Multimodal Imputation of Imaging-derived Phenotypes from Genomic and Blood-based Biomarkers Enhances Common Disease Discovery

Haoyang Zhang[1], Yan Li[2], Junhong Liu[1], Lizhen Lan[1], Zian Wang[3], Longyu Sun[1], Yuntong Lv[4], Shengxiao Yang[5], Qing Li[1], Mengting Sun[1], Yajing Zhang[6], Binghua Chen[7], Xionghui Zhou[8], Lianming Wu[7], and Chengyan Wang[1(✉)]

[1] Shanghai Pudong Hospital and Human Phenome Institute, Fudan University, Shanghai, China
[2] Department of Radiology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[3] School of Computer Science, Fudan University, Shanghai, China
[4] School of Clinical Medicine, Zhongshan Hospital, Shanghai Medical College, Fudan University, Shanghai, China
[5] School of Data Science, Fudan University, Shanghai, China
[6] GE Healthcare, Beijing, China
[7] Department of Radiology, Ren Ji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China
[8] Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan, China
wangcy@fudan.edu.cn

Abstract. Medical imaging provides a wealth of information about a patient's physical condition, and Imaging-derived phenotypes (IDPs) extracted from medical images have applications in various biomedical tasks such as disease prediction and phenotype association studies. For disease prediction tasks, the collection of multimodal imaging data and the conduct of long-term follow-ups are crucial; however, the low incidence rates of certain diseases make it challenging to acquire large-scale cohort data. On the other hand, cohorts that contain genomics and blood-based biomarkers are relatively extensive. Against this backdrop, large-scale cohort data from the UK Biobank (UKB) were leveraged to construct prediction models for 260 IDPs extracted from common brain MRI and cardiac MRI using machine learning methods combined with genomics and basic blood characteristics. We applied these models to impute IDPs in cohorts missing imaging data and utilized the imputed IDPs for IDP-disease association studies and disease prediction. Association study results demonstrate that the imputed IDPs can reveal numerous IDP-disease associations. Furthermore, the disease prediction models developed using imputed IDPs demonstrated significantly superior performance across 184 common diseases, as evidenced by higher overall AUC values when compared to models utilizing real IDPs (Wilcoxon signed-rank test, $p < 0.001$). These results clearly highlight the significant application value of our IDPs prediction models in the context of disease discovery.

Keywords: Imaging-derived phenotypes, Imputation, Disease discovery

1 Introduction

Medical imaging, a crucial component of modern healthcare, enables visualization of the body's internal structures and functions. Techniques like X-ray, CT, and MRI provide detailed images and are essential for disease diagnosis, treatment planning, and patient monitoring. In biomedical applications such as disease prediction, medical images cannot be directly utilized as information sources but require undergoing image processing and information extraction [1]. Imaging-derived phenotypes (IDPs) are phenotypic information extracted from medical images using specific processing pipelines [2], which possess specific significance and can participate in tasks such as biomedical modeling. Currently, IDPs have been applied to multiple biomedical fields such as phenotype prediction [3, 4] and phenotype association study [5, 6], confirming the significant importance of IDPs in biomedical research.

However, for disease prediction tasks, the collection of multimodal imaging data and long-term follow-up are essential, but are often limited by the low incidence rates of certain diseases. This makes it difficult to obtain large-scale cohort data with complete IDPs. Although many general-purpose data imputation methods, such as K-nearest neighbors (KNN) imputation and regression-model-based imputation [7], have been developed, most of these approaches rely on prior knowledge of partially observed phenotypes. In the field of omics research, imputation methods for partially or completely missing data have been developed for various omics data [8]. For example, Zhou et al. developed a framework for predicting RNA-sequencing data from DNA methylation [9], Ansari et al. developed an approach for optimizing multi-omics data imputation with NMF and GAN synergy [10]. In addition, Xu et al. utilized single nucleotide polymorphism (SNP) data to develop polygenic score maps for multiple omics data and trained polygenic score models using the Bayesian ridge regression model to predict omics features [11]. However, to our best knowledge, there is a lack of effective data prediction methods for cohorts with completely missing IDPs.

