

Comparing Data-Driven Subtypes of Depression Informed by Clinical and Neuroimaging Data: A Registered Report

Kayla Hannon, Setthanan Jarukasemkit, Leda Balogh, Fyzeen Ahmad, Petra Lenzini, Aristeidis Sotiras, and Janine D. Bijsterbosch

ABSTRACT

BACKGROUND: Efforts to elucidate subtypes within depression have yet to establish a consensus. In this study, we aimed to rigorously compare different subtyping approaches in the same participant space to quantitatively test agreement across subtyping approaches and determine whether the different approaches are sensitive to different sources of heterogeneity in depression.

METHODS: We implemented 6 different data-driven subtyping methods developed in previous work using the same UK Biobank participants ($n = 2276$ participants with depression, $n = 1595$ healthy control participants). The 6 approaches include 2 symptom-based, 2 structural neuroimaging-based, and 2 functional neuroimaging-based techniques. The resulting subtypes were compared based on participant assignment, stability, and sensitivity to subtype differences in demographics, general health, clinical characteristics, neuroimaging, trauma, cognition, genetics, and inflammation markers.

RESULTS: We found almost no agreement between the resulting subtypes of the 6 approaches (mean adjusted Rand index [ARI] = 0.006), even within data domains. This finding was largely driven by differences in input feature set (mean ARI = 0.005) rather than clustering algorithm (mean ARI = 0.23). However, each approach had relatively high internal stability across bootstraps (ARI = 0.36–0.89); most approaches performed above null; and most approaches were sensitive to relevant phenotypes within their data domain.

CONCLUSIONS: Despite marginal overlap between approaches, we found the subtyping approaches to be internally consistent. These results explain why previous studies found strong evidence for subtypes within their analysis but with very little convergence between studies. We recommend that in future work, investigators incorporate systematic comparisons between their approach and alternative/previous approaches to facilitate consensus on depression subtypes.

<https://doi.org/10.1016/j.bpsgos.2025.100473>

According to the World Health Organization, depression is the leading contributor to disability globally, with an estimated 4.4% people in the world experiencing depression (1). Despite this high prevalence and negative impact, the basis for depression in the brain is still unclear. A major barrier to understanding the neural basis of depression is the heterogeneity within the depressed population (2).

Although the presence of subtypes in depression has often been investigated (3–12), no consensus has been achieved in the literature regarding depression subtypes. For example, some attempts to replicate subtyping approaches have been unsuccessful (10), and different subtyping approaches have largely failed to converge on comparable depression subtypes (3). The reasons for this lack of agreement are still unclear, partly due to a lack of rigorous comparisons across subtyping findings/approaches. The differences in depression subtypes across studies could be due to differences in inclusion/exclusion criteria across studies. Alternatively, different subtyping

approaches may capture different domains of heterogeneity. For example, studies that adopt data-driven subtyping analysis driven by symptom features may be capturing symptom heterogeneity but not neuroimaging heterogeneity, whereas studies that adopt subtypes based on neuroimaging features may not be sensitive to symptom heterogeneity (13).

In this study, we aimed to compare 6 previously published data-driven approaches for identifying depression subtypes (4–9) (based on both symptom and neuroimaging features) by applying these previously used approaches in the same large-scale cohort. After applying each of the 6 subtyping approaches in the same cohort, we compared the resulting subtypes on the commonality of their groupings (i.e., tested how consistently a participant was assigned to the same subtype) and the stability of resulting subtypes. We leveraged the UK Biobank (UKB) cohort, which has released both symptom and neuroimaging data from >40,000 participants (14,15). Furthermore, we compared the 6 sets of subtypes in

terms of group differences for depression-relevant variables that were not used to inform the subtypes (such as demographics, genetics, cognition). Therefore, the results are expected to provide insights into the relative sensitivity of different subtyping approaches to depression-related factors that are indicative of risk (e.g., trauma) and/or etiological mechanisms (e.g., inflammation).

We note that this study is not intended to be a true replication of each of the original 6 studies; instead, the primary goal is the comparison across resulting subtypes. Although we followed the feature selection and algorithmic choices of the original subtyping approaches when applicable, we diverged when we needed to maximize our ability to compare across resulting subtypes as described in the [Methods and Materials](#). Specifically, our translations of the original studies differed in terms of the demographic characteristics of the sample (because the UKB is a middle- to older-aged cohort), the inclusion/exclusion criteria of the depression sample (which was kept identical across the 6 studies to enable comparisons), the symptom measures (which were mapped onto available UKB variables as closely as possible), the neuroimaging preprocessing (the UKB-released preprocessed data were used for all studies to avoid potential preprocessing differences influencing our comparisons), and covariates included in the analyses (which were kept consistent across approaches).

This study elucidated important drivers of inconsistency between data-driven approaches used to identify depression subtypes. Although the subtypes derived from each approach had marginal overlap above chance, each approach was internally stable and sensitive to relevant phenotypes. Our findings established that the major driver of subtypes was the input feature set. Here, even relatively subtle differences between feature sets derived from the same domain (i.e., clinical, structural, functional neuroimaging) resulted in substantial differences between subtypes. Differences in the clustering approach had a more minor impact on inconsistency. Taken together, these findings suggest that future work would benefit from including comparisons between multiple feature types and/or clustering approaches to help identify consistent subtypes.

METHODS AND MATERIALS

This work was performed as a registered report. The stage 1 planned analyses and associated hypotheses are provided in [Table S1](#) and were preregistered on the Open Science Framework (<https://osf.io/w54da/>). Any deviation from the stage 1 planned analyses are explicitly stated in the descriptions below, and explanations for changes are provided in [Table S7](#).

