PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (http://bmjopen.bmj.com/site/about/resources/checklist.pdf) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

| TITLE (PROVISIONAL) | An Integrated Module of Multidimensional Omics for Peripheral Biomarkers (iMORE) in Major Depressive Disorder Patients: Rationale and Design of a Prospective Multicenter Cohort Study |
| AUTHORs            | Li, Huafang; Zheng, Yuzhen; Zhang, Linna; He, Shen; Xie, Zuoquan; Zhang, Jing; Ge, Changrong; Sun, Guangqiang; Huang, Jingjing |

VERSION 1 – REVIEW

| REVIEWER          | Lutz, Pierre-Eric |
|                  | Institut des neurosciences cellulaires et intégratives, CNRS - INCI UPR3212 |
| REVIEW RETURNED  | 15-Sep-2022 |

GENERAL COMMENTS

This is the protocol for an interesting study designed to investigate biomarkers of antidepressant treatment response in 2 phases: a first observation phase in 150 patients with MDD and 50 healthy controls, and a second validation phase with 50 patients and 50 controls. Along with clinical evaluation (MADRS, HAMD-17), the following biomarkers will be assessed: >400 plasma cytokines (Quantibody Human Cytokine Antibody Array), >600 plasma metabolites (LC-MS), and gut microbiome taxonomy (16S sequencing). Feature selection will be conducted (Deep Feature Selection), followed by dimensionality reduction (autoencoder) and evaluation of accuracy (AUC). Overall, the manuscript is clear and the study well-described. I only have minor comments that the authors may wish to consider.

1) It would probably be helpful to specify in the abstract which multi-dimensional integration and machine learning methods will be applied, and to separate the presentation of strengths and limitations of the study.

2) The following could be clarified or discussed:
   - How many features will be prioritized for classification?
   - Will the exact same biochemical and sequencing methods be used for the 2 phases of the study? Then, in addition to testing in Phase 2 the replication of models constructed in Phase 1, it would also be possible to conduct an additional analysis using pooled subjects from the 2 Phases. Is this planned?
   - If I understand correctly, the treatment response part of the study will use only MDD patients (separated in responders and non-responders), while healthy controls will only be used for the biological subtyping part of the study? These are essentially 2 distinct studies, which could be clarified. Also, Ns could be clarified accordingly: N=150+50 for the treatment response study, N=200+100 for the subtyping study?
An increasing number of studies perform multiomic or multimodal investigations of MDD. I am curious to know how the authors prioritized the methods they will combine: personal experience with each approach (practical justification, which is OK), rationale based on previous studies conducted in other medical fields? Within this line, perhaps it would be interesting if, in the introduction, the authors could provide some background on current knowledge regarding putative relationships and cross-talk mechanisms among cytokines and metabolites in the blood, the gut microbiome, and the brain.

- Bibliographic reference for autoencoder?
- Justification for choosing autoencoder and Deep Feature Selection (primarily developed for genomics data) for dimensionality reduction & feature selection, rather than other methods (scikit, JIVE, RGCCA, etc)?

<table>
<thead>
<tr>
<th>REVIEWER</th>
<th>Issler, Orna</th>
</tr>
</thead>
<tbody>
<tr>
<td>Icahn School of Medicine at Mount Sinai</td>
<td></td>
</tr>
<tr>
<td>REVIEW RETURNED</td>
<td>19-Sep-2022</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GENERAL COMMENTS</th>
</tr>
</thead>
</table>
| This proposal sets an important goal: identifying biomarkers for depression diagnosis and predicting treatment response. It acknowledges the need to combine multiple data points to identify features with a predictive value that is also scalable for actual clinical use. It proposes to collect rich data sets, including clinical data along with blood immune markers and microbiome data from the stool of depressed and control subjects before and after SSRi/SNRI treatment. The authors will then perform machine learning data analysis to identify features with diagnostic value and test the results in a replication cohort. Overall, this is a good proposal clearly written. Several topics require discussion and clarification:

1. A large subset of the subjects in this study was recruited during the COVID pandemic. The pandemic has been a major stressor that has increased depression rates both in people with a history of depression and naïve patients. What efforts are being made to account for this specific cohort's potential confounder of the pandemic? Please discuss.

