] general fluid intelligence,[15,14] pre-clinical Alzheimer's disease,[16] and schizophrenia.[17,18] Classification based on 4D fMRI images, not FC, is also an active area of research.[19] Naturally, the ability to predict cognition-related endophenotypes or pre-clinical disease status is an exciting avenue for translational applications.

Although fMRI offers unmatched ability to observe neural activity in vivo in human subjects, there are two questions that must be addressed when interpreting the results of predictive studies. First, are these studies meant to establish a groundwork for a clinical system such as, e.g., an AI-based breast cancer screening tool?[20] If so, then these studies must be validated in a randomized trial with thousands of subjects.[21] By contrast, fewer than 1% of fMRI studies in 2017 and 2018 enrolled more than 100 subjects, with most recruiting less than 30.[22] A large number of subjects is needed partly because, in the past, fMRI has faced several replicability crises.[23] For example, Bennett et al. (2010) made the case that multiple comparison correction was indispensable in

Further author information: (Send correspondence to Anton Orlichenko)
Anton Orlichenko: E-mail: aorlichenko@tulane.edu

fMRI by revealing emotion-associated voxels in a dead salmon.[24] It is suspicious that fMRI-based predictions of schizophrenia status achieve above 90% accuracy with fewer than 100 training subjects,[17,18] whereas genome-wide association studies find single-nucleotide polymorphisms (SNPs) explain only 23% of schizophrenia variance,[25] and prediction studies based on SNPs in the UK Biobank report a maximum AUC of 0.71.[26] In fact, recent studies[27] and many recent posters at OHBM 2023 give a classification accuracy for schizophrenia diagnosis using FC (often times cited as the best metric) of 70-80%.[28,29,30]

If these pipelines are not meant to be introduced clinically, are they meant to provide mechanistic insights into human cognition? This is more likely to be the case, but there is sometimes a very loose interpretation of what FC actually is. For example, the UKB description of fMRI processing[31] makes the point that, compared to PC, FC "has various practical and interpretational disadvantages including an inability to differentiate between directly connected nodes and nodes that are only connected via an intermediate node."[32] Many recent studies also implicitly assume that connectivity in the context of fMRI implies physical connections.[33]

In reality, there is no signal traveling from node to node: fMRI essentially measures blood flow,[1] and any correlation-based metric is only looking at how much the bandpass-filtered BOLD signals between two regions are in sync.[10] Thus, at first order, fMRI is not measuring the electrical activity of neurons or the release of neurotransmitters, although reframing the problem as connectivity or graph edges may be a useful construct.[34] On the other hand, fMRI has identified several robust characteristics of BOLD signal. In particular, it has been shown that, on average, FC intensity decreases progressing from children to young adults,[35] and that females have greater relative intra-default mode network (DMN) connectivity compared to males.[36,37] FC has also shown some ability to predict race, even between different datasets,[38] and it has been used in mechanistic studies of aggression related to olfactory stimulus.[39]

One thing that fMRI-based FC is very good at is identifying the same subject from different scans, referred to as fingerprinting.[40] FC can easily achieve 60% fingerprinting accuracy,[40] and with post-processing, fingerprinting accuracy can become greater than 95%.[41,42] Some studies have explicitly aimed to improve prediction of Alzheimer's disease by maximizing identifiability after processing with PCA.[43] At the same time, recent work shows that confounder elimination[44] or intentional but undetectable data manipulation[45] can greatly improve prediction performance.

In this work, we present a procedure by which FC-based prediction results may be unintentionally inflated. By including different scans of the same subject in both train and test sets, the machine learning algorithm learns to memorize subjects from different scans rather than to select task-specific features. Connectivities or regions that are reported in such studies may confuse other researchers when surveying the literature. Abu-Mostafa et al. (2012) gives the example of a machine learning algorithm that was able to predict exchange rate direction 52.1% of the time, resulting in a theoretical profit of 22% over 2 years.[46] In live trading, the program actually lost money. The reported problem was that the training set was normalized using the statistics of the entire cohort, and this was enough to poison the results.[46] If these problems show up in finance, where money is on the line, then we posit it may be prudent to look for them in scientific procedures as well.

2. METHODS

We first present the procedure for exploiting identifiability by treating independent scans as independent subjects. Second, we describe what we mean by identifiability or fingerprinting. Third, we give a brief review of how we derive FC or dFC from 4D fMRI volumes. Finally, we list revelant characteristics of the four datasets used in this study.

2.1 Procedure for Exploiting Identifiability

The procedure for exploiting identifiability is simple, and example code demonstrating exploitation of identifiability is provided in the link in the footnote.* When multiple longitudinal or contemporaneous scans of the same subject are available, treat these scans as independent subjects when creating training and test sets. If only one subject scan is available, use dynamic functional connectivity to create FC from multiple non-overlapping windows, and treat these FC matrices as independent subjects. In our experiments, to highlight the maximum