Group Definition

The UKB is an epidemiological cohort from which imaging and clinical measures have been collected for more than 40,000 participants (15). Previous work developed criteria for a case-control definition of individuals likely experiencing depression based on measures available in the UKB (16), which we used to identify individuals with depression in this work ([Table S2](#)). We excluded individuals who had schizophrenia, bipolar disorder, psychotic symptoms, obsessive-compulsive disorder,

posttraumatic stress disorder, Huntington's disease, Alzheimer's disease, epilepsy, or stroke ([Table S3](#)) and anyone who had missing data for the clinical and imaging features used in the main analysis ([Table S4](#)). The resulting group of participants with depression had a sample size of 2276, and the healthy control group had a sample size of 1595. Notably, the sample size of individuals with depression was larger than all 6 previous studies, suggesting that our work was well powered.

Subtyping Approaches

[Figure 1](#) provides an overview of the 6 different subtyping approaches developed in previous work that were implemented in the same cohort (4–9). These 6 data-driven subtyping approaches were chosen to cover input features across 3 different input feature domains, namely clinical, structural magnetic resonance imaging (sMRI), and resting-state functional MRI (rfMRI) (see [Supplemental Methods Section 5](#)). The clustering algorithms utilized are latent class analysis (LCA), high-dimensional data clustering (HDDC), k-means, subgroup-group iterative multiple model estimation (s-GIMME), and Ward's hierarchical clustering.

Statistical Analyses

Once we applied each of the 6 subtyping approaches in the same UKB cohort, our primary goal in this article was to systematically compare the resulting subtypes. In the description below, we use the term subtyping approach or approach to refer to each of the 6 studies that were adopted (see [Figure 1](#)). We use the term domain to refer to the 3 classes of features used to inform the subtyping approaches (e.g., clinical features, sMRI features, fMRI features). We use the term subject groupings to refer to the subtype label assignments across individual participants and subtypes to refer to the set of subject groupings derived from a specific subtyping approach.

Participant Assignment. To determine whether different subtyping approaches resulted in the same subject groupings (i.e., assigned the same sets of participants to the same subtypes), we calculated the adjusted Rand index (ARI) to compare each pair of subtyping approaches. Permutation testing was performed over 2000 permutations where subtype labels were permuted between participants within each approach, and the 15 ARIs between each possible pair of subtype approaches were calculated per permutation. Uncorrected *p* values were calculated for each of the 15 original ARIs compared with their matched permuted null distribution, and Bonferroni correction was used to correct for multiple comparisons (i.e., $p_{\text{uncorrected}} \times 15$, which is mathematically equivalent to dividing the alpha threshold of .05 by 15).

Fixing Number of Subtypes Across Approaches. In the stage 1 registered report, we had planned to “perform subgroup comparisons for both the optimized number of subtypes and the matched number of subtypes to the original published papers.” This plan was extended to repeat all subtyping approaches to derive a fixed number of subtypes (where possible; e.g., s-GIMME does not support researcher-specified dimensionality definition). A fixed number of

Comparing Data-Driven Subtypes of Depression


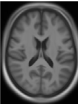
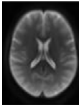
Clinical 		Structural MRI 		Resting state fMRI 	
Clinical + LCA	Clinical + HDDC	sMRI + Kmeans	sMRI + HYDRA	rfMRI + sGIMME	rfMRI + Wards
12 depression questionnaire items	17 depression and anxiety items	FSL Voxel Based Morphometry	UKB internal FreeSurfer processing	Time-series for 13 a priori ROIs	Connectivity matrix between 413 ROIs
Add 1 to ensure all natural numbers	Z-score questionnaire items	ICA → 30 covariance networks	61 cortical + 19 subcortical ROIs		Feature selection down to 150 ROIs
		12 matched ICs covering 10 networks	Mean gray matter volume (GMV)		rCCA against 17 depression questions
Latent class analysis on 12 items	HDDC on 17 items	K-means clustering on 12 network loadings	HYDRA on 80 GMVs	sGIMME on 13 ROI timeseries	Ward's clustering on 2 rCCA mode scores
Lowest AIC & BIC between 1-5 classes	Lowest BIC from 1-10 subtypes	Optimal silhouette, CH index, gap for k = 2-6	Bootstrapped ARI between 2-8 subtypes	Walktrap optimization of subtype number	VRC and silhouette score

Figure 1. Overview of 6 subtyping approaches. The blue shaded elements describe processing/intermediate steps, the orange shaded element describes the key subtyping step, and the gray shaded element describes the steps to determine the optimal number of subtypes. See [Supplemental Methods Section 5](#) for a detailed description of the subtyping approaches. AIC, Akaike information criterion; ARI, adjusted Rand index; BIC, Bayesian information criterion; CH, Calinski-Harabasz; HDDC, high-dimensional data clustering; HYDRA, heterogeneity through discriminative analysis; ICA, independent component analysis; LCA, latent class analysis; rCCA, regularized canonical correlation analysis; rfMRI, resting-state functional magnetic resonance imaging; ROI, region of interest; sGIMME, subgroup-group iterative multiple model estimation; sMRI, structural MRI; UKB, UK Biobank; VRC, variance ratio criterion.

subtypes ($k = 2,3,5,6$) was chosen, encompassing every k solution found in the main results across the 6 approaches. The subject agreement ARI comparison was repeated for each k to determine whether relative performance changed. The reason for this change from the stage 1 submission was that it is possible that some of the comparative analyses may systematically vary as a function of the number of subgroups (for example, as a result of differences in power). Therefore, systematically varying k for all approaches allowed for like-for-like comparisons.

Subtype Assignment After Swapping Inputs Between Approaches (Added After Stage 1). If participant assignments differ between approaches, this may be either due to differences in the input features or due to differences in the subtyping approach. To assess which of these sources of potential differences was driving results, we repeated each of the subtyping approaches on input features from each of the other approaches. Notably, rfMRI+s-GIMME was excluded from this comparison because none of the other subtyping approaches can take time-series data as input. Furthermore, LCA was not performed on sMRI or rfMRI inputs because it requires a limited set of natural numbers as inputs.

