Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Identify Unique Subtypes in Recurrent Major Depression Disorder

By

Zixi Zhu
MPH

Biostatistics and Bioinformatics Department

Zhaohui Qin
Committee Chair

Yijuan Hu
Committee Member

Identify Unique Subtypes in Recurrent Major Depression Disorder

By

Zixi Zhu

MPH
Emory University
2024

Thesis Committee Chair: Zhaohui Qin, PHD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
MPH
in Biostatistics and Bioinformatics Department
2024

Abstract

Identify Unique Subtypes in Recurrent Major Depression Disorder

By Zixi Zhu

Major Depressive Disorder (MDD) is a multifaceted mental health condition with varying symptomatology and treatment responses among individuals. This study aims to elucidate the heterogeneity within MDD by identifying distinct subcategories characterized by symptom severity and treatment resistance. Leveraging data from 922 patients, we employed hierarchical clustering to reveal significant differences in anxiety and depression symptoms, as identified through the PHQ-9 and GAD-7 questionnaires. Furthermore, the application of Hadamard Autoencoders facilitated the imputation of missing data and feature extraction, highlighting the importance of specific questionnaire items in distinguishing between MDD subtypes. Our findings suggest the existence of a severe MDD subtype, potentially analogous to Treatment-Resistant Depression (TRD), exhibiting pronounced anxiety and depression symptoms and differential responses to conventional treatments. The identification of these subcategories underscores the need for personalized treatment approaches in MDD management. Future research should integrate treatment response data to further refine MDD subtyping and enhance therapeutic strategies.

Identify Unique Subtypes in Recurrent Major Depression Disorder

By

Zixi Zhu

MPH
Emory University
2024

Thesis Committee Chair: Zhaohui Qin, PHD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Biostatistics and Bioinformatics Department
2024

Introduction

Major Depressive Disorder (MDD) presents as a multifaceted phenotype, with its familial patterns largely attributed to genetic underpinnings[1]. It is associated with a myriad of physiological changes, including neurohormonal secretion fluctuations, structural alterations in brain anatomy, and deviations in immune function and inflammatory markers. This disorder is the product of an intricate interplay between genetic predispositions and environmental factors. Whole transcriptome analyses, which assess the mRNA expression levels across genes in pertinent tissues, have illuminated altered expression patterns in diseases. These changes encapsulate the influence of both common and rare genetic variations, environmental contributors, the inherent effects of the disease itself, and the interplay between genetics and environment.

Depression's complexity allows for its categorization into clinically relevant subtypes based on symptomatology, etiology, progression, and comorbidity with other psychiatric conditions[2]. Among these, Treatment-Resistant Depression (TRD) stands out as a distinct subtype in MDD, characterized by its unique treatment response profile[3].

The interaction between MDD and the immune system is particularly noteworthy. The disease course and onset are potentially influenced by immune cells and molecules through various mechanisms. These include pro-inflammatory signaling pathways, alterations in the hypothalamic-pituitary-adrenal (HPA) axis, dysregulation of serotonergic and noradrenergic

neurotransmitters, neuroinflammation, and immune dysfunction within the meninges[4].

This study builds upon a previous whole-transcriptome analysis of MDD in a cohort of 922 individuals of European ancestry (463 cases and 459 controls), utilizing RNA sequencing data from whole blood samples[5]. Our objectives are threefold: firstly, to identify potential subcategories within MDD that exhibit distinct phenotypic behaviors; secondly, to investigate differences in immune cell fractions among these subcategories; and thirdly, to explore variations in RNA expression and alternative splicing events within the subtypes. This multifaceted approach aims to deepen our understanding of MDD's biological underpinnings and pave the way for tailored therapeutic strategies.

Methods

Data Acquisition and Utilization

The dataset utilized in our study was derived from an extensive recruitment and assessment effort conducted by an external research team, in accordance with IRB-approved protocols. The team collaborated with Knowledge Networks, Inc. (Menlo Park, CA), employing a national panel of approximately 60,000 individuals to recruit both cases (individuals with recurrent or chronic MDD) and controls. This process entailed a rigorous online screening, comprehensive telephone interviews, and detailed clinical assessments to ensure a robust selection of participants, culminating in a cohort of 922 individuals (463 cases and 459 controls).

The RNA sequencing data, integral to our analysis, was generated by the same team. Blood samples were meticulously processed, employing state-of-the-art techniques for RNA extraction and sequencing to ensure the highest quality of data. This included the use of PaxGene tubes for RNA stabilization, GLOBINclear Kits for globin RNA depletion, and Illumina TrueSeq kits for the RNA purification and library preparation processes[6]. Sequencing was conducted on the Illumina HiSeq 2000 platform, yielding detailed transcriptomic profiles across the cohort.

Our study leverages this rich dataset to further explore and elucidate the molecular underpinnings of MDD, building upon the foundational work established by the original researchers.