*https://github.com/aorliche/fc-identifiability-exploit

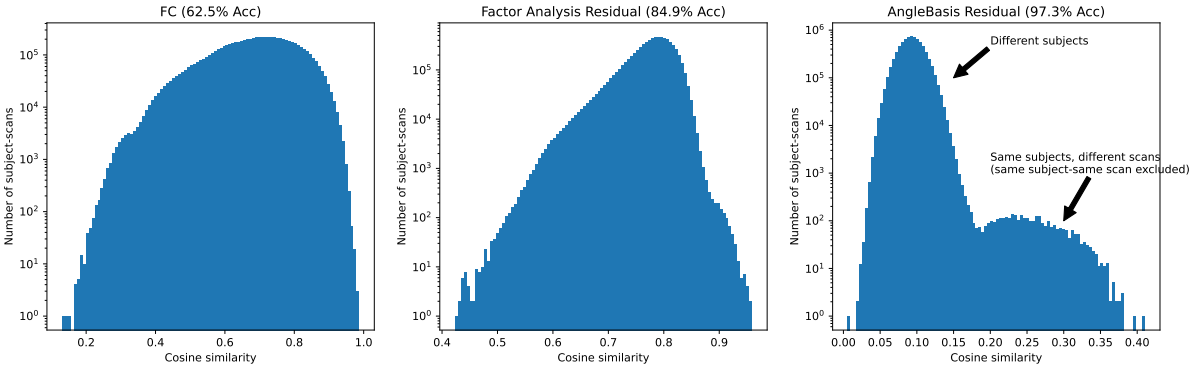

Figure 1. Demonstration of identifiability/fingerprinting with plain FC (62.5% left) vs with FC factor analysis residual (84.9% middle) vs with FC angle basis residual (97.3% right). Among 1,529 subjects having 3,843 scans, same-subject, different-scan FC has the highest cosine similarity among all scan pairs 62.5% of the time. Scans from the PNC dataset. Reproduced from Orlichenko et al. (2023).[42]

possible gap in prediction, each subject has one scan in the training set and one scan in the test set. A random distribution of scans will achieve an accuracy somewhere between the double-dipping and legitimate results.

2.2 Identifiability

We define successful identification (identifiability) of a subject as a same-subject, different-scan FC pair having a higher cosine similarity (Equation 1) compared to all other scan pairs in the cohort, where $\mathbf{a}$ and $\mathbf{b}$ are vectorized subject FCs. An alternative is to use Euclidean distance as the similarity metric; the numbers we present, however, are based on cosine similarity.

$$\text{sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^{\mathrm{T}}\mathbf{b}}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2} \tag{1}$$

We see in Figure 1 that plain FC has 62.5% identiability among 3,843 subjects in the PNC dataset. With preprocessing, this number can be increased to 97.3%.[40,41,42]

2.3 Functional and Dynamic Functional Connectivity

First, we register 4D fMRI volumes into MNI space using SPM12.[†] Second, we identify regions of interest and extract the BOLD signal from those regions. These regions may be defined either using ICA[47] or a template. We use the Power264 template in this work.[48] Third, we bandpass filter these timeseries within a 0.01 to 0.15 Hz envelope. This removes both low-frequency scanner drift and high-frequency noise as well as heartbeat and breathing signal. Finally, we calculate the Pearson correlation between the timeseries of each region to find the region-to-region FC. This symmetric matrix is reduced to the upper right triangle and vectorized. The entire procedure is illustrated in Figure 2.

When no longitudinal or contemporaneous scans are available for a subject, we create multiple FC matrices from the same scan using windowing in time. Non-overlapping windows of the bandpass-filtered timeseries are used to create multiple FC matrices. In this study we use a window size of $N = 50$ repetition times (TRs).

2.4 Predictive Models

We use simple logistic and ridge regression models for all predictive tasks. The scikit-learn implementation[49] is used in all cases.[‡] All prediction tasks are performed with an 80/20 training/test split, over 20 bootstrap iterations. The optimal hyperparameter (there is only one for both logistic and ridge regression) is chosen via grid search with grid locations at powers-of-ten intervals, i.e., on a logarithmic grid.

[†] http://www.fil.ion.ucl.ac.uk/spm/software/spm12/

[‡] https://scikit-learn.org/stable/

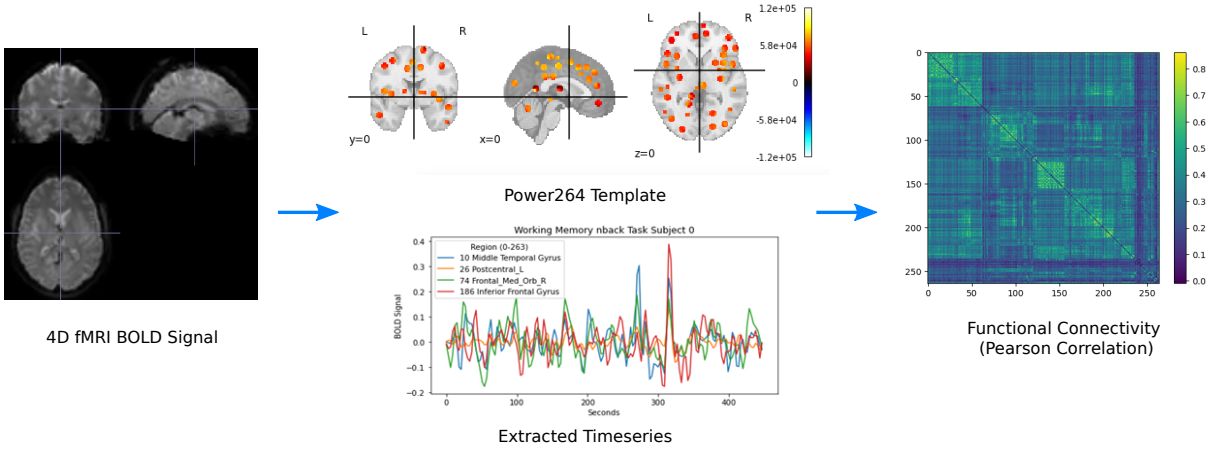

Figure 2. Illustration of the pipeline for creating FC matrices from fMRI data.