Subtype Stability Across Bootstraps. To determine the stability of the resulting subtypes within each subtyping approach, we repeated the identical subtyping approach across 100 bootstraps. Only the clustering algorithm was repeated (i.e., the steps in orange shading in [Figure 1](#)), and the subtype number was set to match the original results. We did

not repeat all processing/intermediate steps performed before clustering or the optimization of the number of subtypes (i.e., steps in blue and gray shading in [Figure 1](#)) to keep the comparisons fair between approaches and to keep computational resources feasible. For each bootstrap, 80% of the participants were drawn from the original sample (matched 80% across approaches), and the ARI between the bootstrapped subtypes and the full sample subtypes (excluding participants missing from the random 80% bootstrap) was calculated. Confidence intervals from the bootstrap distributions were used to compare stability across subtyping approaches.

Null Model. To determine the effect of random noise on clustering solutions, we synthesized a null dataset by sampling each feature from a Gaussian distribution. We created 4 sets of synthesized null data with numbers of features representative of the empirical data (i.e., 2 features similar in size to canonical correlation analysis features from rfMRI+Ward, 12 features similar in size to the clinically informed subtyping approaches and sMRI+kmeans, 17 features similar in size to clinical+HDDC, and 80 features similar in size to sMRI+HYDRA), all with 2276 cases and 1595 controls. The synthesized data were intended as a null comparison, and therefore no systematic differences between cases and controls or among groups of cases were simulated. Each subtyping method was performed on all 4 synthesized feature sets and compared using a Student's t test to the synthesized feature set matched in size. Bonferroni correction was applied to p values resulting from the Student's t test to control for 5 comparisons. Bootstrapping was performed to assess the

stability of subtype estimation in null data (as above, drawing 80% of null participants for each of the 100 bootstraps). We repeated the clustering algorithm (the step in orange in Figure 1) for each bootstrap. After each bootstrap, we recorded the silhouette score and ARI to develop a null distribution of clustering solutions (separately for each feature set), which were compared against the true silhouette score and ARI from empirical data. This null modeling approach allowed us to differentiate the impact of analytical choices from the impact of differences in source data. Null model testing was not performed for rfMRI+s-GIMME because s-GIMME requires time-series data, and none of the other subtyping approaches are suited for such data. Importantly, the purpose of the null is to compare the subtyping approaches, and therefore we did not create a separate null dataset to input into s-GIMME.

Subtype Stability Across Random Seed Initializations (Added After Stage 1). Several of the subtyping approaches are nondeterministic and may therefore vary as a function of the random seed initialization, namely k-means and HDDC. To determine the stability of the resulting subtypes as a function of initialization, we repeated the identical clustering algorithm and determination of optimal cluster number across 100 iterations using the total participant sample in each iteration but varying the seed initialization (0:99). Both the clustering algorithm and optimization of the number of subtypes (i.e., steps in orange and gray shading in Figure 1) were repeated. The optimal cluster number and silhouette score were recorded for each iteration to assess stability.

Sensitivity to Subgroup Differences in Unseen Variables. A common challenge in data-driven subtyping analyses is that algorithms will optimize their cost function to output subtypes, but these resulting subject groupings may or may not be sensitive to meaningful clinical or biological distinctions. Therefore, it is common for data-driven subtyping efforts to compare the resulting subtypes on unseen variables (i.e., measures that are thought to capture meaningful sources of variability between individuals with depression but that were not used by the method to derive the subtypes). Subtypes were not compared on seen variables (i.e., the features that informed the subtypes), but it is important to note that some of the variables are very similar to the seen variables (such as the same questionnaire items obtained on previous visits) and therefore should be treated with caution.

In each of the 6 articles, a different set of validation/characterization methods was used. Informed by these previous approaches, we compared subtypes based on categories of demographics, general health, clinical characteristics, neuroimaging, trauma, cognition, genetics, and inflammation (see Supplemental Methods Section 6). An important advantage of our approach compared with previous studies that adopted a single subtyping approach is that we can directly compare sensitivity to each category between subtyping approaches and examine which subtyping approaches are more sensitive to subgroup differences in each category. Sensitivity analyses were performed using all available data (i.e., excluding participants with missing variables). Only phenotypes with data for at least 50% of the participants were included.

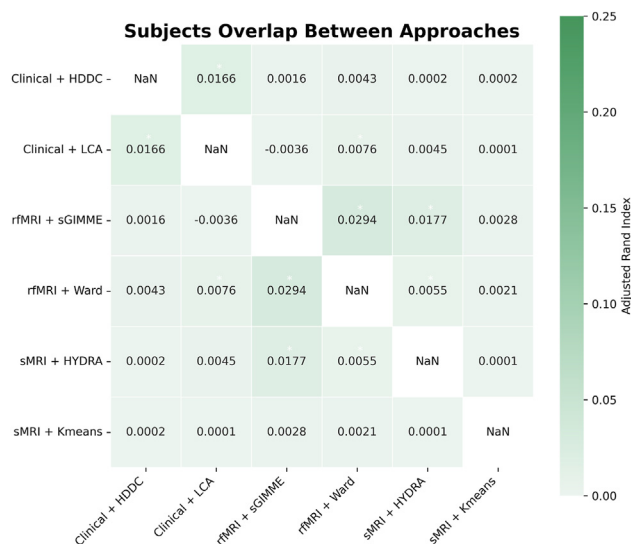


Figure 2. Agreement of subject groupings between approaches. We calculated the adjusted Rand index (ARI) between the cluster assignments of every approach to measure how much the cluster assignments agree above chance. An ARI of 0 indicates overlap above chance with the highest possible ARI value equal to 1. Significant findings (based on permutation testing) are highlighted with an asterisk (*). For a description of the subtyping approaches, see Figure 1. HDDC, high-dimensional data clustering; HYDRA, heterogeneity through discriminative analysis; LCA, latent class analysis; NaN, not a number; rfMRI, resting-state functional magnetic resonance imaging; sMRI, structural MRI.