Alignment and quantification of RNA-seq reads

In this study, we prioritized the initial FASTQ file from each sample for analysis. Quality control and summarization of per-sample statistics were adeptly managed using MultiQC, which generated an interactive HTML report for comprehensive review[7]. The integrity of the RNA-seq data was scrutinized using fastp (Version 0.23.2), leading to the exclusion of sample LD0053 due to a compromised FASTQ file.

Alignment of RNA-seq reads to the GRCh38 primary assembly, annotated with UCSC genes, was accomplished using STAR (version 2.7.10b)[8]. This process employed a two-pass mapping strategy (–twopassMode Basic) and incorporated strand specificity based on intron motifs (--outSAMstrandField intronMotif --outFilterIntronMotifs RemoveNoncanonical) to

ensure accurate alignment. Furthermore, STAR was also utilized to quantify gene counts (--quantMode GeneCounts), providing a detailed account of gene expression levels across samples.

Cell type deconvolution and validation

To estimate cell type fractions within each sample, we employed CIBERSORTx[9], inputting transcripts per million (TPM) values in accordance with the guidelines provided in the tool's documentation. The reference dataset utilized for this analysis was immunoStates[10], which facilitated the identification of twenty distinct cell types. Additionally, this dataset provided the set of cell type-specific (CTS) genes crucial for the deconvolution process.

Exon read counts

To quantify exon reads, we utilized the summarizeOverlaps() function from the GenomicAlignments package. Annotation was facilitated by a GTF file, which was constructed using the Table Browser feature of the UCSC Genome Browser.

Missing value imputation and dimensionality reduction

Our clinical dataset comprised 921 samples(462 MDD samples and 459 control samples) and 967 clinical variables, encompassing results from both online screening and telephone interviews. We excluded variables with over 20% missing values, retaining 438 variables for analysis. Notably, 86 of these variables were exclusive to samples with MDD.

Traditionally employed in denoising and inpainting within image processing, the Hadamard Autoencoder was adapted in our study for missing value imputation and dimensionality reduction, leveraging Self-Supervised learning techniques. This approach aimed to

reconstruct unobserved dataset entries by utilizing the spatial information of adjacent observed entries, as defined by a specific cost function[11].

For the development of our autoencoder network, we opted for PyTorch, given its robust capabilities in deep learning. Data normalization between 0 and 1 was achieved using the MinMaxScaler() function from the scikit-learn library. Our network architecture consisted of a two-layer deep autoencoder, configured with a batch size of 32, a learning rate of 0.0025, and employed the Adam optimizer. Non-linear activation was introduced via ReLU functions in hidden layers and a Sigmoid function in the output layer to accommodate the normalized data range. Training was conducted using mini-batch gradient descent.

To prepare the data matrix for analysis and model validation, rows with missing values were omitted. For testing, 20% of the original entries were randomly set to zero, simulating missing-at-random (MAR) conditions, and were masked during training. To enhance data robustness, each row in the dataset was duplicated 20 times, with 40% of original entries in each copy being randomly replaced with the column mean, providing an additional supervisory signal for the model.

Ultimately, one Hadamard Autoencoder was trained using all 921 samples, addressing missing values across 438 variables. The Mean Absolute Error (MAE) for imputation with this model was recorded at 0.0132, significantly outperforming the mean imputation MAE of 0.0354. A separate Hadamard Autoencoder was trained using 462 MDD samples to impute

missing values in the 86 MDD-specific variables, resulting in 64 encoded features for MDD samples. The MAE for this MDD-specific imputation was 0.0229.

Statistical analysis

Among the 462 samples with Major Depressive Disorder (MDD), a substantial proportion, 77.7% (359 samples), were from female participants. To explore the underlying structure of the data, we applied hierarchical clustering to the 64 encoded features derived from the Hadamard Autoencoder. This analysis was performed separately for female and male samples, utilizing the complete linkage method to assess similarities between clusters.

Feature Ranking

Following data preprocessing, a total of 524 features remained to characterize the 462 Major Depressive Disorder (MDD) samples. Our approach to understanding these features involved leveraging machine learning models to quantify and rank their importance, aiming to identify the most effective feature extraction methods that succinctly capture the essence of each cluster. Features deemed most critical were those with the highest importance scores.

To ascertain feature importance, we employed permutation importance—a technique that evaluates the decrease in model performance when a given feature's values are randomly shuffled. This method is particularly favored for its straightforward interpretation within the context of machine learning models. However, it's noteworthy that permutation importance may be less effective in situations of multicollinearity, where features exhibit high correlation or linear dependence.

Multicollinearity can obscure the unique contribution of individual features, thereby complicating efficient data analysis and model interpretation[12]. To address this, we initiated our analysis by mitigating multicollinearity through the use of phi correlations for binary features and Spearman rank-order correlations for continuous and ordinal features. Hierarchical clustering based on these correlations allowed us to identify groups of collinear features, from which we selected a representative feature from each group, thus forming a new feature set devoid of multicollinearity concerns.