2. A detailed power analysis should be included in the manuscript. What is the predicted effect size for the components of the proposed multi-omics analysis? To avoid overfitting machine learning analysis, a large data set is required. Is the proposed sample size sufficient?

3. Please include more information regarding the blood and stool collection that can affect the results. For example, time of day, were subjects fasting, and so on.

4. It will be beneficial to include pilot data and/or cite other papers by the authors to show the general feasibility and the team's specific expertise in the omics and informatics approaches proposed.

5. GWAS studies linking specific SNPs with depression risk in the Han Chinese population. Do the authors plan to integrate genetic data in their analysis? Will the results of this study be tested in other ethnic groups?

6. The proposed criteria for response and remission sound very dichotomic; I would advise including also % change by antidepressant treatment and not just categorizing as responders vs nonresponses. Moreover, please ensure to include as many clinical features as possible in machine learning.
7. Typically, the term proteomics is used to describe unbiased proteome-wide mass spectrometry analysis, while in this study, it will be used to study a large set of targeted plasma cytokines. I would change the name of this analysis to more accurately describe the data that will be collected. Moreover, I would add more methods and details regarding the immunophenotyping analysis: What type of cells will be profiled and more technical information regarding the FACS.

8. Please clarify why one of the exclusion criteria is “Fertile not using contraception”. There is no justification not to include women that are not taking hormones to prevent pregnancy in such a study.

**VERSION 1 – AUTHOR RESPONSE**

Reviewer #1

Comments:
1) It would probably be helpful to specify in the abstract which multi-dimensional integration and machine learning methods will be applied, and to separate the presentation of strengths and limitations of the study.

Responds to the reviewers' comments:
Sincerely thank you for your suggestions. We have added details in the abstract about the multi-omics integration and machine learning methods applied (Page3, line 2-4).
To clearly demonstrate the strengths and limitations of the study, we have revised it into an independent paragraph in the discussion section (Page17, line 19-22).

2) Comments: How many features will be prioritized for classification?

Responds to the reviewers' comments:
Sincerely thank you for your significant comment. Features for constructing models are selected from three main types of data, sociodemographic factors, clinical characteristics, and omics data. According to literature (Squarcina L, et al. Deep learning for the prediction of treatment response in depression. J Affect Disord. 2021 doi: 10.1016/j.jad.2020.11.104), about 20 to more than 100 features would be selected from about 200 variables. Since our data contain more dimensions, we may have more features for classification.

3) Comments: Will the exact same biochemical and sequencing methods be used for the 2 phases of the study? Then, in addition to testing in Phase 2 the replication of models constructed in Phase 1, it would also be possible to conduct an additional analysis using pooled subjects from the 2 Phases. Is this planned?

Responds to the reviewers' comments:
Sincerely thank you for your significant comment. Our research team has established a fixed collaboration with the central laboratory on the biochemical measurement, so in the absence of unexpected circumstances, assays used in the two phases of the study remain the same. Models built in Phase 1 will be validated in Phase 2. At the same time, the multi-stage study design will allow us to continue to optimize the results of the previous stage. For example, the prognostic data for all stages could be cross-validated with more complexity to obtain more stable results.
4) Comments: If I understand correctly, the treatment response part of the study will use only MDD patients (separated in responders and non-responders), while healthy controls will only be used for the biological subtyping part of the study? These are essentially 2 distinct studies, which could be clarified. Also, Ns could be clarified accordingly: N=150+50 for the treatment response study, N=200+100 for the subtyping study?

Responds to the reviewers' comments:
Sincerely thank you for your significant comment. Your understanding is correct, and we would like to add some information. Data from MDD subjects are used in the study for efficacy response, which is the primary purpose. Meanwhile, the data are used to analyze the biological phenotypes of depression, such as the biological characteristics of patients with significant sleep disturbance or severe anxiety, and their differences in prognosis. To explore the diagnostic markers of MDD, data are used from the healthy controls. In order to make the most of the sample, we might be inclined to consider this as different content within one study. However, we realized some of our expressions in the manuscript could be misleading. In order to describe this clearly, we have made the following modifications (Page 16, line 10-16):
Diagnostic biomarkers for MDD are explored using data from 100 healthy controls including initial assessment in Stage 2. Biological phenotyping of MDD is based on the variables of 150 MDD subjects in all stages by integration analysis, such as biological signatures with severe sleep disturbances or anxiety symptoms.