To address the challenge of missing IDPs data in large-scale cohorts, referring to the research approach of Xu et al. [11], we integrated more widely available SNP data and low-cost basic blood characteristics (including blood counts and blood biochemistry) and adopted an ensemble learning approach [12] to build multiple two-layer stacked models to impute IDPs. Predictive models for 260 brain T1, brain SWI, and cardiac MRI-derived phenotypes were developed and achieved good IDPs prediction performance. Subsequently, we applied our models to the UKB European population cohort without IDPs, predicted their brain and cardiac IDPs. Utilizing the real IDPs cohort and the imputed IDPs cohort, an IDP-disease association study was first conducted. It was found that the imputed IDPs were capable of identifying numerous IDPs-disease associations. This reveals that the imputed IDPs can offer insights into the associations between IDPs and diseases within cohorts lacking IDPs. These insights are advantageous for the implementation of explanatory research. Finally, we utilized the imputed IDPs for building disease prediction models, and compared the performance of disease prediction models based on imputed IDPs with those based on real IDPs. The evaluation results indicated that models based on imputed IDPs demonstrated better disease prediction performance, highlighting the value of imputed IDPs in application.

2 Materials and Methods

2.1 Data Collection

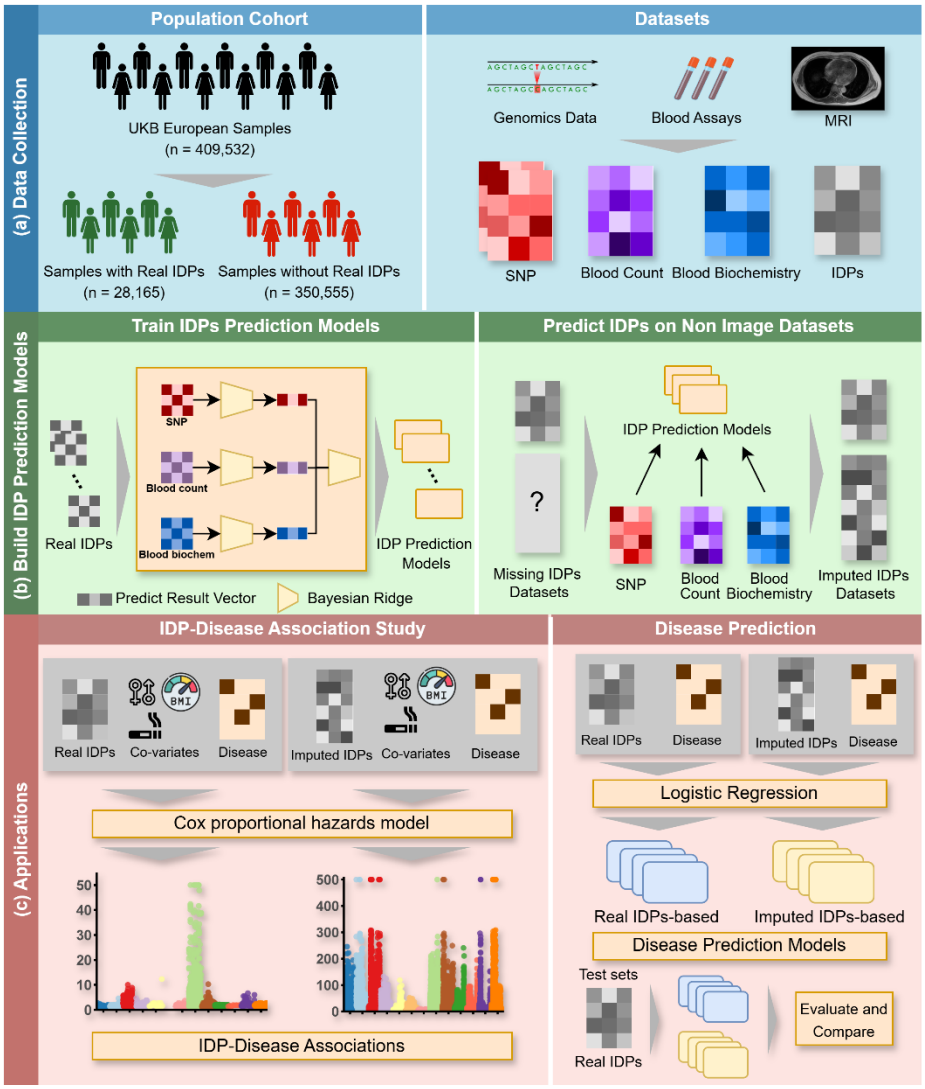

Fig. 1. Overall workflow of the study. (a) Collection of multimodal datasets. (b) Build and apply IDPs prediction models. (c) Downstream applications of imputed IDPs cohorts, including IDP-Disease association study and disease prediction.