This refined feature set served as the input for our cluster identification model, which was built using an XGBoost classifier tasked with predicting cluster membership. The dataset was split into training and testing subsets through a stratified approach to ensure representative distribution of clusters. Feature importance was then determined using the ELI5 library to perform permutation importance calculations. This process was repeated fifty times with the XGBoost classifier to average the importance scores, ensuring robustness in our feature ranking. The top 10 features with the highest average importance were designated as key features, offering valuable insights into cluster characteristics.

Results

Our analysis encompassed data from 462 individuals diagnosed with Major Depressive Disorder (MDD), with 249 (53.9%) participants hailing from Columbia University and the remaining 213 (46.1%) from Johns Hopkins University. The cohort had an average age of

44.79 ± 10.70 years, and a predominant proportion, 77.7%, were female.

Utilizing hierarchical clustering, we stratified both female and male MDD patients into two distinct clusters (Figure 1). Among female participants, 238 patients (66.3%) were classified into Cluster 1, while 121 patients (33.7%) fell into Cluster 2. For male participants, Cluster 1 comprised 71 patients (69.6%), and Cluster 2 included 31 patients (30.4%).

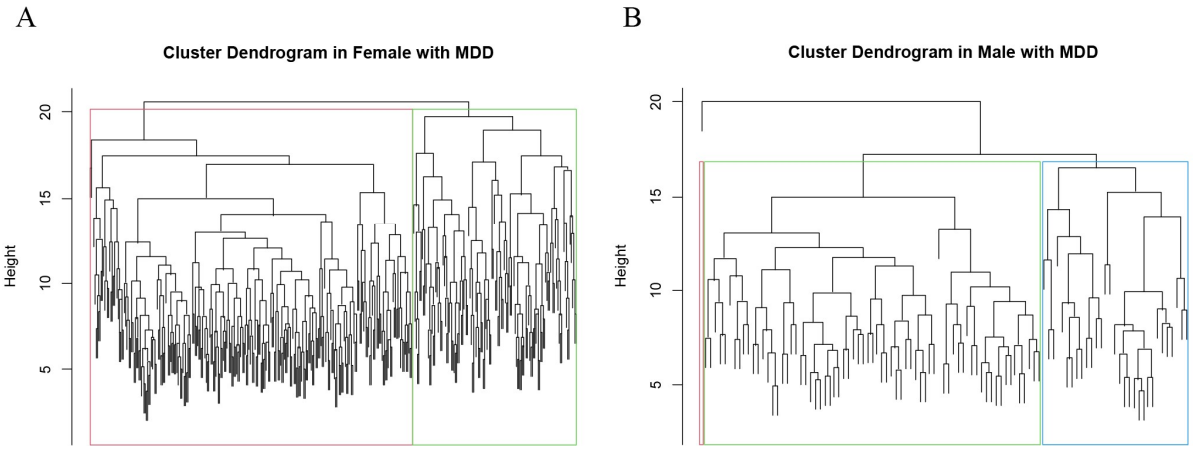

Figure 1

A. Dendrogram shows subclusters in female samples with MDD

B. Dendrogram shows subclusters in male samples with MDD

Feature ranking was informed by the average importance weights derived from permutation importance, highlighting the variates that most significantly delineate the clusters within our MDD patient cohort. The top ten variates, as determined by this ranking, elucidate the differences between the clusters (Tables 1, 2, 3). Additionally, groups of collinear variates were pinpointed through hierarchical clustering, further refining our analysis.

Table 1

Top10 variables displaying the greatest difference among clusters in female

Top 10 reserved variates	Correlated variates removed	average importance weights
GAD_4	NA	0.0593
GAD_3	NA	0.0405
PHQ_7	NA	0.0330
Q4_Currently_Smoke	NA	0.0327
PHQ_4	NA	0.0326
GAD_2	GAD_TOT Gad_Tot	0.0317
Smokebefore	SmokeBefore Q4a_Cigs_Typ_Day_Now	0.0293
AGE_ONSET_MDD	Age_Onset_Mdd Logaao	0.0264
PHQ_3	NA	0.0223
DEATH_WORST	NA	0.0212

Table 2

Top10 variables displaying the greatest difference among clusters in male

Feature name	Highly corelated features	average importance weights
PHQ_7	NA	0.1140
STIM_LEVEL	Stim_Level	0.0623
ALCOHOL_PROBLEM	Alcohol_Abuse Anyalcdx Anyalcdrug	0.0570
Neglect	GRQ6_Q6d_GV_neglect_1.4	0.05349
Ppwork_12work_37not	PPWORK	0.0506
Alc_Dependence	Alcsubdep	0.0352
Bmi_Current	Bmi_Max Qc_weight Qd_max_weight	0.0215
Emot_Abuse	GRQ6_Q6c_GV_emot_abuse_ 1.4 Tot_Abuse7 Abusef1	0.0213
Factor(4)	NA	0.0200
HALLUC_LEVEL	Halluc_Level H1a_LSD_GV	0.0191