5) Comments: An increasing number of studies perform multiomic or multimodal investigations of MDD. I am curious to know how the authors prioritized the methods they will combine: personal experience with each approach (practical justification, which is OK), rationale based on previous studies conducted in other medical fields? Within this line, perhaps it would be interesting if, in the introduction, the authors could provide some background on current knowledge regarding putative relationships and cross-talk mechanisms among cytokines and metabolites in the blood, the gut microbiome, and the brain.

Responds to the reviewers' comments:
Sincerely thank you for your important suggestions. As researchers, we hope to include as many dimensions as possible, which is also the current research trend. Because this is a cohort study in the real world, we selected metabolome, cytokines, gut microbiome, and clinical features that are more likely to reflect environmental influences, given clinical application and sample size. As described in previous studies (Rush AJ, et al. Selecting Among Second-Step Antidepressant Medication Monotherapies: Predictive Value of Clinical, Demographic, or First-Step Treatment Features. Archives of General Psychiatry, 2008.), it is common to screen clinical characteristics based on clinical experience or literature research. We are less sure about the preferred method for other omics, as data vary widely from different studies. However, we will pay close attention to neuroimmunology and gut microbiota-brain axis markers in the dataset, which should guide our analysis direction. Some current knowledge about the gut microbiota-brain axis and neuroimmunology have been added as background (Page 6, line 4-13):
6) Comments: Bibliographic reference for autoencoder? Justification for choosing autoencoder and Deep Feature Selection (primarily developed for genomics data) for dimensionality reduction & feature selection, rather than other methods (scikit, JIVE, RGCCA, etc)?

Responds to the reviewers' comments:
Sincerely thank you for your important suggestions. We have revised and added some information about autoencoders, including references in the manuscript (Page 15, line 15-18).

Autoencoder is a nonlinear factorization technique with multi-layer neural network structures to learn data representation by reducing dimensionality, widely used in deep learning. The latent variables from autoencoder can take advantage of learning that can capture nonlinear features. Although traditional dimensionality reduction (e.g., PCA, canonical correlation analysis) techniques can also be used, such algorithms are linear-based, which may not fully take advantage of deep learning (Kang M, et al. A roadmap for multi-omics data integration using deep learning. Brief Bioinform. 2022;23).

We admit that we do not yet know the best fit for each dimension of the data and the algorithms you mentioned are very valuable for reference!

Taking the pipeline method widely used in the Scikit package as an example, its algorithm includes a linear model and RandomForestClassifier, which we think can be classified as a traditional feature selection. Regularized Generalized Canonical Correlation Analysis (RGCCA) is a framework for modeling linear relationships between several blocks of variables observed on the same set of individuals. Considering a network of connections, the objective of RGCCA is to find linear combinations of block variables. If we understand correctly, Joint and Individual Variation Explained (JIVE) represents an extension of Principal Component Analysis and has advantages over popular two-block methods such as Canonical Correlation Analysis and Partial Least Squares (Lock EF, et al. JOINT AND INDIVIDUAL VARIATION EXPLAINED (JIVE) FOR INTEGRATED ANALYSIS OF MULTIPLE DATA TYPES. Ann Appl Stat. 2013 doi: 10.1214/12-AOAS597). We think this algorithm is also worth trying.

Therefore, considering the complexity of data, we may consider deep learning methods more. Some studies also supported the superiority of autoencoder over other algorithms (Li Zhang, et al. Deep Learning-Based Multi-Omics Data Integration Reveals Two Prognostic Subtypes in High-Risk NeuroblastomaFront. Genet. 2018 doi: 10.3389/fgene.2018.00477.).

However, we also realized that the Deep Feature Selection might not be a good example for reference, just as you have suggested, and we have revised the description in the section. We will use more than one algorithm or model in our research. Comparison of results from multiple algorithmic models will increase the interpretability of our findings.