Datasets. A total of 260 IDPs were selected, including 164 T1-weighted structural brain MRI-derived IDPs, 14 susceptibility-weighted brain MRI-derived IDPs, and 82 cardiac

and aortic structure/function IDPs from the UKB imaging pipeline. Sixty-one blood-based traits from the UKB blood assay were also included, comprising 31 blood count features and 30 blood biochemical features. We performed z-score standardization on the IDPs data and blood-based traits data and applied KNN imputation to impute the missing values.

For disease association and prediction analysis, we identified 184 diseases within the cohort possessing confirmed IDPs, each with at least 50 incident cases. These diseases spanned diverse categories to evaluate the utility of imputed IDPs in disease-related research comprehensively.

Publicly available Genome-Wide Association Study (GWAS) summary statistics for 82 cardiac IDPs [13] and 178 brain IDPs [14] were obtained to support IDPs prediction analyses. We selected SNPs from the GWAS summary statistics for each IDP with p-values less than 0.001 and used the reference genome SNP file from Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA GWAS) [15] to process linkage disequilibrium with PLINK2 [16], filter SNPs, and extract the SNP dosage matrices.

Quality Control. We utilized population cohort data from the UK Biobank. Initially, we selected samples from individuals of European ancestry. Individuals with missing values exceeding 20% in either blood-based or imaging-derived phenotypic data were excluded. Only samples with both blood and genomic data were included. Based on the availability of IDPs, the cohort was divided into two groups: 28,615 individuals with confirmed IDPs and 350,555 individuals lacking IDPs for downstream analyses.

2.2 IDPs Prediction Models

In constructing the IDPs prediction models, we adopted an ensemble learning approach to build a two-layer stacked model. In the first layer, we trained a Bayesian ridge regression (BR) model for SNP data, blood count data, and biochemical data separately. In the second layer, we used the inputs from the three models in the training set as features to train a BR model. A separate stacked model was trained for each IDP. We employed five-fold cross-validation and calculated the R2 and Spearman correlation for each IDP prediction model as performance metrics.

2.3 Disease Discovery Analysis

Phenotype-Disease Association Study. We employed Cox proportional hazards regression models, adjusting for age, sex, body mass index (BMI), and smoking status as covariates, to examine the associations between both real and imputed IDPs and the 184 previously identified diseases. The statistical significance of these associations was determined using a false discovery rate (FDR) threshold of 5% for multiple testing correction.

Disease Prediction Models. We employed L1-regularized logistic regression models to construct disease prediction models for the 184 diseases in both the imputed IDPs cohort and the real IDPs cohort. In the real IDPs cohort, the training set and test set were divided in a 7:3 ratio. In the imputed IDPs cohort, all samples were used as the training set. The test sets divided from the real IDPs cohort are utilized for model testing in both models. To address sample imbalance in the training set, random down-sampling was used to equalize the number of minority and majority class samples. Due to the randomness of down-sampling and dataset splitting, we ran the process ten times with different random seeds to ensure the reliability of the results. We used AUC, accuracy, specificity, and recall as evaluation metrics.