2.5 Datasets

We verify the potential of identiability to skew results in a favorable manner on four different datasets: the UK Biobank, the Philadelphia Neurodevelopmental Cohort, the Bipolar and Schizophrenia Network for Intermediate Phenotypes, and an OpenNeuro Fibromyalgia dataset.

2.5.1 UK Biobank (UKB)

We have processed the scans of more than 40,000 UK Biobank[50] subjects using SPM12. Of these, 2,722 subjects have two longitudinal scans, taken approximately two years apart. An additional 154 subjects have the second scan but not the first, resulting from quality control or a failure in our pipeline during pre-processing. We use the longitudinal subjects to predict age and genetic sex. We also predict age and sex on the non-longitudinal cohort in order to provide a baseline for model performance without double-dipping.

2.5.2 Philadelphia Neurodevelopmental Cohort (PNC)

The Philadelphia Neurodevelopmental Cohort is a dataset of 9,267 children and young adults aged 8-23 years old containing demographics, cognitive battery, questionaire responses, and SNP data.[51] Among the cohort, 1,529 subjects have fMRI scans with up to 3 scanner tasks: resting state, working memory (nback), and emotion identification (emoid).[52] The data includes Wide Range Achievement Test (WRAT) scores[53] that have had the effects of age regressed out. It has previously been shown that the ability to predict WRAT score from FC was mostly due to the different distribution of WRAT scores among races and the ability to predict race from FC.[38]

2.5.3 OpenNeuro Fibromyalgia Dataset (Fibro)

We include a 66-subject dataset of 33 female fibromyalgia patients and 33 female healthy controls from the OpenNeuro repository,[54] study identifier ds004144.[55] Out of the entire cohort, 65 subjects have two different scans: resting state and epr. A variety of medication, demographic, and questionnaire data are available.

2.5.4 Bipolar and Schizophrenia Network for Intermediate Phenotypes (BSNIP)

The Bipolar and Schizophrenia Network for Intermediate Phenotypes is a large study of schizophrenia, bipolar, and schizoaffective disorder patients; relatives of patients; and healthy controls from several sites.[56] Our data contains 199 schizophrenia patients and 243 healthy controls. Patient sex is slightly skewed toward males in the schizophrenia group. Only one scan is available per subject, necessitating use of dFC in order to exploit identifiability.

3. RESULTS

We present a summary of our results in Table 3, before highlighting results in each individual dataset.

Dataset	Task	Null Model	Best Prediction	Double-Dipping Prediction	Unintentional Improvement
UKB	Sex (Accuracy)	0.528	0.82 ± 0.02	0.89 ± 0.006	**7**%
UKB	Age (RMSE)	7.68	5.92 ± 0.03	4.97 ± 0.50	**12.4**%
PNC	WRAT (RMSE)	14.6	14.45 ± 0.92	11.0 ± 0.28	**23.6**%
Fibro	Diagnosis (Accuracy)	0.51	0.61 ± 0.17	0.86 ± 0.038	**25**%
BSNIP	Diagnosis (Accuracy)	0.5	0.78 ± 0.05	0.89 ± 0.04	**11**%

Table 1. Prediction results with and without erroneous exploitation of identifiability. Best Prediction for UKB is reported for 1350 subjects in the training set, whereas for all other datasets it is reported with the maximum number of training subjects available.

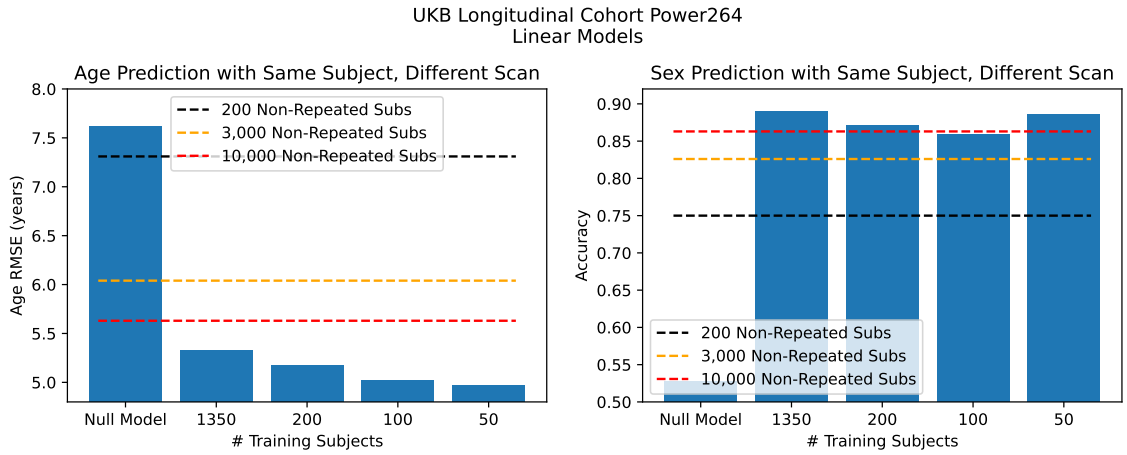

Figure 3. Prediction performance in the UKB with and without misuse of identifiability. Age prediction (left) and sex prediction (right). We find misused identifiability can lead to superb results with very small number of training subjects.

3.1 UKB

The accuracy of UKB predictions with and without unintentional identifiability enhancement (double-dipping) is shown in Figure 3. We find that misusing identifiability in the longitudinal cohort leads to prediction performance with 50 subjects not matched by 10,000 training subjects in the full cohort.