Reviewer# 2
1) Comments: A large subset of the subjects in this study was recruited during the COVID pandemic. The pandemic has been a major stressor that has increased depression rates both in people with a history of depression and naïve patients. What efforts are being made to account for this specific cohort's potential confounder of the pandemic? Please discuss.

Responds to the reviewers' comments:
Sincerely thank you for the important suggestions. We fully understand your concerns. As far as we know, the impact of the new crown epidemic varies greatly in different countries and regions. The city of Shanghai, China, where this study was conducted, and its neighboring areas were less affected by the COVID-19 epidemic before March 2022. The study was conducted normally and the recruitment of most patients was completed before then. In the next 3 months, due to the epidemic, parts of the recruitment and follow-up were affected. What is interesting is that the impact on patients is also quite different. We observed that not all patients deteriorated due to the epidemic, and some patients were emotionally stable due to resting at home.
Therefore, we will perform sensitivity analysis and mediation analysis on the data of patients enrolled during the epidemic for the primary outcome. If the results suggest that there is a large risk from the confounder, we will conduct a mediation analysis, specifically calculating the OR value of the impact of the epidemic on patient outcomes, and make corresponding adjustments.

2) Comments: A detailed power analysis should be included in the manuscript. What is the predicted effect size for the components of the proposed multi-omics analysis? To avoid overfitting machine learning analysis, a large data set is required. Is the proposed sample size sufficient?

Responds to the reviewers' comments:
Sincerely thank you for your significant comment. Since multi-omics analysis uses high-dimensional data, it may be difficult to accurately calculate the sample size required to build a model. And we have added some power analysis for sample size (Page 13, line 4-10):

According to the AUC range of previous machine learning prediction models (Squarcina L, et al. Deep learning for the prediction of treatment response in depression. J Affect Disord. 2021;281:618-22.), we estimate the AUC of the model developed to be at least 0.70. The sample size included in Stage 2 can achieve more than 80% power to detect a difference of 0.10 between the AUC value (assuming a 50% response rate of MDD patients in the cohort), using a two-sided z-test at a significance level of 0.05. For the main parameter of interest, an effect size criteria for models is applied that odds ratios should exceed 1.2.

As we discussed in the protocol, high dimensionality and a relatively small sample size are issues to tackle in this study because more types of omics are involved, similar to other omics studies. Several common approaches would help to avoid overfitting problems, such as leave-one-out, and training can be terminated early when overfitting occurs.

3) Comments: Please include more information regarding the blood and stool collection that can affect the results. For example- time of day, were subjects fasting, and so on.

Responds to the reviewers' comments:
Sincerely thank you for your significant comment. We have added more detailed information about the standard procedure of samples collection (Page 10, line 15-19).

4) Comments: It will be beneficial to include pilot data and/or cite other papers by the authors to show the general feasibility and the team's specific expertise in the omics and informatics approaches proposed.

Responds to the reviewers' comments:
Sincerely thank you for the important suggestions. We have published results on plasma metabolomic and transcriptomics studies of mental disorders (schizophrenia and MDD), which we have agreed to cite appropriately in the context of background (Page 6, line 4-7):

5) Comments: GWAS studies linking specific SNPs with depression risk in the Han Chinese population. Do the authors plan to integrate genetic data in their analysis? Will the results of this study be tested in other ethnic groups?

Responds to the reviewers' comments:
Sincerely thank you for the important suggestions. The famous study CONVERGE (Cai N, et al. 11,670 whole-genome sequences representative of the Han Chinese population from the CONVERGE project. Sci Data. 2017.) showed evidence of association of two SNPs (SIRT1 and LHPP) on Han Chinese population, and the trans-ancestry genetic correlation between the PGC (Psychiatric Genomics Consortium) and CONVERGE GWAS indicating there are likely population differences in MDD genetic etiology (Bigdeli TB, et al. Genetic effects influencing risk for major depressive disorder in China and Europe. Transl Psychiatry. 2017. doi: 10.1038/tp.2016.292.). These studies were based on extremely large sample sizes. Given the modest heritability of MDD, compared with other psychiatric disorders, risk alleles are likely to have small effect sizes (Sullivan PF, et al. Genetic epidemiology of major depression: review and meta-analysis. Am J Psychiatry. 2000.). Limited by the actual sample size, we estimate that it may be difficult to capture stable gene-level differences.