3 Results

3.1 IDPs Prediction Model

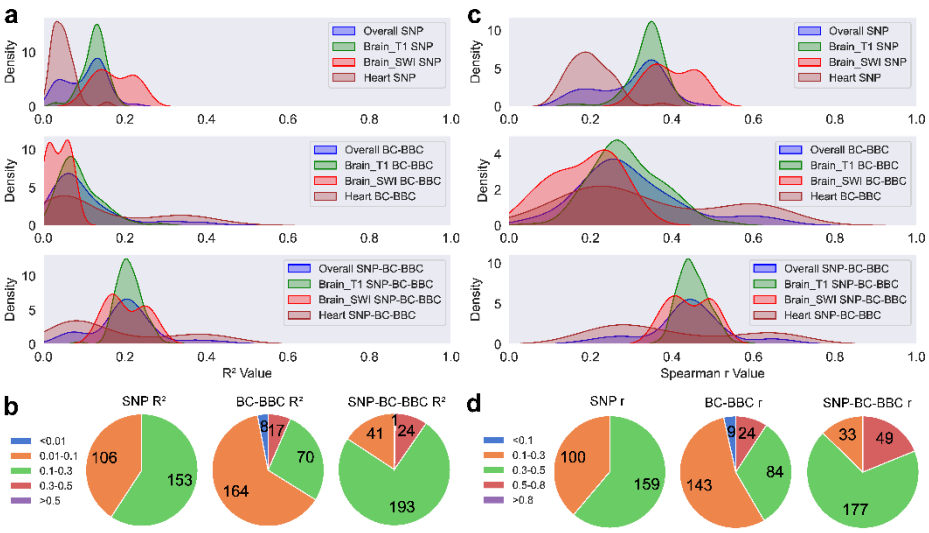

Fig. 2. Evaluation results of the IDP prediction model. SNP, BC, and BBC denote single nucleotide polymorphism data, blood count data, and blood biochemistry data, respectively. Hyphenated terms indicate data integration. Panels (a) and (b) show the distribution of R^2 values and Spearman correlation coefficients for brain T1, brain SWI, cardiac, and overall IDPs predictions using SNP data alone, blood data alone, and the integration of SNP and blood data. Panels (c) and (d) show the interval distribution of R^2 values and Spearman correlation coefficients for overall IDP predictions and imputation, respectively.

We developed a two-layer stacked model incorporating Bayesian ridge regression (BR) to predict 260 MRI-derived brain and cardiac IDPs using SNPs and blood-based biomarkers. Model performance was rigorously evaluated through five-fold cross-validation, with comprehensive results presented in Figure 2.

For comparative analysis, we constructed three distinct model configurations: SNP-only models, blood biomarker-only models, and integrated models combining both modalities. As shown in figure 2, the integrated models demonstrated superior predictive performance compared to single-modality approaches, with coefficients of determination (R^2) predominantly distributed between 0.1-0.5. Spearman correlation coefficients between imputed and real IDPs were predominantly in the range of 0.3 to 0.6, with select instances approaching 0.8. Notably, cardiac IDPs exhibited enhanced predictability in integrated models compared to cerebral IDPs, as evidenced in Figure 2a. These results substantiate our biological hypothesis regarding the combined genetic and blood regulation of IDPs, while demonstrating the feasibility of multimodal data prediction for imaging-lacking cohorts.

3.2 IDP-Disease Association Study

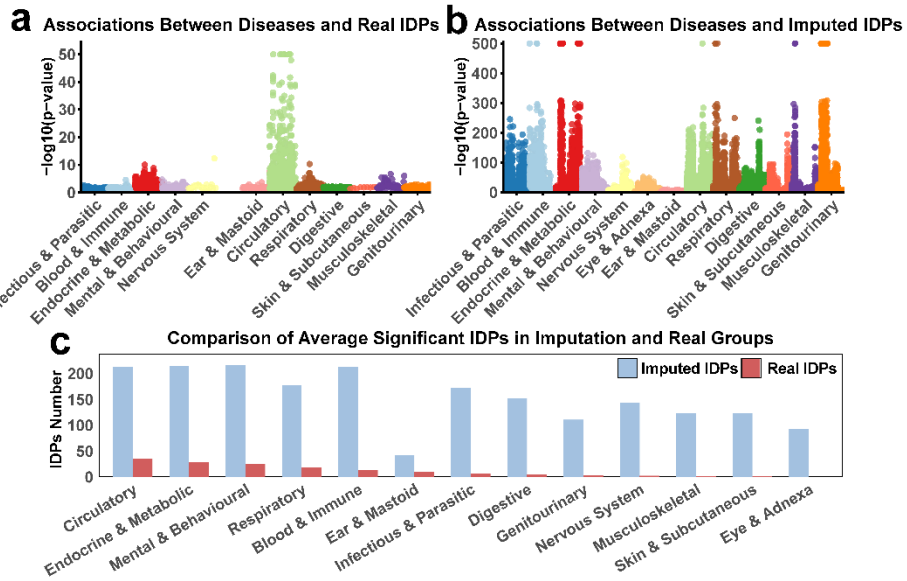

Fig. 3. Results of the IDP-disease association study. (a) Disease-IDP correlation results derived from real IDPs data. (b) Disease-IDP correlation results derived from imputed IDPs data. The -log10(p - value) of associations that exceeded 50 (for real IDPs) or 500 (for imputed IDPs) or were displayed as 0 are set to 50 (for real IDPs) or 500 (for imputed IDPs). (c) Comparison of average number of Significant IDPs between imputed IDPs and real IDPs, disease types were ranked according to number of significant real IDPs.