3.2 PNC

Figure 4 (top) shows the possibility of incorrectly attributing achievement score prediction to fMRI because of a race confound.[38] In fact, an even greater prediction accuracy can be achieved by treating independent scans as different subjects, as seen in Figure 4 (bottom).

3.3 Fibromyalgia

We see in Figure 5 that in cohorts with a limited number of subjects, such as the Fibromyalgia dataset, the difference between proper and improper placement of scans in training and test sets in term of prediction accuracy is maximized.

3.4 BSNIP

As the BSNIP dataset only provides one scan per subject, identifiability enhancement must rely on the use of dynamic functional connectivity, with different windows treated as independent subjects. Identiability enhancement (Figure 6) leads to predictive accuracy results in keeping with some of the larger values found in the literature.

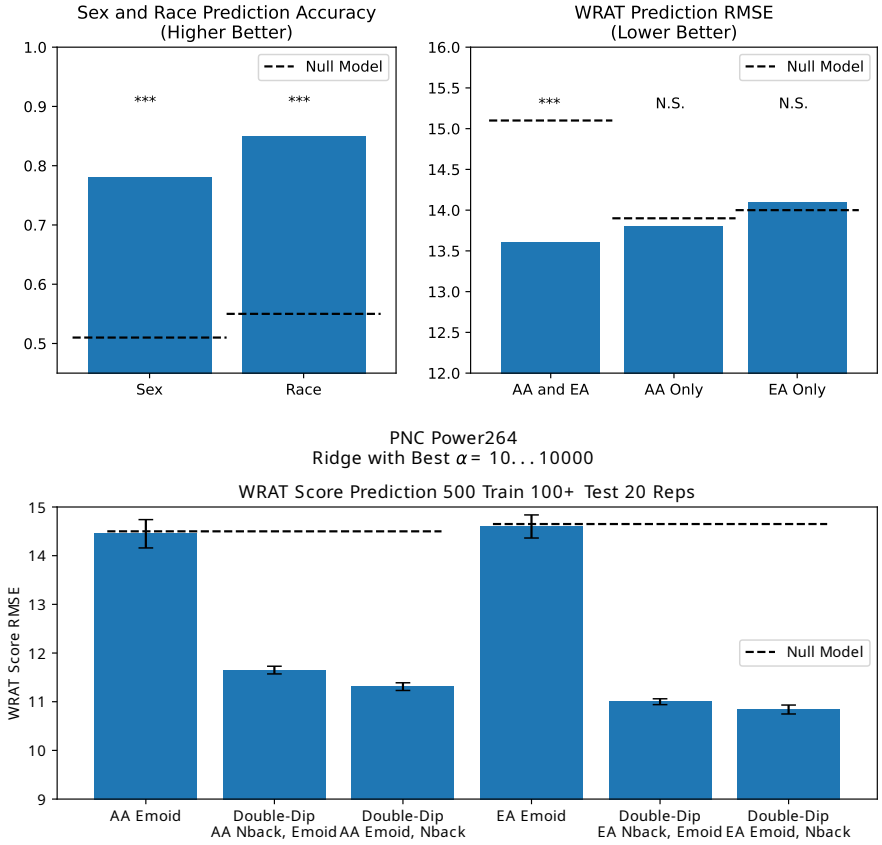

Figure 4. Incorrect attribution of FC ability to predict race as FC ability to predict achievement score (top), and the greater predictive accuracy enhancement possible by treating independent scans as independent subjects (bottom). EA refers to European Ancestry and AA refers to African Ancestry. Top graph from Orlichenko et al. (2023).[38]

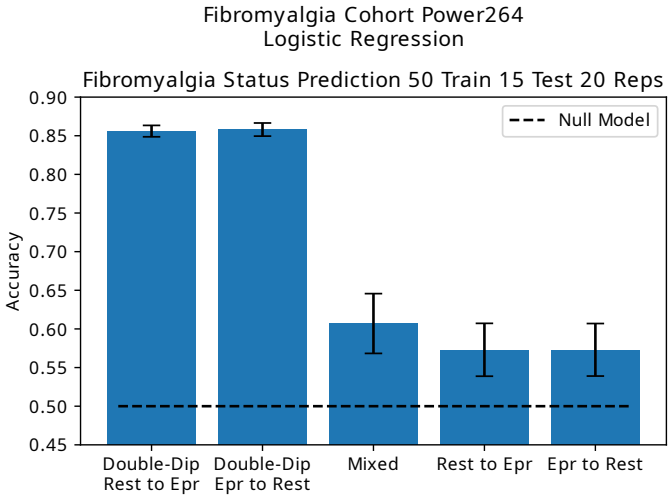

Figure 5. Prediction accuracy in the Fibromyalgia dataset using resting state scans as the training set and epr scans as the test set (and vice versa) compared to performing prediction on only one set of scans.