Table 3

Top10 variables displaying the greatest difference among clusters both in male and female

Feature name	Highly corelated features	average importance weights
gender	Sex PPGENDER	0.0811
GAD_4	NA	0.0479
Qa_height_feet	NA	0.0363
GAD_3	NA	0.0286
Q4_Currently_Smoke	NA	0.0187
PHQ_7	NA	0.0182
PHQ_4	NA	0.0168
Qb+height_inches	NA	0.0152
Smokebefore	SmokeBefore Q4a_Cigs_Typ_Day_Now	0.0152
GAD_2	GAD_TOT Gad_Tot	0.0143

Cluster analysis for female patients

The Generalized Anxiety Disorder 7-item (GAD-7) scale, a self-administered questionnaire designed for the screening and severity assessment of generalized anxiety disorder (GAD)[13], revealed significant disparities between clusters in responses to three specific questions: Q2 ("Not being able to stop or control worrying"), Q3 ("Worrying too much about different things"), and Q4 ("Trouble relaxing") (Figure 2). Additionally, the overall GAD-7 scores varied markedly between clusters, indicating distinct anxiety profiles within the female MDD cohort.

Similarly, the nine-item Patient Health Questionnaire (PHQ-9), a widely-used scale for evaluating depressive symptoms and serving as a diagnostic tool in primary care settings[14], highlighted significant differences in responses to questions Q3 ("Trouble falling or staying asleep, or sleeping too much"), Q4 ("Feeling tired or having little energy"), and Q7 ("Trouble concentrating on things, such as reading the newspaper or watching television") across

clusters (Figure 2). These findings suggest varied depression symptomatology among the clusters.

Consistent with literature linking smoking to both depression and anxiety, our analysis found a notable contrast in smoking behaviors between clusters[15]. In Cluster 1, a vast majority (91.2%) of MDD patients reported not currently smoking, whereas Cluster 2 had a significantly higher proportion (33.1%) of current smokers. Moreover, the mean age was higher in Cluster 1 compared to Cluster 2. Additionally, the prevalence of general thoughts about death was significantly different between clusters, with 53.8% in Cluster 1 and a higher rate of 81.8% in Cluster 2 experiencing such thoughts (Figure 2).