RESULTS

Participant Assignment

The ARI revealed minimal similarity of depression subtypes between subtyping approaches (Figure 2). Although the maximum ARI was 0.0294, suggesting very limited subtype similarity, permutation testing indicated that the 2 clinically driven approaches overlapped significantly above chance ($p = .0075$), and the 2 rfMRI driven approaches overlapped significantly above chance ($p = .0075$), consistent with our hypothesis of higher within-domain subtype overlap. Furthermore, rfMRI+Ward and sMRI+HYDRA as well as rfMRI+Ward and clinical+LCA overlapped significantly ($p = .0450$ and $p = .0075$, respectively). Taken together, these results revealed that the subtyping approaches did not achieve the same subject groupings even when applied within the same cohort. Table S8 provides the cluster evaluation criterion for each approach, and Table S9 shows the sample sizes of resulting subtypes.

Fixing Number of Subtypes Across Approaches

Similar findings to Participant Assignment above were observed (Figure S3) when fixing the cluster number for all approaches, indicating that differences in cluster number did not drive this disagreement.

Subtype Assignment After Swapping Inputs Between Approaches

The lack of subtype assignment similarity observed in Figure 2 could be due to differences in the input features and/or

Comparing Data-Driven Subtypes of Depression

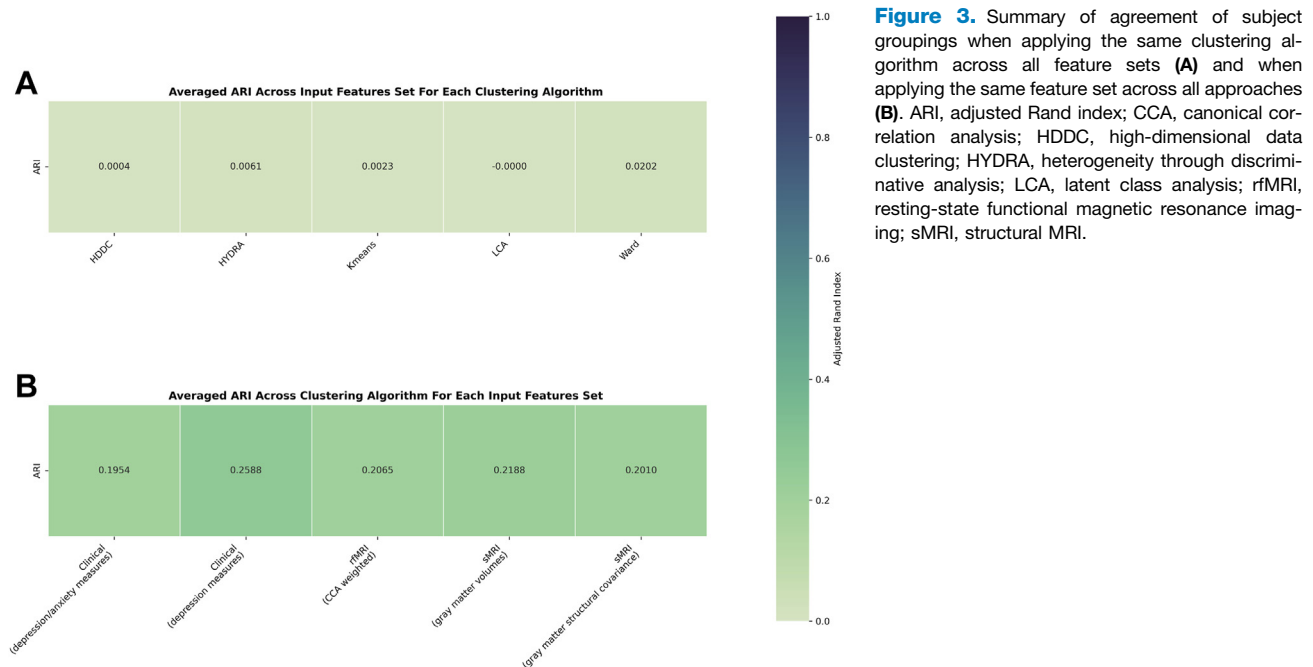


Figure 3. Summary of agreement of subject groupings when applying the same clustering algorithm across all feature sets (**A**) and when applying the same feature set across all approaches (**B**). ARI, adjusted Rand index; CCA, canonical correlation analysis; HDDC, high-dimensional data clustering; HYDRA, heterogeneity through discriminative analysis; LCA, latent class analysis; rsfMRI, resting-state functional magnetic resonance imaging; sMRI, structural MRI.

differences in the subtyping approaches. We repeated the subtyping approaches on all possible feature sets where possible. The results revealed substantially increased ARI across subtyping approaches when they were performed on the same feature set (ARI range, -0.0356 to 0.6834) (Figure 3 and Figure S4). However, ARIs remained close to 0 across feature sets within the same subtyping approach (ARI range, -0.0052 to 0.1711) (Figure 3 and Figure S4). These results revealed that differences in subtype assignments were primarily driven by the input features rather than by the cluster algorithm.

Subtype Stability Across Bootstraps

When we assessed the stability within each approach, subtyping approaches achieved a median ARI across bootstraps of $ARI = 0.36$ – 0.89 (Figure 4). Although the highest median stabilities were observed for clinical+LCA ($ARI = 0.89$) and sMRI+kmeans ($ARI = 0.89$), the confidence intervals for all subtyping approaches overlapped such that no subtyping approach was significantly more stable than any other (Table S10).

Null Model

To determine the effect of random noise on clustering solutions, we synthesized null data from a Gaussian distribution and repeated the bootstrapping for stability comparisons, and we calculated the silhouette scores for cluster differentiation comparisons. This null modeling approach allows us to differentiate the impact of analytical choices from the impact of differences in source data. We found that most of the subtype approaches were more stable than the null ($p < .0001$) (Figure 5 and Table S11). However, Ward hierarchical

clustering did not differ significantly from the null at the matched feature dimension (2 features; $p = .3438$) (Figure 5C). Regarding the silhouette score comparisons, all silhouette scores from true data were higher than the null silhouette scores, although the result for HDDC overlapped with the null distribution (Figure 5F).