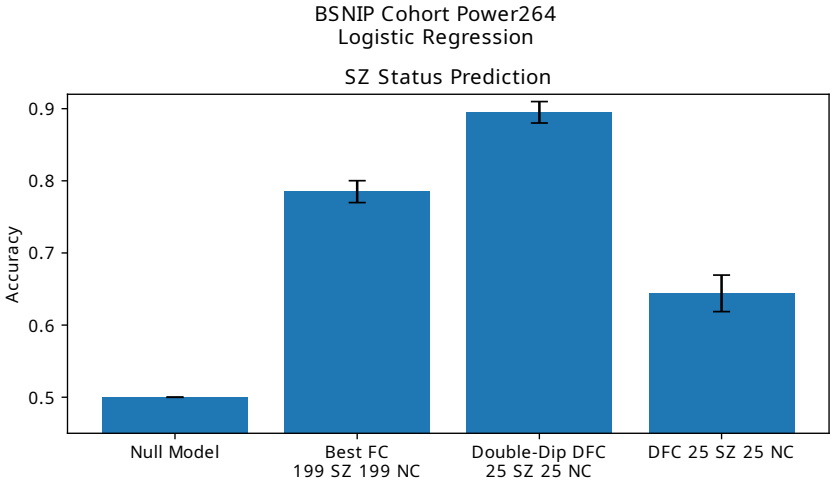

Figure 6. Schizophrenia diagnosis prediction using FC, correctly applied dFC, and dFC with using different connectivity matrices as independent subjects. We find that treating different FC matrices of the same subject as independent subjects leads to skewing of prediction results.

4. DISCUSSION

Various studies have examined the reproducibility of regions identified through fMRI studies.[2357] To our knowledge, few studies have examined the reproducibility of, e.g., fMRI schizophrenia classification results. One multi-site study did report classification accuracies of 79.8 to 97.1%,[58] the lower end of which is consistent with our own findings. Another methodology-based study found various methods to provide between 56.7 and 92.5% classification accuracy.[59] We use schizophrenia as an example, but any type of phenotype prediction may be used instead. While not a predictive study, van den Heuvel et al. (2017) showed that with a small number of subjects, minor manipulation of proportional thresholding can cause group differences to appear or disappear in an fMRI schizophrenia dataset.[60] There is another effect where an artificially inflated 95% classification accuracy existing in the literature may inhibit other researchers from publishing results that only achieve a 70-80% accuracy in their own data.

There have been at least three recent high-profile cases of data manipulation in academia: a Harvard scientist faking data,[61] a Stanford scientist implicitly allowing graduate students to fake data,[62] and evidence of fake graphs in a room temperature semiconductor publication.[63] We are not sure how prevalent this practice is in the fMRI literature, but we think based on evidence presented earlier that some misuse may occur, especially since we provide a procedure by which one may misuse longitudinal or contemporaneous scans unintentionally.

In regards to the UK Biobank, as more data is released, people may have a greater opportunity to reuse longitudinal data from the same individuals for prediction. Another potential unexplored effect is whether identifiability extends to family members. In this case, FC similarity due to relatedness may be mistaken for true FC-phenotype correlations.

4.1 Potential Solutions

The most obvious solution is not treating different scans as independent subjects and not using different dynamic FC windows as independent subjects. Otherwise, we recommend training a model on one dataset and testing on another, thus eliminating the possibility of subject memorization. Additionally, prediction results should be corroborated through different machine learning models and methods. This means that a result should only be considered valid when it is identified by several different models, not just a single newly proposed model, except with good justification. The use of very reduced feature sets (only up to 10 features per subject) may also hinder the ability of complex models to memorize identifiable subject features, even though predictive performance will almost certainly decrease. Finally, the use of a mixup model, as found in the computer science literature, may be explored as a mitigating strategy.[64]

5. CONCLUSION

We find that unintentional treatment of independent scans as independent subjects can greatly increase predictive accuracy. Prediction accuracy is increased by 7 to 25% compared to the best legitimate training procedure, using a small fraction of training subjects. This highlights the importance of reproducibility studies, as well as meaningful physiological interpretations of prediction results in contrast to optimization of prediction accuracy. It would be especially helpful if machine learning studies using neuroimaging data made proposals that could be tested in an independent manner.

6. ACKNOWLEDGEMENTS

The authors would like acknowledge the NIH (grants R01 GM109068, R01 MH104680, R01 MH107354, P20 GM103472, R01 EB020407, R01 EB006841, R56 MH124925) and NSF (grant #1539067) for partial funding support.

fMRI and phenotype data for the PNC dataset came from the Neurodevelopmental Genomics: Trajectories of Complex Phenotypes database of genotypes and phenotypes repository, dbGaP Study Accession ID phs000607.v3.p2. The authors would also like to thank the UK Biobank (UKB application ID 61915), the BSNIP study organizers, and OpenNeuro as well as the Fibromyalgia dataset curators for making data publicly available or available to authorized researchers.

REFERENCES

[1] Belliveau, J. W., Kennedy, D. N., McKinstry, R. C., Buchbinder, B. R., Weisskoff, R. M., Cohen, M. S., Vevea, J. M., Brady, T. J., and Rosen, B. R., "Functional mapping of the human visual cortex by magnetic resonance imaging.," *Science* **254 5032**, 716–9 (1991).

[2] Cox, D. D. and Savoy, R. L., "Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex," *Neuroimage* **19**, 261–270 (June 2003).

[3] Coull, J. T. and Nobre, A. C., "Where and when to pay attention: the neural systems for directing attention to spatial locations and to time intervals as revealed by both PET and fMRI," *J. Neurosci.* **18**, 7426–7435 (Sept. 1998).

[4] Pugh, K. R., Shaywitz, B. A., Shaywitz, S. E., Fulbright, R. K., Byrd, D., Skudlarski, P., Shankweiler, D. P., Katz, L., Constable, R. T., Fletcher, J., Lacadie, C., Marchione, K., and Gore, J. C., "Auditory selective attention: An fMRI investigation," *Neuroimage* **4**, 159–173 (Dec. 1996).