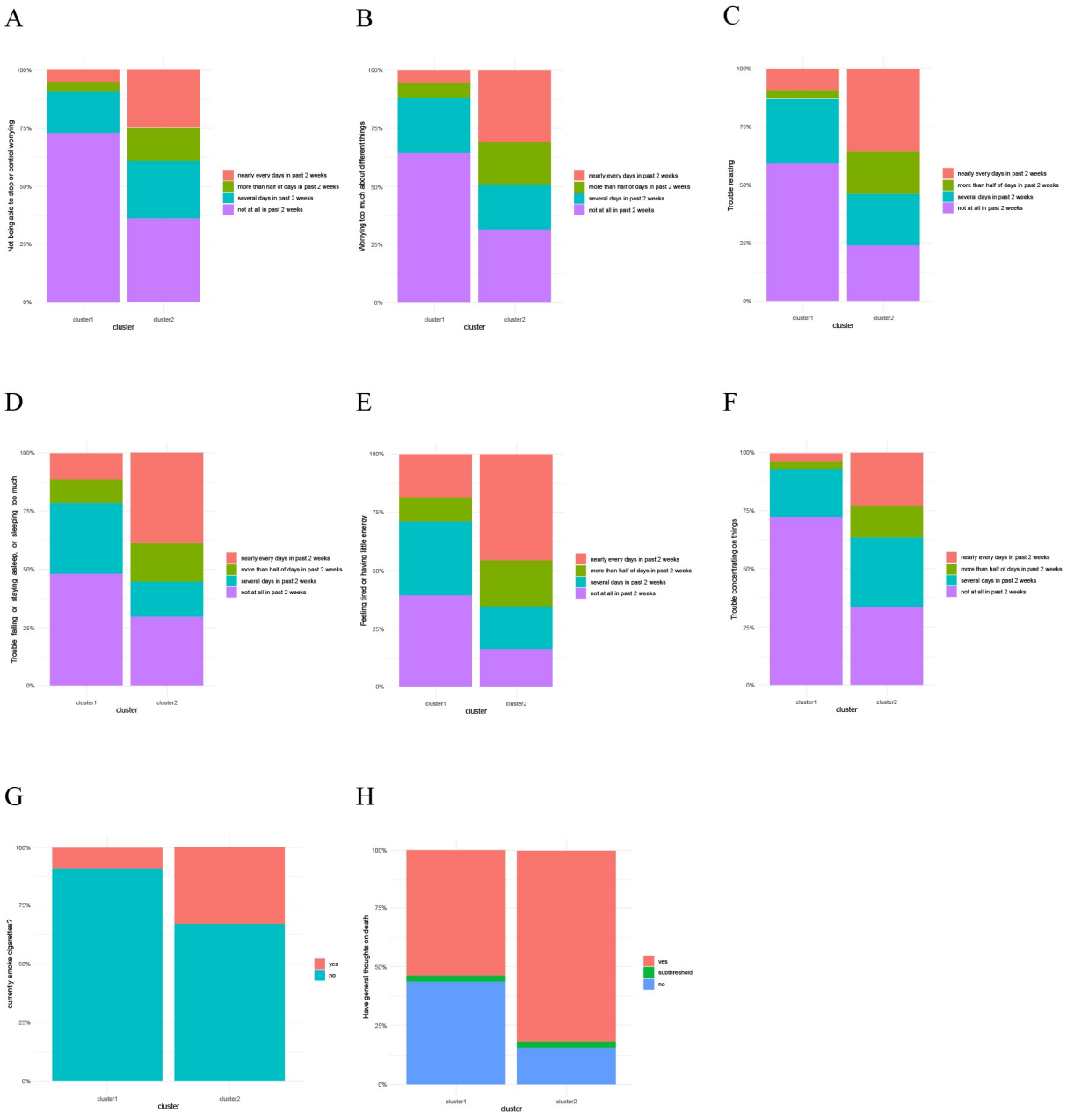

Figure 2

Barplots shows distribution of subclusters of female MDD in key features

A-C shows distribution of subclusters of female MDD at Q2, Q3, Q4 in GAD-7

D-F shows distribution of subclusters of female MDD at Q3, Q4, Q7 in PHQ-9

G-H shows distribution of subclusters of female MDD at smoking and general thoughts on death

Cluster analysis for male patients

Depression frequently co-occurs with addiction to substances, including drugs and alcohol. Substance abuse can exacerbate feelings of loneliness, sadness, and hopelessness, which are commonly associated with depression. Our findings reveal marked differences between clusters in terms of substance abuse, encompassing alcohol, stimulants, and hallucinogens (Figure 3).

Furthermore, there is a well-documented link between childhood trauma and the subsequent development of Major Depressive Disorder (MDD) in adulthood[16]. Our analysis indicated significant variances across clusters concerning experiences of childhood trauma, specifically neglect and emotional abuse (Figure 3).

Consistent with observations in female patients, question Q7 from the PHQ-9 ("Trouble concentrating on things, such as reading the newspaper or watching television") demonstrated notable differences across clusters in male MDD patients. Additionally, employment status varied significantly between clusters, with 88.7% of patients in Cluster 1 being employed (either as paid employees or self-employed), compared to only 61.3% in Cluster 2. Body Mass Index (BMI) also differed, with Cluster 2 showing higher mean BMI values than Cluster 1 (Figure 3).