To investigate clinical relevance, we conducted phenome-wide association studies (PheWAS) between IDPs and 184 diseases (prevalence > 50 cases in the 28,165-subject imaging cohort). Analyses encompassed both real IDPs from 28,165 subjects with MRI data and imputed IDPs extrapolated to 350,555 subjects (Figure 3). The imputed IDPs cohort exhibited substantially more significant associations ($P < 0.05$, FDR-corrected)

than the observed cohort. While real IDPs primarily revealed associations with circulatory system diseases, imputed IDPs demonstrated polygenic associations across multiple disease categories. We also compared the average number of significantly associated IDPs per disease for real and imputed IDPs across disease types. Figure 2(c) shows a ranking correlation in the average disease-associated IDPs between real and imputed IDPs. Spearman correlation analysis revealed a significant correlation ($p = 0.004$, rho = 0.758), indicating a ranking correlation and overall trend consistency in the large-scale disease-IDPs association results. However, differences exist in the significant IDPs identified by the two types of IDPs.

3.3 Application of imputed IDPs in Disease Prediction

The principal advantage of phenomics cohort expansion lies in its capacity to substantially increase sample size, thereby enhancing disease prediction modeling. We developed disease prediction models for 184 selected clinical conditions using L1-regularized logistic regression, separately employing: 1) imputed IDPs from the expanded cohort of 350,555 subjects, and 2) real IDPs from the original 28,615-sample cohort. Crucially, both models were evaluated on an identical test set with real IDPs measurements to ensure reliable assessment of predictive enhancement through cohort expansion. To mitigate stochastic influences from data partitioning and random subsampling, we conducted ten independent model iterations with varying random seeds. The consolidated evaluation metrics across these iterations are systematically presented in Figure 4.

We integrated the evaluation results of two models across 184 diseases and performed Wilcoxon signed-rank tests to compare the performance metrics (accuracy, AUC, recall, and specificity) between models constructed using real IDPs versus imputed IDPs. The results demonstrated that disease prediction models based on imputed IDPs from large-scale cohorts exhibited significantly superior overall performance compared to those built with real IDPs data, despite our imputed IDPs models being tested against real IDPs that theoretically differ in distribution characteristics. Subsequent disease category-specific AUC comparisons revealed that models utilizing imputed IDPs from large-scale cohorts outperformed those based on small-scale real IDPs data in all disease categories (all p-values < 0.001).

Combined with the outstanding performance in Spearman correlation coefficients of the IDPs prediction models mentioned in Section 3.1, these findings suggest that although substantial improvements remain in the absolute value accuracy of imputed IDPs, their superior rank correlation might enable logistic regression-based disease prediction models to benefit from the positive weighting effects afforded by expanded cohort sizes during training. This mechanism could facilitate the development of more accurate and effective models. This discovery substantiates the value of constructing imputed IDPs-based models, representing a kind of data augmentation [17] method for expanding training samples. These insights may provide valuable guidance for future directions in enhancing the performance of prediction models.