Subtype Stability Across Random Seed Initializations (Added After Stage 1)

The subtyping approaches of HDDC and k-means are initialized with a random seed. We repeated these 2 approaches on 5 input feature sets (except for time-series inputs) to test the sensitivity of the determination of the optimal k number of subtypes to the initialization. The results (Figure S5) revealed that HDDC was relatively unstable for clinical input features sets (optimal k ranging 3–10) and for gray matter volumes (optimal k ranging 6–9), with greater stability for gray matter covariance and functional connectivity (optimal k = 2). K-means clustering achieved strong stability of the optimal k across all feature sets irrespective of initialization (Figure S5).

Sensitivity to Subgroup Differences in Unseen Variables

Subtypes were compared based on categories of demographics, general health, clinical characteristics, neuroimaging, trauma, cognition, genetics, and inflammation to assess sensitivity. The clinically driven subtypes were more sensitive to clinical data, and the imaging-driven subtypes were more sensitive to imaging data as expected, with little cross-domain sensitivity for all approaches. Figure 6B indicates that clinical+HDDC was more sensitive to differences

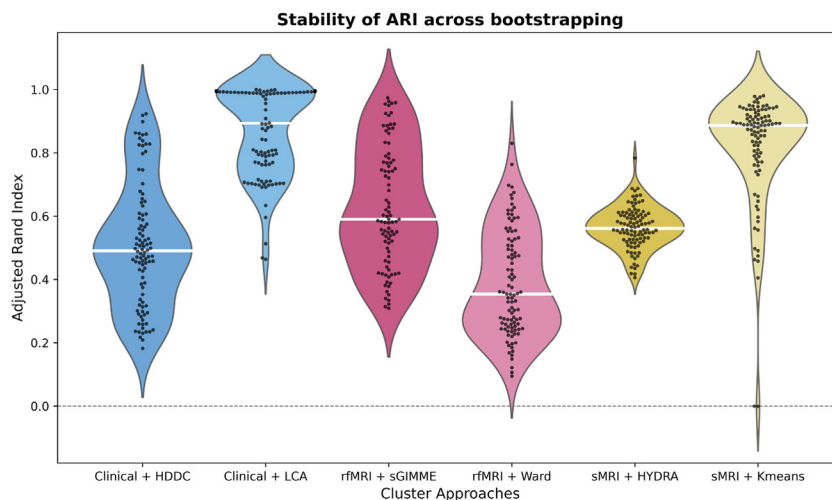


Figure 4. The stability of each approach was assessed using the adjusted Rand index (ARI) between each of the 100 bootstraps using 80% of the data and the original subtype assignments based on the full dataset (i.e., each bootstrap is 1 data point). The participant subselection for each bootstrap was identical for all approaches except for heterogeneity through discriminative analysis (HYDRA) because HYDRA already calculates bootstraps, and so those bootstraps were used for this analysis. For a description of the subtyping approaches, see Figure 1. HDDC, high-dimensional data clustering; LCA, latent class analysis; rfMRI, resting-state functional magnetic resonance imaging; sMRI, structural MRI.

in depression and anxiety questions while clinical+LCA was sensitive to measures of physical health like age, sex, body mass index (BMI), and overall health (Table S12 has all results). The subtypes of clinical+HDDC, rfMRI+Ward, and sMRI+kmeans found a single nucleotide polymorphism (SNP) difference compared with the control group (Figure S6 has details).

DISCUSSION

In this study, we aimed to directly compare depression subtypes identified using 6 different data-driven approaches in the same cohort to assess whether the lack of consensus on depression subtypes in the existing literature is driven by differences in study cohorts, subtype approaches, or input features. In summary, our findings revealed minimal participant agreement between approaches (Figure 2), even between approaches based on input features from the same domain (clinical/sMRI/rfMRI). The similarity between approaches increased when matched input features were used (Figure 3), suggesting that the choice of input features is a critical driver of resulting depression subtypes, even within general feature domains. Furthermore, each approach driven by its original input features was internally relatively stable across bootstraps (Figure 4), outperformed null comparisons (Figure 5), and showed sensitivity to multiple phenotypes (Figure 6). The approaches tended to be more sensitive to phenotypes in their own domain, indicating that each approach was subtyping on different sources of heterogeneity. Therefore, this study has identified core drivers of inconsistency between subtyping approaches as well as key pitfalls for future work to avoid.

We discovered that accounting for participant variability was not sufficient to overcome the almost complete inconsistency between results of the different approaches, even within the same data domain. Further investigation revealed input feature selection and processing choices as the main driver of subtyping results even within the same domains (Figure 3B) and a relatively smaller contribution of clustering

approach (Figure 3A). This finding of the importance of feature selection and processing choices is similar to previous work outside the subtyping field (17) and emphasizes the challenge of analytical flexibility (18,19). Therefore, we recommend that future work incorporate assessments of subtype agreement across input feature choices and include rigorous comparisons to other published subtypes.

Despite comparative disagreement between subtypes, we found the subtyping approaches to be internally consistent. Most of the approaches had relatively high stability across bootstraps (Figure 4) and performed well compared with null data (Figure 5). This stability is an improvement over previous critical assessments of depression subtypes in which relatively lower bootstrap stability (10) have been observed in at least 1 subtyping approach. One possible reason for the improved stability observed is the relatively large sample size ($n = 2276$), which likely improves robustness against potential outliers. Notably, we observed volatility of Ward's hierarchical clustering to small changes in participants across bootstraps, reducing its stability compared with other approaches (Figure 4) and compared with null data (Figure 5). This finding is consistent with previous evidence for volatility of Ward's hierarchical clustering when using a relatively small number of input features (20–22). Therefore, in future work, investigators may wish to avoid Ward's hierarchical clustering (especially in small feature sets) unless they have reason to believe that subtypes are hierarchically nested. In general, we recommend that future work maximize sample sizes and incorporate consistency analyses across bootstraps and against null data to rigorously determine the reliability of subtypes. Notably, the stability of subtypes on low-dimensional unstructured null data displayed high median ARI for all clustering approaches, suggesting that comparisons to null data are especially important when there are a small number of input features.