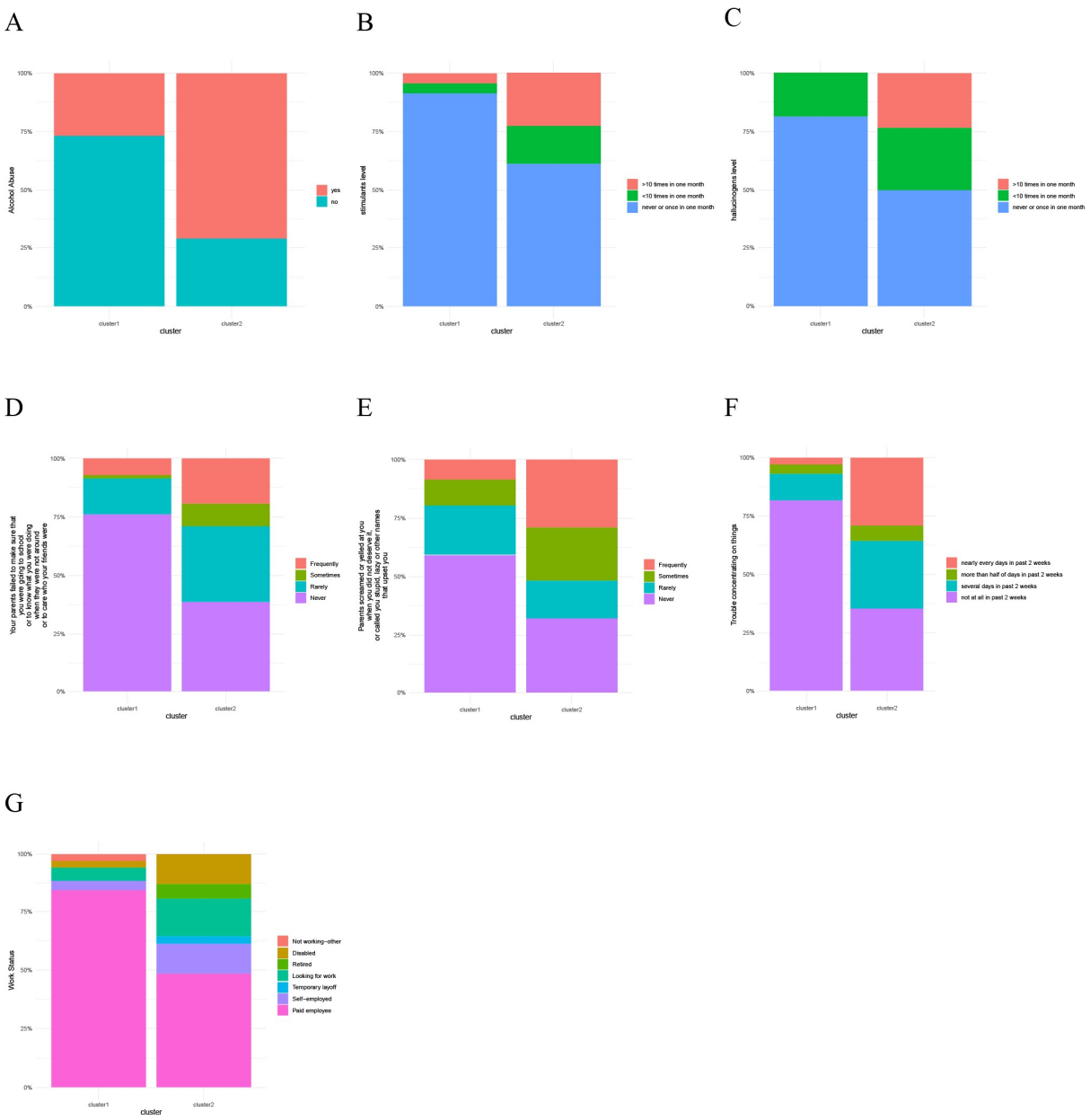

Figure 3

Barplots shows distribution of subclusters of male MDD in key features

A-C shows distribution of subclusters of male MDD in alcohol, stimulants and hallucinogens abuse

D-E shows distribution of subclusters of male MDD in neglect abuse and emotion abuse

F shows distribution of subclusters of male MDD at Q7 in PHQ-9

G shows distribution of subclusters of male MDD in work status

Overall cluster analysis

The integration of cluster data from both female and male MDD patients revealed nuanced differences across the combined four clusters (Figure 4). Notably, Cluster 2 among female patients exhibited more pronounced anxiety symptoms, particularly in responses to questions Q2, Q3, and Q4 of the GAD-7 scale. Furthermore, this cluster also recorded a significantly higher average total GAD-7 score in comparison to other clusters. Specifically, the mean ± standard deviation (SD) for Cluster 2 in females was 8.50 ± 5.50, whereas the corresponding mean scores ± SD for Cluster 2 in females, along with Clusters 1 and 2 in males, were 3.29 ± 3.51, 3.48 ± 4.32, and 4.81 ± 5.37, respectively.

Moreover, both Cluster 2 in females and Cluster 2 in males demonstrated more severe symptoms of depression, as evidenced by their responses to questions Q4 and Q7 of the PHQ-9. Additionally, these clusters showed a higher prevalence of current smoking compared to their counterparts (Figure 5).

Lastly, an analysis of immune cell composition revealed that the mean proportion of CD16+ monocytes was notably higher in Cluster 1 among males than in other clusters. Conversely, the mean proportion of naive B cells was observed to be lower in Cluster 1 among females compared to Cluster 2 in females and Cluster 1 in males.