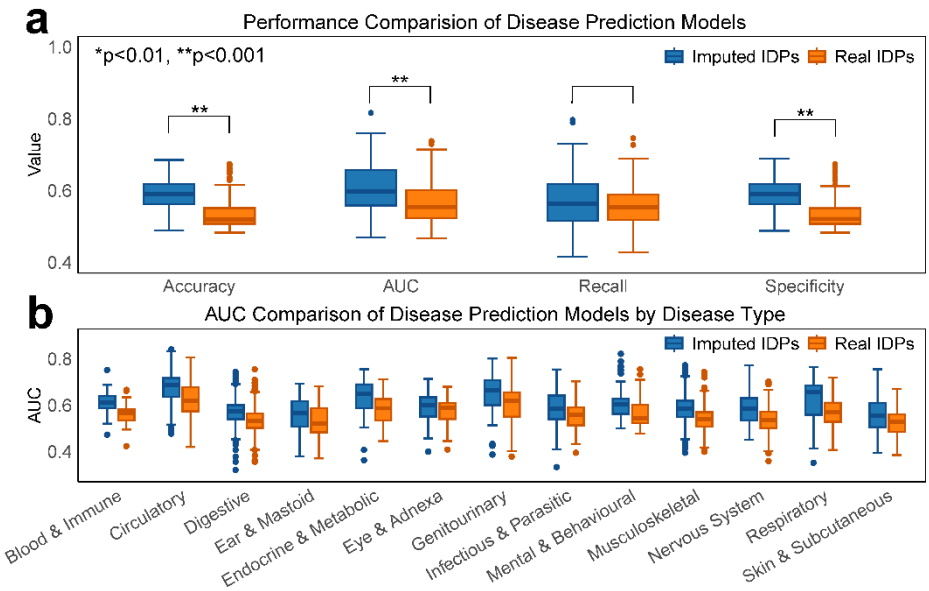

Fig. 4. Comparison of evaluation results of disease prediction models built on 184 diseases using real IDPs and imputed IDPs. (a) Overall evaluation metrics of the two models across 184 diseases, with values averaged over ten repeated random seeds. (b) Comparison of the distribution of disease prediction AUC for each disease type between the two types of models.

4 Discussion and Conclusion

Drawing upon previous research frameworks that utilized genomic SNP signatures for multi-omics prediction [11], this study established predictive models for 260 IDPs under the biological hypothesis that human IDPs are jointly influenced by genetic and environmental factors. Leveraging multimodal data from the UK Biobank (UKB), we integrated SNPs and fundamental blood-based biomarkers to develop robust predictive models with satisfactory performance. These models were subsequently applied to large-scale population cohorts lacking direct IDP measurements to generate predicted IDP values for downstream analyses.

Our results demonstrated the utility of imputed IDPs in two critical disease discovery applications: First, the imputed IDPs successfully identified substantial IDP-disease associations, confirming their capacity to expand phenotype-disease insights for cohorts without imaging data. Second, disease prediction models constructed using imputed IDPs exhibited significantly superior performance compared to those based on real IDPs. We infer this enhancement primarily stems from the expanded training sample size enabled by imputed IDPs, effectively achieving feature-based data augmentation.

Nevertheless, this study has inherent limitations. First, we did not compare the performance of various machine learning models, and the existing model architecture may not be optimal. Second, the lack of classification and in-depth discussion on IDPs has limited the research depth. Finally, while expanding the sample size likely contributes

to the improved performance of prediction-based disease models, our preliminary analysis did not fully explore other potential mechanisms underlying this phenomenon. Future studies could enhance IDP prediction efficacy by comparing the performance of multiple models. Additionally, investigating the reasons for varying prediction difficulties among different IDPs, conducting in-depth discussions on the roles of different modalities, analyzing the contribution of specific SNPs in prediction, and exploring how IDP prediction improves the performance of disease prediction models would enhance both the research depth and biological significance.

In summary, this study established predictive models for 260 IDPs by integrating genetic and blood-based biomarkers, demonstrating their utility in expanding phenotype-disease insights and enhancing disease prediction performance, thereby validating their application potential in the field of disease discovery.