In addition to moderate to good internal consistency, many of the subtypes were sensitive to relevant phenotypes. The subtypes were much more sensitive to the phenotypes in their own domain (clinically driven subtypes being more sensitive to

Comparing Data-Driven Subtypes of Depression

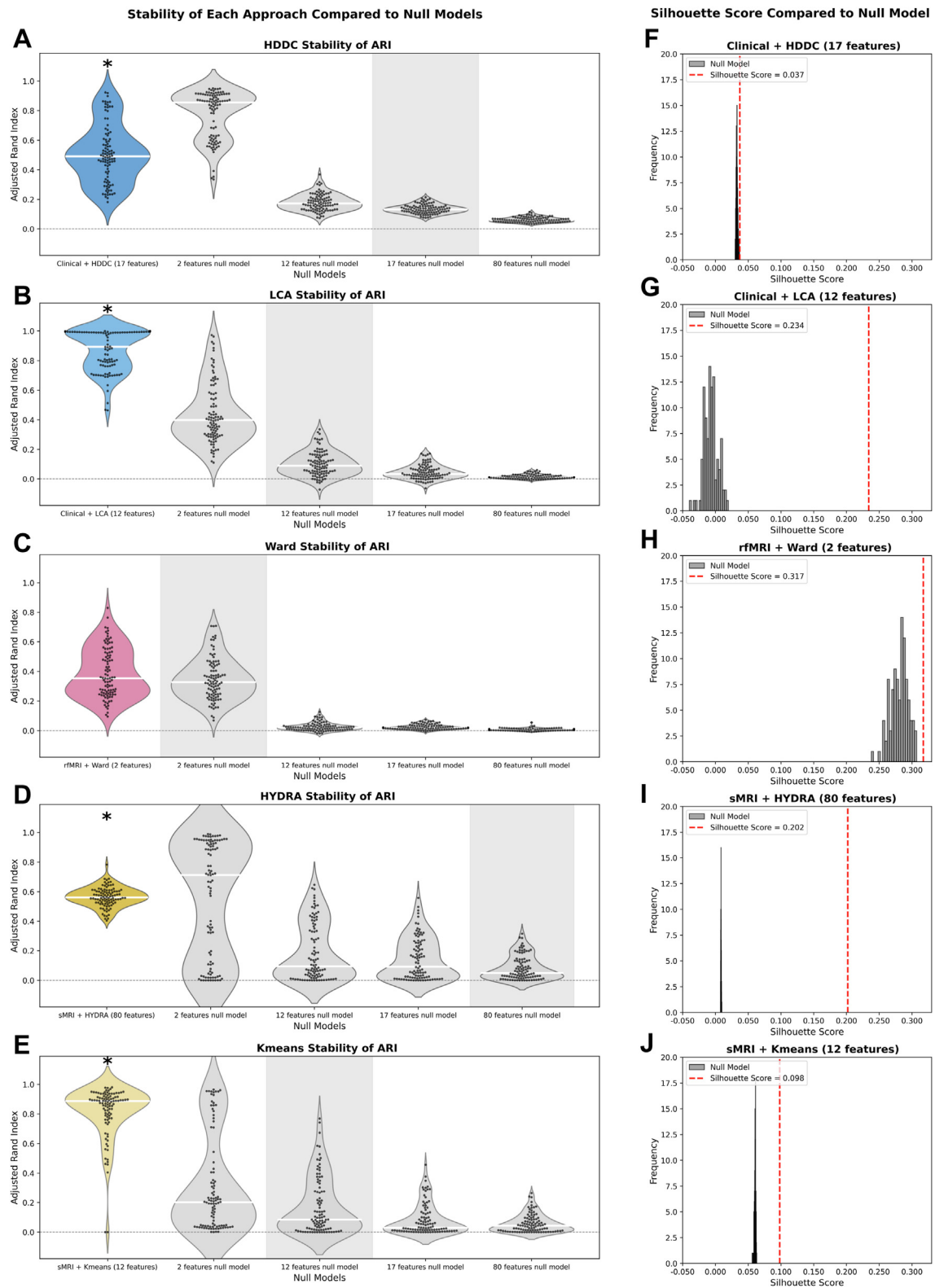


Figure 5. Approaches compared with the null. (A–E) Stability of each approach, adjusted Rand index (ARI) over 100 bootstraps compared with nulls. The null feature set most comparable to the true feature set is highlighted in gray. (F–J) True silhouette score of each approach’s clustering result compared with the null (100 bootstraps). *Indicates significantly different than simulated null highlighted in gray. HDHC, high-dimensional data clustering; HYDRA, heterogeneity through discriminative analysis; LCA, latent class analysis; rfMRI, resting-state functional magnetic resonance imaging; sMRI, structural MRI.