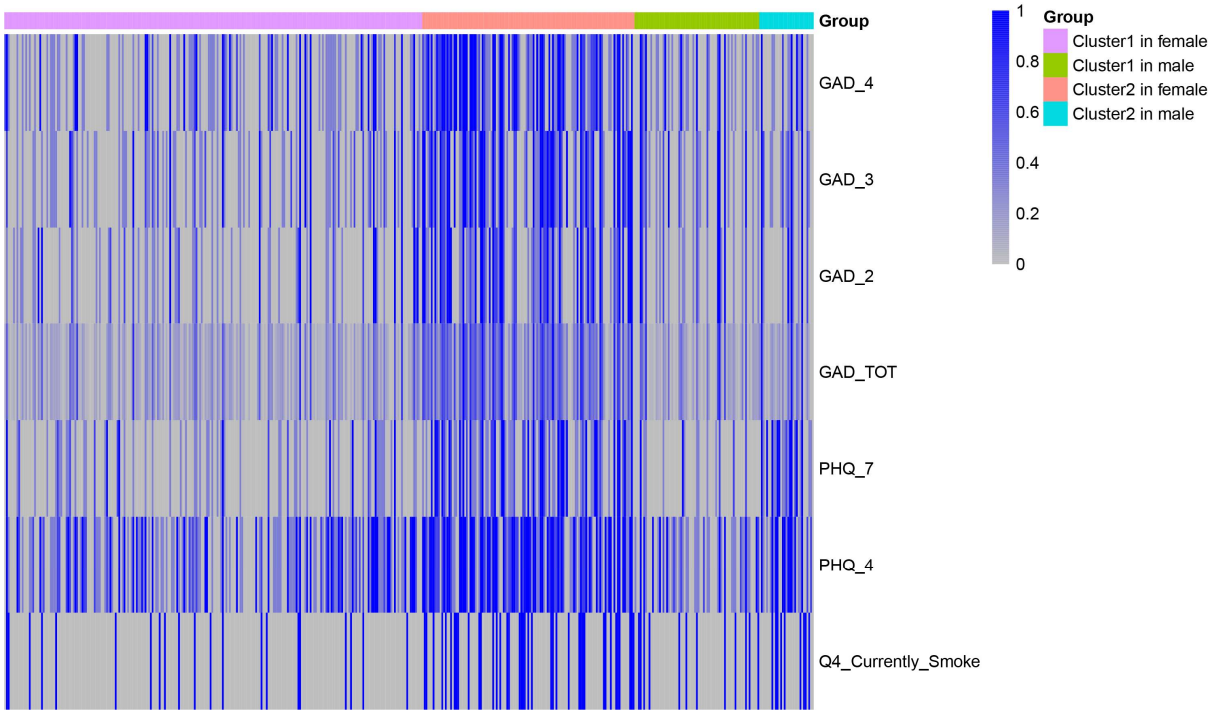

Figure 4

This heatmap displays symptom severity of key features across four MDD subclusters. The horizontal axis represents MDD samples, and the vertical axis denotes names of key features. Color intensity of blue indicates normalized expression values of key features.