Acknowledgments. This work was supported in part by the Shanghai Municipal Science and Technology Major Project (No.2023SHZDZX02A05), and the Shanghai Rising-Star Program (No.24QA2703300). The computations in this research were performed using the CFFF platform of Fudan University.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L.: Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nature neuroscience 19, 1523-1536 (2016)
2. Elliott, L.T., Sharp, K., Alfaro-Almagro, F., Shi, S., Miller, K.L., Douaud, G., Marchini, J., Smith, S.M.: Genome-wide association studies of brain imaging phenotypes in UK Biobank. Nature 562, 210-216 (2018)
3. Leiby, J.S., Lee, M.E., Shivakumar, M., Choe, E.K., Kim, D.: Deep learning imaging phenotype can classify metabolic syndrome and is predictive of cardiometabolic disorders. Journal of Translational Medicine 22, 434 (2024)
4. Gong, W., Beckmann, C.F., Smith, S.M.: Phenotype discovery from population brain imaging. Medical Image Analysis 71, 102050 (2021)
5. Ning, C., Fan, L., Jin, M., Wang, W., Hu, Z., Cai, Y., Chen, L., Lu, Z., Zhang, M., Chen, C., Li, Y., Zhang, F., Wang, W., Liu, Y., Chen, S., Jiang, Y., He, C., Wang, Z., Chen, X., Li, H., Li, G., Ma, Q., Geng, H., Tian, W., Zhang, H., Liu, B., Xia, Q., Yang, X., Liu, Z., Li, B., Zhu, Y., Li, X., Zhang, S., Tian, J., Miao, X.: Genome-wide association analysis of left ventricular imaging-derived phenotypes identifies 72 risk loci and yields genetic insights into hypertrophic cardiomyopathy. Nat Commun 14, 7900 (2023)
6. Liu, Y., Shen, O., Zhu, H., He, Y., Chang, X., Sun, L., Jia, Y., Sun, H., Wang, Y., Xu, Q., Guo, D., Shi, M., Zheng, J., Zhu, Z.: Associations between brain imaging–derived phenotypes and cognitive functions. Cerebral Cortex 34, (2024)

7. Jadhav, A., Pramod, D., Ramanathan, K.: Comparison of performance of data imputation methods for numeric dataset. Applied Artificial Intelligence 33, 913-933 (2019)
8. Song, M., Greenbaum, J., Luttrell IV, J., Zhou, W., Wu, C., Shen, H., Gong, P., Zhang, C., Deng, H.-W.: A review of integrative imputation for multi-omics datasets. Frontiers in genetics 11, 570255 (2020)
9. Zhou, X., Chai, H., Zhao, H., Luo, C.-H., Yang, Y.: Imputing missing RNA-sequencing data from DNA methylation by using a transfer learning–based neural network. GigaScience 9, (2020)
10. Ansari, M.I., Ahmed, K.T., Zhang, W.: Optimizing multi-omics data imputation with NMF and GAN synergy. Bioinformatics 40, (2024)
11. Xu, Y., Ritchie, S.C., Liang, Y., Timmers, P., Pietzner, M., Lannelongue, L., Lambert, S.A., Tahir, U.A., May-Wilson, S., Foguet, C., Johansson, Å., Surendran, P., Nath, A.P., Persyn, E., Peters, J.E., Oliver-Williams, C., Deng, S., Prins, B., Luan, J., Bomba, L., Soranzo, N., Di Angelantonio, E., Pirastu, N., Tai, E.S., van Dam, R.M., Parkinson, H., Davenport, E.E., Paul, D.S., Yau, C., Gerszten, R.E., Mälarstig, A., Danesh, J., Sim, X., Langenberg, C., Wilson, J.F., Butterworth, A.S., Inouye, M.: An atlas of genetic scores to predict multi-omic traits. Nature 616, 123-131 (2023)
12. Mienye, I.D., Sun, Y.: A survey of ensemble learning: Concepts, algorithms, applications, and prospects. IEEE Access 10, 99129-99149 (2022)
13. Zhao, B., Li, T., Fan, Z., Yang, Y., Shu, J., Yang, X., Wang, X., Luo, T., Tang, J., Xiong, D., Wu, Z., Li, B., Chen, J., Shan, Y., Tomlinson, C., Zhu, Z., Li, Y., Stein, J.L., Zhu, H.: Heart-brain connections: Phenotypic and genetic insights from magnetic resonance images. Science 380, abn6598 (2023)
14. Smith, S.M., Douaud, G., Chen, W., Hanayik, T., Alfaro-Almagro, F., Sharp, K., Elliott, L.T.: An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank. Nat Neurosci 24, 737-745 (2021)
15. Watanabe, K., Taskesen, E., van Bochoven, A., Posthuma, D.: Functional mapping and annotation of genetic associations with FUMA. Nature Communications 8, 1826 (2017)
16. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J.: Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4, (2015)
17. Maharana, K., Mondal, S., Nemade, B.: A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings 3, 91-99 (2022)