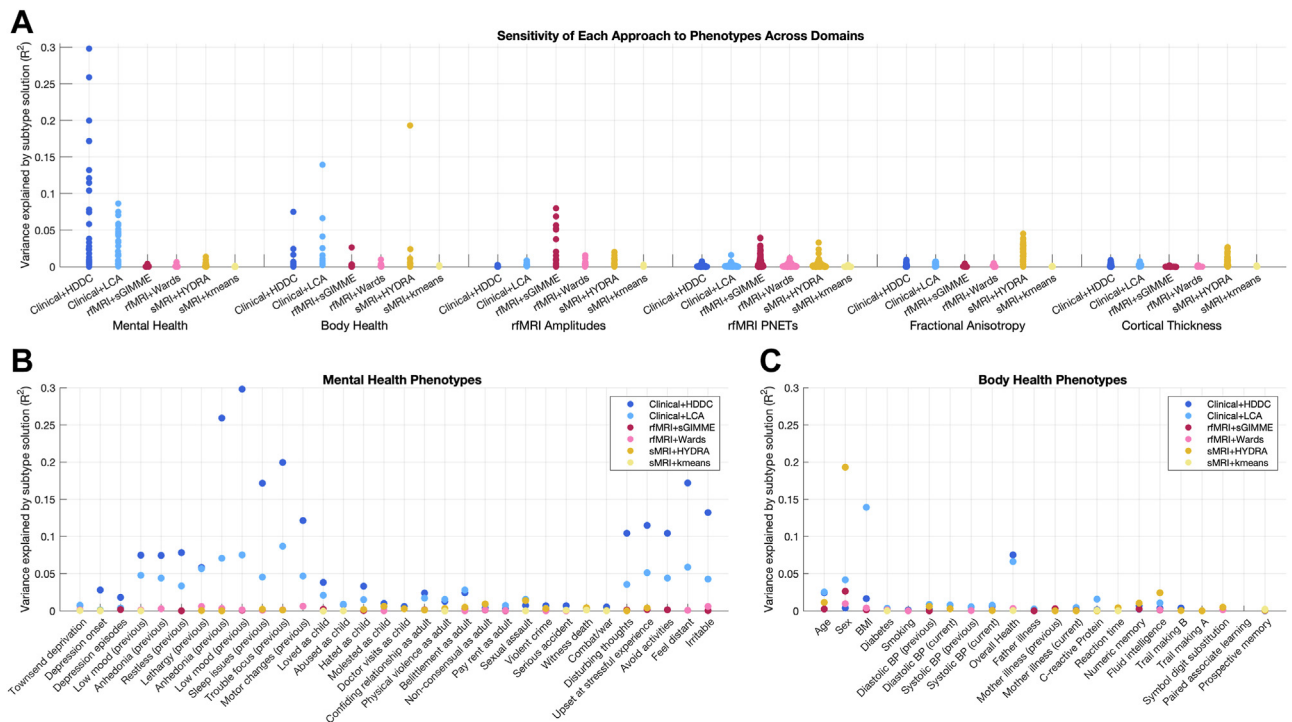


Figure 6. Variance explained of each approach across out-of-analysis features. Panel (A) provides the variance explained by each subtyping solution for all phenotype domains (mental health, body health, resting-state independent component analysis [ICA] network amplitudes and partial correlations, fractional anisotropy from diffusion-tensor imaging [DTI], and cortical thickness from FreeSurfer). Panels (B) and (C) denote the variance explained of all mental health and body health phenotypes that have at least 1 significant comparison, respectively. BP, blood pressure; HDDC, high-dimensional data clustering; HYDRA, heterogeneity through discriminative analysis; LCA, latent class analysis; PNET, partial network correlation; rfMRI, resting-state functional magnetic resonance imaging; sMRI, structural MRI.

clinical data and neuroimaging subtypes being more sensitive to their neuroimaging domain), indicating that the subtyping approaches were likely subtyping on different sources of heterogeneity (13). This explains why these approaches can be internally consistent and sensitive but not agree with each other. Based on that evidence, this study provides an explanation for why previous subtyping efforts do not converge despite each article showcasing internal stability and sensitivity. Differences in input feature sets lead to entirely different but potentially internally consistent subtypes. Therefore, we recommend that future work systematically investigate and compare different sources of heterogeneity and/or consider multimodal input features to potentially optimize cross-domain consistency.

Most of the sensitivity results explained relatively little variance ($R^2 < 0.1$). We note that the clinical+HDDC approach was sensitive to a range of mental health variables, in particular those related to anhedonia and trauma responses (Figure 6B). Subtypes differed significantly on sex for sMRI+HYDRA (Figure 6C). Notably, the original study (7) controlled for sex differences, which was not implemented here because several of the original articles wanted to evaluate sex differences between subtypes. Lastly, BMI differed significantly between the subtypes identified with clinical+LCA (Figure 6C), which replicates their findings (4).

Lastly, it is important to note that some data-driven subtyping approaches are nondeterministic, meaning that the results are

influenced by a random initiation and therefore may change over repeat runs. Importantly, our findings showed that of the nondeterministic approaches, HDDC changed the optimal cluster number on each run for many input feature sets (Figure S5). Therefore, we recommend that future work that uses nondeterministic approaches to assess consistency across iterations.

We should note some limitations of this work. Firstly, we diverged from the original studies in some decisions (see [Supplemental Methods Section 5](#)) to maximize the comparability of findings across approaches. Therefore, any differences in findings from the original articles should not be overinterpreted. Secondly, although we adopted a clinical definition for the depression cohort (16), the degree of depression severity in the UKB cohort is relatively low, which may impact subtype performance. However, this should impact all approaches similarly and therefore is not expected to impact comparisons.

Conclusions

This study revealed key drivers of disagreement between subtyping approaches. Because input feature set is a critical driver of disagreement, we recommend that investigators assess subtype agreement across input feature space in future work and compare their work with previous subtypes. Given the impact of clustering algorithm on subtype solutions and the possibility of clustering algorithms to embed structure in noise at low dimensions, we suggest that in future work,

investigators evaluate the stability of their results across bootstraps and compared with null data. Finally, given that subtyping approaches in the current study were sensitive only to the source of heterogeneity that they parsed, we recommend that future work investigate several sources of heterogeneity.

ACKNOWLEDGMENTS AND DISCLOSURES

JDB is supported by the National Institute of Mental Health (Grant Nos. R01 MH128286 and R01 MH132962). Computations were performed using the facilities of the Washington University Research Computing and Informatics Facility, which has received funding from National Institutes of Health S10 program (Grant Nos. 1S10OD025200-01A1 and 1S10OD030477-01).

We thank UKB and the UKB participants for making the resource data possible and the data processing team at Oxford University for producing the shared processed data. This research was performed under UKB application No. 47267. We thank Mahshid Naghashzadeh for her work on understanding the methodology of the Wen *et al.* study. We thank Dr. Price and Dr. Liston for their support in helping us properly apply the methodologies of the Price *et al.* and Drysdale *et al.* studies, respectively.