[5] Phan, K. L., Wager, T., Taylor, S. F., and Liberzon, I., "Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI," *Neuroimage* **16**, 331–348 (June 2002).

[6] Ochsner, K. N., Bunge, S. A., Gross, J. J., and Gabrieli, J. D. E., "Rethinking feelings: An fmri study of the cognitive regulation of emotion," *Journal of Cognitive Neuroscience* **14**(8), 1215–1229 (2002).

[7] Koelsch, S., Fritz, T., Cramon, V, D. Y., Müller, K., and Friederici, A. D., "Investigating emotion with music: an fMRI study," *Hum. Brain Mapp.* **27**, 239–250 (Mar. 2006).

[8] Hernandez, A. E., Dapretto, M., Mazziotta, J., and Bookheimer, S., "Language switching and language representation in Spanish-English bilinguals: an fMRI study," *Neuroimage* **14**, 510–520 (Aug. 2001).

[9] Szaflarski, J. P., Holland, S. K., Schmithorst, V. J., and Byars, A. W., "fMRI study of language lateralization in children and adults," *Hum. Brain Mapp.* **27**, 202–212 (Mar. 2006).

[10] van den Heuvel, M. P. and Hulshoff Pol, H. E., "Exploring the brain network: a review on resting-state fMRI functional connectivity," *Eur. Neuropsychopharmacol.* **20**, 519–534 (Aug. 2010).

[11] Orlichenko, A., Qu, G., Zhang, G., Patel, B., Wilson, T. W., Stephen, J. M., Calhoun, V. D., and Wang, Y.-P., "Latent similarity identifies important functional connections for phenotype prediction," *IEEE Transactions on Biomedical Engineering* **70**(6), 1979–1989 (2023).

[12] Hu, W., Cai, B., Zhang, A., Calhoun, V. D., and Wang, Y.-P., "Deep collaborative learning with application to the study of multimodal brain development," *IEEE Transactions on Biomedical Engineering* **66**(12), 3346–3359 (2019).

[13] İçer, S., İrem Acer, and Baş, A., "Gender-based functional connectivity differences in brain networks in childhood," *Computer Methods and Programs in Biomedicine* **192**, 105444 (2020).

[14] Sen, B. and Parhi, K. K., "Predicting biological gender and intelligence from fmri via dynamic functional connectivity," *IEEE Transactions on Biomedical Engineering* **68**(3), 815–825 (2021).

[15] Qu, G., Xiao, L., Hu, W., Wang, J., Zhang, K., Calhoun, V. D., and ping Wang, Y., "Ensemble manifold regularized multi-modal graph convolutional network for cognitive ability prediction," *IEEE Transactions on Biomedical Engineering* **68**, 3564–3573 (2021).

[16] Millar, P. R., Luckett, P. H., Gordon, B. A., and Benzinger, T. L., "Predicting brain age from functional connectivity in symptomatic and preclinical alzheimer disease," *NeuroImage* **256**, 119228 (2022).

[17] Wang, S., Zhan, Y., Zhang, Y., Lyu, L., Lyu, H., Wang, G., Wu, R., Zhao, J., and Guo, W., "Abnormal long- and short-range functional connectivity in adolescent-onset schizophrenia patients: A resting-state fMRI study," *Progress in Neuro-Psychopharmacology and Biological Psychiatry* **81**, 445–451 (Feb. 2018).

[18] Rashid, B., Arbabshirani, M. R., Damaraju, E., Cetin, M. S., Miller, R., Pearlson, G. D., and Calhoun, V. D., "Classification of schizophrenia and bipolar patients using static and dynamic resting-state fMRI brain connectivity," *Neuroimage* **134**, 645–657 (July 2016).

[19] Shamrat, F. M. J. M., Akter, S., Azam, S., Karim, A., Ghosh, P., Tasnim, Z., Hasib, K. M., De Boer, F., and Ahmed, K., "Alzheimernet: An effective deep learning based proposition for alzheimer's disease stages classification from functional brain changes in magnetic resonance images," *IEEE Access* **11**, 16376–16395 (2023).

[20] McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., Ledsam, J. R., Melnick, D., Mostofi, H., Peng, L., Reicher, J. J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K. C., De Fauw, J., and Shetty, S., "International evaluation of an AI system for breast cancer screening," *Nature* **577**, 89–94 (Jan. 2020).

[21] Salehinejad, H., Kitamura, J., Ditkofsky, N. G., Lin, A. W., Bharatha, A., Suthiphosuwan, S., Lin, H.-M., Wilson, J. R., Mamdani, M., and Colak, E., "A real-world demonstration of machine learning generalizability in the detection of intracranial hemorrhage on head computerized tomography," *Scientific Reports* **11** (2021).

[22] Szucs, D. and Ioannidis, J. P., "Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals," *NeuroImage* **221**, 117164 (2020).

[23] Chen, X., Lu, B., and Yan, C.-G., "Reproducibility of R-fMRI metrics on the impact of different strategies for multiple comparison correction and sample sizes," *Hum. Brain Mapp.* **39**, 300–318 (Jan. 2018).

[24] Bennett, C. M. and Miller, M. B., "How reliable are the results from functional magnetic resonance imaging?," *Ann. N. Y. Acad. Sci.* **1191**, 133–155 (Mar. 2010).