In addition, the main purpose of this study is to find biomarkers for predicting efficacy. Considering that the current genomics research may not be enough to distinguish the prognosis of depression, we tend to focus more on biological markers that may represent potential differences in phenotypes, such as metabolome, gut microbiome, etc. However, all biological samples will be well preserved, and genomics research could be conducted under specific plans in the future. If available, we will be glad to cooperate with relevant foreign researchers to apply our research results to other ethnic groups to increase representation.

6) Comments: The proposed criteria for response and remission sound very dichotomic; I would advise including also % change by antidepressant treatment and not just categorizing as responders vs nonresponses. Moreover, please ensure to include as many clinical features as possible in machine learning.

Responds to the reviewers' comments:
Sincerely thank you for the important suggestions. Response and remission rates are widely used as primary outcome measures in drug efficacy studies, including prognostic modeling studies. However, we also realized response rate is not enough to reflect the rich heterogeneity, so clinical phenotypes are also an important content of this study. Your suggestion is of great value, that is, the biological characteristics of differentiated therapeutic effects, which we will try to explore in specific data analysis.

As a real-world-based omics study, we are sure to fully collect the clinical characteristics of patients, such as disease course, family history, drug exposure, living habits, etc. Abundant clinical experience in our team helps us include as much meaningful clinical information as possible to satisfy some interpretability of the results.

7) Comments: Typically, the term proteomics is used to describe unbiased proteome-wide mass spectrometry analysis, while in this study, it will be used to study a large set of targeted plasma cytokines. I would change the name of this analysis to more accurately describe the data that will be collected. Moreover, I would add more methods and details regarding the immunophenotyping analysis: What type of cells will be profiled and more technical information regarding the FACS.

Responds to the reviewers' comments:
Sincerely thank you for the important suggestions. We agree that the term cytokines describe our research more accurately. We have made changes to the corresponding presentation in our
manuscript. In addition, details regarding the immunophenotyping analysis have been added to the manuscript (Page 11, line 5-12):

Human blood samples collected in EDTA tubes are centrifuged at 500 g for 10 minutes. Blood cells are subjected to red blood cell lysis at room temperature, then washed with staining buffer, filtered through a 70 μm strainer and counted. Metal-labeled antibodies are used for staining according to the manufacturer's instructions (Fluidigm Science, San Francisco, USA). Cells are fixed with 1.6% paraformaldehyde and are processed in Ir-Interchelator (Fluidigm) and then incubated at 2-8°C. Before the acquisition, cells are resuspended with Cell Acquisition Solution (Fluidigm) containing diluted EQ Four Element Calibration beads (Fluidigm) and filtered through a 35 μm nylon mesh filter cap. Finally, cells are obtained on a Helios Mass Cytometer (Fluidigm) and are exported and analyzed using Cytobank analysis software.

8) Comments: Please clarify why one of the exclusion criteria is “Fertile not using contraception”. There is no justification not to include women that are not taking hormones to prevent pregnancy in such a study.

Responds to the reviewers' comments:
Sincerely thank you for the important suggestions. We aim to exclude women who currently plan to conceive during this study period, to reduce the dropout rate and other complex factors. We are sorry that our presentation lacked clarity. We have changed the description in the corresponding section (Page11, Table 2).

**VERSION 2 – REVIEW**

<table>
<thead>
<tr>
<th>REVIEWER</th>
<th>Lutz, Pierre-Eric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Institut des neurosciences cellulaires et intégratives, CNRS - INCI UPR3212</td>
</tr>
<tr>
<td>REVIEW RETURNED</td>
<td>28-Oct-2022</td>
</tr>
</tbody>
</table>

| GENERAL COMMENTS    | The authors have addressed my comments. I thank them and recommend publication. |

<table>
<thead>
<tr>
<th>REVIEWER</th>
<th>Issler, Orna</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Icahn School of Medicine at Mount Sinai</td>
</tr>
<tr>
<td>REVIEW RETURNED</td>
<td>03-Nov-2022</td>
</tr>
</tbody>
</table>

| GENERAL COMMENTS    | The authors have addressed this reviewer's concerns in the revised version of the manuscript. Good luck with your study! |