Data are released through the UKB and downloaded locally. We confirm that none of the analyses described in the [Methods and Materials](#) section were started prior to submission of this Registered Report (Stage 1). All analysis code for this article is at <https://github.com/PersonomicsLab/CompareSubtypes>. UKB data (14,16) are available following an access application process; for more information, please see: <https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access>.

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

ARTICLE INFORMATION

From the Department of Radiology, Washington University School of Medicine, St. Louis, Missouri (KH, SJ, LB, FA, PL, AS, JDB); Department of Internal Medicine, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand (SJ); University of Amsterdam, Amsterdam, the Netherlands (LB); University of Minnesota, Minneapolis, Minnesota (FA); and Institute for Informatics, Data Science, & Biostatistics, Washington University in St. Louis, St. Louis, Missouri (AS).

LB is currently affiliated with the Department of Radiology, Washington University School of Medicine, St. Louis, Missouri.

Address correspondence to Kayla Hannon, B.S., at khannon@wustl.edu, or Janine D. Bijsterbosch, Ph.D., at janine.bijsterbosch@wustl.edu.

Received Apr 27, 2023; revised Jan 21, 2025; accepted Feb 13, 2025.

Supplementary material cited in this article is available online at <https://doi.org/10.1016/j.bpsgos.2025.100473>.

REFERENCES

- World Health Organization (2017): Depression and other common mental disorders: Global Health Estimates. Available at: <https://www.who.int/publications/i/item/depression-global-health-estimates>. Accessed March 28, 2025.
- Kennis M, Gerritsen L, van Dalen M, Williams A, Cuijpers P, Bockting C (2020): Prospective biomarkers of major depressive disorder: a systematic review and meta-analysis. *Mol Psychiatry* 25:321–338.
- Beijers L, Wardenaar KJ, van Loo HM, Schoevers RA (2019): Data-driven biological subtypes of depression: Systematic review of biological approaches to depression subtyping. *Mol Psychiatry* 24:888–900.
- Lamers F, de Jonge P, Nolen WA, Smit JH, Zitman FG, Beekman ATF, Penninx BWJH (2010): Identifying depressive subtypes in a large cohort study: Results from the Netherlands study of depression and anxiety (NESDA). *J Clin Psychiatry* 71:1582–1589.
- Maglanoc LA, Landrø NI, Jonassen R, Kaufmann T, Córdova-Palomera A, Hilland E, Westlye LT (2019): Data-driven clustering reveals a link between symptoms and functional brain connectivity in depression. *Biol Psychiatry Cogn Neurosci Neuroimaging* 4:16–26.
- Price RB, Gates K, Kraynak TE, Thase ME, Siegle GJ (2017): Data-driven subgroups in depression derived from directed functional connectivity paths at rest. *Neuropsychopharmacology* 42:2623–2632.
- Wen J, Fu CHY, Tosun D, Veturi Y, Yang Z, Abdulkadir A, *et al.* (2022): Characterizing heterogeneity in neuroimaging, cognition, clinical symptoms, and genetics among patients with late-life depression. *JAMA Psychiatry* 79:464–474.
- Yang XH, Huang J, Lan Y, Zhu CY, Liu XQ, Wang YF, *et al.* (2016): Diminished caudate and superior temporal gyrus responses to effort-based decision making in patients with first-episode major depressive disorder. *Prog Neuropsychopharmacol Biol Psychiatry* 64:52–59.
- Drysdale AT, Grosenick L, Downar J, Dunlop K, Mansouri F, Meng Y, *et al.* (2017): Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat Med* 23:28–38.
- Dinga R, Schmaal L, Penninx BWJH, van Tol MJ, Veltman DJ, van Velzen L, *et al.* (2019): Evaluating the evidence for biotypes of depression: Methodological replication and extension of Drysdale *et al.* *Neuroimage Clin* 22:101796.
- Grosenick L, Shi TC, Gunning FM, Dubin MJ, Downar J, Liston C (2019): Functional and optogenetic approaches to discovering stable subtype-specific circuit mechanisms in depression. *Biol Psychiatry Cogn Neurosci Neuroimaging* 4:554–566.
- Tozzi L, Zhang X, Pines A, Olmsted AM, Zhai ES, Anene ET, *et al.* (2024): Personalized brain circuit scores identify clinically distinct biotypes in depression and anxiety. *Nat Med* 30:2076–2087.
- Hannon K, Easley T, Zhang W, Lew D, Thornton V, Sotiras A, *et al.* (2022): Heterogeneity in depression: Evidence for distinct clinical and neurobiological profiles. *medRxiv* <https://doi.org/10.1101/2022.12.07.22283225>.
- Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, *et al.* (2015): 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.
- Littlejohns TJ, Holliday J, Gibson LM, Garratt S, Oesingmann N, Alfaro-Almagro F, *et al.* (2020): The UK biobank imaging enhancement of 100,000 participants: Rationale, data collection, management and future directions. *Nat Commun* 11:2624.
- Smith DJ, Nicholl BI, Cullen B, Martin D, Ul-Haq Z, Evans J, *et al.* (2013): Prevalence and characteristics of probable major depression and bipolar disorder within UK Biobank: Cross-sectional study of 172,751 participants. *PLoS One* 8:e75362.
- Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, *et al.* (2020): Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* 582:84–88.
- Li X, Bianchini Esper N, Ai L, Giavasis S, Jin H, Feczko E, *et al.* (2024): Moving beyond processing- and analysis-related variation in resting-state functional brain imaging. *Nat Hum Behav* 8:2003–2017.
- Bijsterbosch J (2022): The role of analytical flexibility in determining mental health biomarkers. *Biol Psychiatry Glob Open Sci* 2:316–318.
- Saunders A, Ashlock D, Houghten S (2018): Hierarchical clustering and tree stability. St. Louis, Missouri: Presented at the 2018 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology.
- Mucha H-J (2007): On validation of hierarchical clustering. In: Decker R, Lenz H-J, editors. *Advances in Data Analysis*. Berlin: Springer, 115–122.
- Smith SP, Dubes R (1980): Stability of a hierarchical clustering. *Pattern Recognit* 12:177–187.