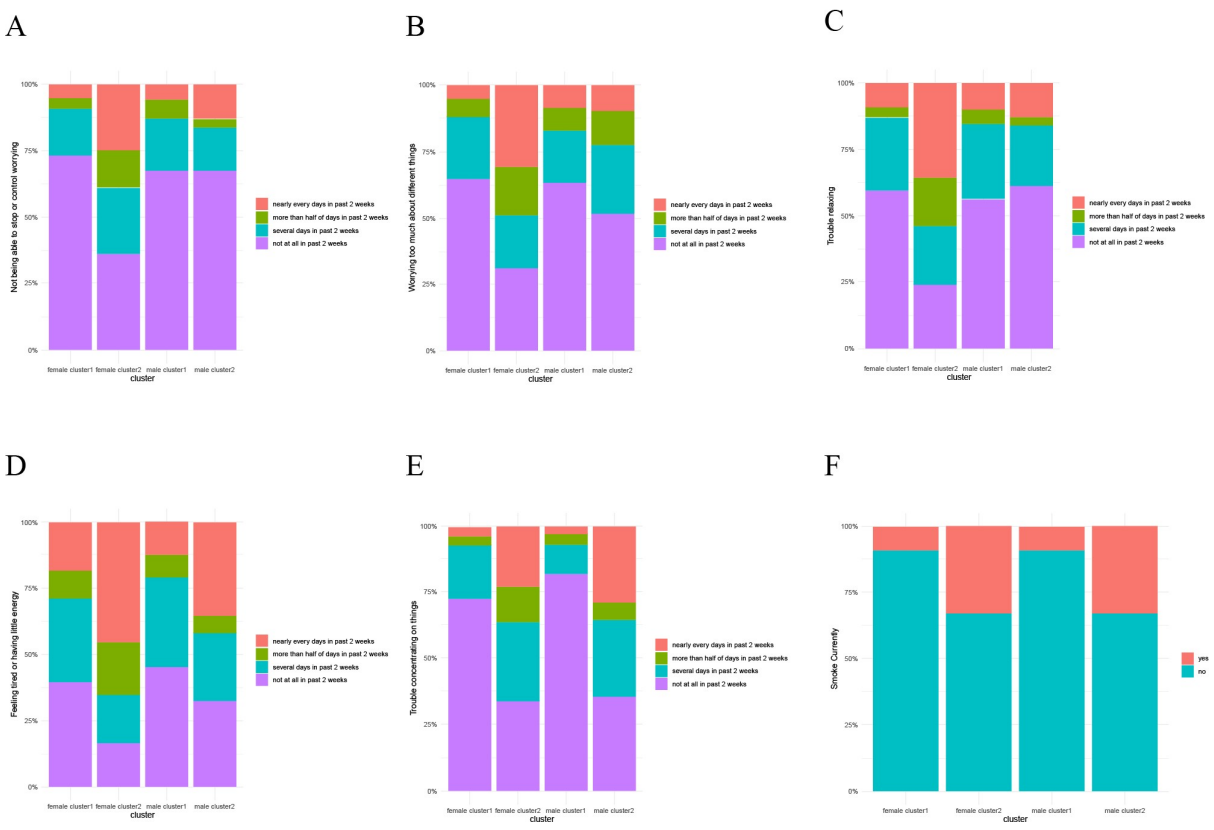

Figure 5

Barplots shows distribution of subclusters of all gender MDD in key features

A-C shows distribution of subclusters MDD at Q2, Q3, Q4 in GAD-7

D-E shows distribution of subclusters MDD at Q4, Q7 in PHQ-9

F shows distribution of subclusters MDD in smoking

Discussion

Our study has elucidated a distinct subtype within Major Depressive Disorder (MDD) that manifests with heightened symptom severity across both male and female cohorts. This particularly severe MDD subtype may share characteristics with Treatment-Resistant Depression (TRD), notably an attenuated response to conventional therapeutic interventions.

The ability to identify such subtypes holds substantial promise for enhancing the clinical management of MDD, enabling more tailored and potentially more effective treatment strategies.

Crucially, the differentiation of MDD subtypes in this study hinged on specific items from established questionnaires, including the Structured Clinical Interview for DSM-IV, the PHQ-9, and the GAD-7. These targeted questions could serve as valuable tools for assessing the severity of MDD in clinical settings, facilitating more nuanced patient evaluations.

However, several areas warrant further exploration and refinement. Notably, our sample exhibited a gender imbalance, with a higher representation of female participants, which, despite aligning with the broader epidemiological trend of MDD's higher prevalence among women, might compromise the generalizability of our findings, especially concerning male patients and overall gender comparisons. Moreover, the lack of data on treatment outcomes in our study limits our ability to conclusively link the identified severe MDD subtype with TRD, underscoring the need for incorporating treatment response information in future research.

Additionally, our investigation did not reveal a significant association between the severe MDD subtype and variations in immune cell fractions. This finding prompts a shift in focus towards molecular investigations, particularly exploring RNA expression differences and alternative splicing patterns, which may offer deeper insights into the biological underpinnings of MDD subtypes.

Reference

[1] *P F Sullivan, K S Neale Mc Fau - Kendler, K S Kendler. Genetic epidemiology of major depression: review and meta-analysis [J]. (0002-953X (Print)).*

[2] *R M Hirschfeld. Major depression, dysthymia and depressive personality disorder [J]. (0960-5371 (Print)).*

[3] *G A Fava. Can long-term treatment with antidepressant drugs worsen the course of depression? [J]. (0160-6689 (Print)).*

[4] *E Sarno, A J Moeser, A J Robison. Neuroimmunology of depression [J]. (1557-8925 (Electronic)).*

[5] *S Mostafavi, A Battle, X Zhu, et al. Type I interferon signaling genes in recurrent major depression: increased expression detected by whole-blood RNA sequencing [J]. (1476-5578 (Electronic)).*

[6] *S Debey-Pascher, F Hofmann a Fau - Kreusch, G Kreusch F Fau - Schuler, et al. RNA-stabilized whole blood samples but not peripheral blood mononuclear cells can be stored for prolonged time periods prior to transcriptome analysis [J]. (1943-7811 (Electronic)).*

[7] *P Ewels, M Magnusson, S Lundin, et al. MultiQC: summarize analysis results for multiple tools and samples in a single report [J]. (1367-4811 (Electronic)).*

[8] *A Dobin, F Davis Ca Fau - Schlesinger, J Schlesinger F Fau - Drenkow, et al. STAR: ultrafast universal RNA-seq aligner [J]. (1367-4811 (Electronic)).*

[9] *A a-O Newman, C B Steen, C L Liu, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry [J]. (1546-1696 (Electronic)).*

[10] *F a-O X Vallania, A Tam, S Lofgren, et al. Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases [J]. (2041-1723 (Electronic)).*

[11] *R Karkare, R Paffenroth, G J a P A Mahindre. Blind image denoising and inpainting using robust hadamard autoencoders [J]. 2021.*

[12] *A Senawi, H-L Wei, S A Billings. A new maximum relevance-minimum multicollinearity (MRmMC) method for feature selection and ranking [J]. Pattern Recognition, 2017, 67: 47-61.*

[13] *R L Spitzer, J B W Kroenke K Fau - Williams, B Williams Jb Fau - Löwe, et al. A brief measure for assessing generalized anxiety disorder: the GAD-7 [J]. (0003-9926 (Print)).*

[14] *K Kroenke, J B Spitzer Rl Fau - Williams, J B Williams. The PHQ-9: validity of a brief depression severity measure [J]. (0884-8734 (Print)).*

[15] *M Fluharty, A E Taylor, M Grabski, et al. The Association of Cigarette Smoking With Depression and Anxiety: A Systematic Review [J]. (1469-994X (Electronic)).*

[16] J Lemoult, K L Humphreys, A Tracy, et al. Meta-analysis: Exposure to Early Life Stress and Risk for Depression in Childhood and Adolescence [J]. (1527-5418 (Electronic)).