[25] Lee, S. H., The Schizophrenia Psychiatric Genome-Wide Association Study Consortium (PGC-SCZ), DeCandia, T. R., Ripke, S., Yang, J., Sullivan, P. F., Goddard, M. E., Keller, M. C., Visscher, P. M., Wray, N. R., The International Schizophrenia Consortium (ISC), and The Molecular Genetics of Schizophrenia Collaboration (MGS), "Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs," *Nat. Genet.* **44**, 247–250 (Mar. 2012).

[26] Bracher-Smith, M., Rees, E., Menzies, G., Walters, J. T. R., O'Donovan, M. C., Owen, M. J., Kirov, G., and Escott-Price, V., "Machine learning for prediction of schizophrenia using genetic and demographic factors in the UK biobank," *Schizophr. Res.* **246**, 156–164 (Aug. 2022).

[27] Caputi, L., Pidnebesna, A., and Hlinka, J., "Promises and pitfalls of topological data analysis for brain connectivity analysis," *Neuroimage* **238**, 118245 (Sept. 2021).

[28] Buckova, B. R., Erus, G., Spaniel, F., Davatzikos, C., and Hlinka, J., "Multimodal analysis of second-level neuroimaging features to identify first-episode schizophrenia," Poster presented at OHBM 2023 (2023).

[29] Popov, P., Mahmood, U., Kolesnikov, S., and Plis, S., "An mlp that could: A simple model with remarkable accuracy on fmri prediction tasks," Poster presented at OHBM 2023 (2023).

[30] Kanyal, A., Kandula, S., Calhoun, V., and Ye, D. H., "Deep learning on multimodal neuroimaging data for schizophrenia classification," Poster presented at OHBM 2023 (2023).

[31] Smith, S. M., Alfaro-Almagro, F., and Miller, K. L., "Uk biobank brain imaging documentation," tech. rep., UK Biobank (September 2022).

[32] Smith, S. M., "The future of FMRI connectivity," *Neuroimage* **62**, 1257–1266 (Aug. 2012).

[33] Bastos, A. M. and Schoffelen, J.-M., "A tutorial review of functional connectivity analysis methods and their interpretational pitfalls," *Front. Syst. Neurosci.* **9**, 175 (2015).

[34] Cui, H., Dai, W., Zhu, Y., Kan, X., Gu, A. A. C., Lukemire, J., Zhan, L., He, L., Guo, Y., and Yang, C., "Braingb: A benchmark for brain network analysis with graph neural networks," *IEEE Transactions on Medical Imaging* **42**(2), 493–506 (2023).

[35] Dosenbach, N. U. F., Nardos, B., Cohen, A. L., Fair, D. A., Power, J. D., Church, J. A., Nelson, S. M., Wig, G. S., Vogel, A. C., Lessov-Schlaggar, C. N., Barnes, K. A., Dubis, J. W., Feczko, E., Coalson, R. S., Pruett, Jr, J. R., Barch, D. M., Petersen, S. E., and Schlaggar, B. L., "Prediction of individual brain maturity using fMRI," *Science* **329**, 1358–1361 (Sept. 2010).

[36] Mak, L. E., Minuzzi, L., MacQueen, G., Hall, G., Kennedy, S. H., and Milev, R., "The default mode network in healthy individuals: A systematic review and meta-analysis," *Brain Connect.* **7**, 25–33 (Feb. 2017).

[37] Ficek-Tani, B., Horien, C., Ju, S., Xu, W., Li, N., Lacadie, C., Shen, X., Scheinos, D., Constable, T., and Fredericks, C., "Sex differences in default mode network connectivity in healthy aging adults," *Cereb. Cortex* (Dec. 2022).

[38] Orlichenko, A., Daly, G., Liu, A., Shen, H., Deng, H.-W., and Wang, Y.-P., "ImageNomer: developing an fMRI and omics visualization tool to detect racial bias in functional connectivity," (2023).

[39] Mishor, E., Amir, D., Weiss, T., Honigstein, D., Weissbrod, A., Livne, E., Gorodisky, L., Karagach, S., Ravia, A., Snitz, K., Karawani, D., Zirler, R., Weissgross, R., Soroka, T., Endevelt-Shapira, Y., Agron, S., Rozenkrantz, L., Reshef, N., Furman-Haran, E., Breer, H., Strotmann, J., Uebi, T., Ozaki, M., and Sobel, N., "Sniffing the human body volatile hexadecanal blocks aggression in men but triggers aggression in women," *Sci. Adv.* **7**, eabg1530 (Nov. 2021).

[40] Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., Papademetris, X., and Constable, R. T., "Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity," *Nat. Neurosci.* **18**, 1664–1671 (Nov. 2015).

[41] Cai, B., Zhang, G., Hu, W., Zhang, A., Zille, P., Zhang, Y., Stephen, J. M., Wilson, T. W., Calhoun, V. D., and Wang, Y.-P., "Refined measure of functional connectomes for improved identifiability and prediction," *Hum. Brain Mapp.* **40**, 4843–4858 (Nov. 2019).

[42] Orlichenko, A., Qu, G., Zhou, Z., Ding, Z., and Wang, Y.-P., "Angle basis: A generative model and decomposition for functional connectivity," (2023).

[43] Svaldi, D. O., Goñi, J., Abbas, K., Amico, E., Clark, D. G., Muralidharan, C., Dzemidzic, M., West, J. D., Risacher, S. L., Saykin, A. J., and Apostolova, L. G., "Optimizing differential identifiability improves connectome predictive modeling of cognitive deficits from functional connectivity in alzheimer's disease," *Hum. Brain Mapp.* **42**, 3500–3516 (Aug. 2021).

[44] Hamdan, S., Love, B. C., von Polier, G. G., Weis, S., Schwender, H., Eickhoff, S. B., and Patil, K. R., "Confound-leakage: Confound removal in machine learning leads to leakage," (2022).

[45] Rosenblatt, M., Rodriguez, R. X., Westwater, M. L., Dai, W., Horien, C., Greene, A. S., Constable, R. T., Noble, S., and Scheinost, D., "Connectome-based machine learning models are vulnerable to subtle data manipulations," *Patterns (N. Y.)* , 100756 (May 2023).

[46] Abu-Mostafa, Y. S., Magdon-Ismail, M., and Lin, H.-T., [*Learning from Data*], AMLBook (2012).

[47] Calhoun, V. D., Liu, J., and Adali, T., "A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data," *Neuroimage* **45**, S163–72 (Mar. 2009).

[48] Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., and Petersen, S. E., "Functional network organization of the human brain," *Neuron* **72**, 665–678 (Nov. 2011).

[49] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[50] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., and Collins, R., "UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS Med.* **12**, e1001779 (Mar. 2015).

[51] Glessner, J. T., Reilly, M. P., Kim, C. E., Takahashi, N., Albano, A., Hou, C., Bradfield, J. P., Zhang, H., Sleiman, P. M. A., Flory, J. H., Imielinski, M., Frackelton, E. C., Chiavacci, R., Thomas, K. A., Garris, M., Otieno, F. G., Davidson, M., Weiser, M., Reichenberg, A., Davis, K. L., Friedman, J. I., Cappola, T. P., Margulies, K. B., Rader, D. J., Grant, S. F. A., Buxbaum, J. D., Gur, R. E., and Hakonarson, H., "Strong synaptic transmission impact by copy number variations in schizophrenia," *Proc. Natl. Acad. Sci. U. S. A.* **107**, 10584–10589 (June 2010).

[52] Satterthwaite, T. D., Elliott, M. A., Ruparel, K., Loughead, J., Prabhakaran, K., Calkins, M. E., Hopson, R., Jackson, C., Keefe, J., Riley, M., Mentch, F. D., Sleiman, P. M. A., Verma, R., Davatzikos, C., Hakonarson, H., Gur, R. C., and Gur, R. E., "Neuroimaging of the philadelphia neurodevelopmental cohort," *NeuroImage* **86**, 544–553 (2014).

[53] Sayegh, P., Arentoft, A., Thaler, N. S., Dean, A. C., and Thames, A. D., "Quality of education predicts performance on the wide range achievement test-4th edition word reading subtest.," *Archives of clinical neuropsychology : the official journal of the National Academy of Neuropsychologists* **29 8**, 731–6 (2014).

[54] Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., Hardcastle, N., Wexler, J., Esteban, O., Goncavles, M., Jwa, A., and Poldrack, R., "The OpenNeuro resource for sharing of neuroscience data," *Elife* **10** (Oct. 2021).

[55] Balducci, T., Rasgado-Toledo, J., Valencia, A., van Tol, M.-J., Aleman, A., and Garza-Villarreal, E. A., "A behavioral, clinical and brain imaging dataset with focus on emotion regulation of females with fibromyalgia," (2022). doi:10.18112/openneuro.ds004144.v1.0.1.

[56] Tamminga, C. A., Pearlson, G., Keshavan, M., Sweeney, J., Clementz, B., and Thaker, G., "Bipolar and schizophrenia network for intermediate phenotypes: outcomes across the psychosis continuum," *Schizophr. Bull.* **40 Suppl 2**, S131–7 (Mar. 2014).

[57] Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., and Hariri, A. R., "What is the test-retest reliability of common task-functional MRI measures? new empirical evidence and a meta-analysis," *Psychol. Sci.* **31**, 792–806 (July 2020).

[58] Salman, M. S., Verner, E., Bockholt, H. J., Fu, Z., Misiura, M., Baker, B. T., Osuch, E., Sui, J., and Calhoun, V. D., "Multi-study evaluation of neuroimaging-based prediction of medication class in mood disorders," *Psychiatry Res. Neuroimaging* **333**, 111655 (Aug. 2023).

[59] Zhang, Y., Zhang, H., Xiao, L., Bai, Y., Calhoun, V. D., and Wang, Y.-P., "Multi-modal imaging genetics data fusion via a hypergraph-based manifold regularization: Application to schizophrenia study," *IEEE Transactions on Medical Imaging* **41**(9), 2263–2272 (2022).

[60] van den Heuvel, M. P., de Lange, S. C., Zalesky, A., Seguin, C., Yeo, B. T. T., and Schmidt, R., "Proportional thresholding in resting-state fMRI functional connectivity networks and consequences for patient-control connectome studies: Issues and recommendations," *Neuroimage* **152**, 437–449 (May 2017).

[61] Simmons, J., Nelson, L., and Simonsohn, U., "[109] data falsificada (part 1): "clusterfake"." https://datacolada.org/109 (2023). Accessed: 2023-08-02.

[62] Baker, T., "Stanford president's research under investigation for scientific misconduct, university admits 'mistakes'." https://stanforddaily.com/2022/11/29/stanford-presidents-research-under-investigation-for-scientific-misconduct-university-admits-mistake (2023). Accessed: 2023-08-02.

[63] Marel, D. v. d. and Hirsch, J. E., "Room-temperature superconductivity — or not? comment on *Nature* 586, 373 (2020) by e. snider *et al.*," *Int. J. Mod. Phys. B* **37** (Feb. 2023).

[64] Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D., "mixup: Beyond empirical risk minimization," (2017).