CS-C4100 Digital Health and Human Behavior Final Project: Depression Analysis

December 9, 2023

1 Introduction

Mental illness and depression has and always been a serious problem in society. There have been numerous old and new researches addressing the issue and potential cure for it [1–6]. Although the actual cure for the illness is very important, prediction and detection of future patients is also an integral part of the treatment. For that reason, the aim of this project is to analyze the pattern of patients suffer from depression, as well as their background and the types of depression they suffered. I will then try to build a prediction model that can hopefully predict the depression state of the patients.

The detailed of the report is as follow. Section 2 will give a concrete definition and the goal of each analyses. Section 3 will give a detailed description of the data used in the analysis. Section 4 will present the methods to approach the analysis. Section 5 and 6 will present the results of the analysis, as well as their implication. Finally, section 7 will be the conclusion of the project.

2 Problem Formulation

The goals of this project is to analyse the behavior and characteristics of patients with and without depression, and to predict future depression state. To do that effectively, I would need to outlines the specific problems I want to solve. Those problems are:

- Individual analysis of a patient with depression. This is to see how their daily or hourly activities of them look like. Then I will do the same for a patient without depression and compared the results with each other

- Perform group analysis on the patients with depression. I want to see if there are any correlation or common factors that are shared between depressed patients. Unfortunately, due to the lack of data (more details in Section 3), I cannot do the same with the controlled group

- Finally, I want to build a simple regression model to predict the Montgomery-Asberg Depression Rating Scale (MADRS) score at a later time using currently available data. The MADRS score is a metric to evaluate how severely depressed a patient is. The higher the score, the more severe. More information can be found in [7]

In the following sections, the patients with depression will be referred to as *conditioned* patients or *condition group* and patients without depression as *controlled* patients or *control group*.

3 Data

The full dataset consists of 2 folders and an independent file. The 2 folders are data of the condition and control group. In each of the folders, there are multiple datasets corresponding to multiple patients. Each individual dataset consists of the timestamp, the date, and the activity measured by an actigraph. These datasets would be used in the individual analysis.

The independent file is a dataset contains the background information and some mental illness that the patient has. The detailed data are:

- The patient identifier in the form of patient number

- The number of days the patient has taken measurements

- The gender of the patients, encoded as binary value: 1 for female and 2 for male

- Bipolar types encoded as numerical categories: 1 for bipolar type II, 2 for unipolar depressive, and 3 for bipolar type I

- Melancholia type of patient: 1 for having melancholia and 2 for no melancholia
- Whether or not the patient is an inpatient (i.e., patients who live in the hospital for treatments)
- Education of the patient, grouped by number of years the patient has received education
- Whether or not the patient is married, coinhabiting, or single
- Whether or not the patient is studying/employed or not
- The MADRS score at the start of the measurement and the MADRS score at the end of the measurements

One particular problem with this MADRS score dataset is that for the controlled patients, all values aside from the first three are NaN (Not a number). For the depression indicators, it is reasonable to infer that their values are all 0. However, the same cannot be said for information like Working, Marriage, or Education since everyone would have different values for these. This effectively making the data for the controlled patients unusable and need to be dropped.

4 Methods

4.1 Data cleaning

The data requires some cleaning before being able to used. For the individual datasets, they are quite cleaned already. A quick check reveals that there are no NaN values. Thus, the only preprocessing procedures were to convert the date values and timestamp into the correct `datetime` format. I also concatenate all of the patients data into a single file for easier processing at later steps.

On the other hand, the MADRS dataset needs more attentions. The first thing to do is to drop all the controlled patients data due to reasons stated in Section 3. Among the conditioned patients, there are three patients who does not have data on the melancholy type and some does not have data on the education. I dropped the patients without melancholy data since this

is important data (it is used in building the prediction model). But for the patients without education data, they can be ignored since education was not as important compared to the melancholy data. After dropping the rows, the data were ready to be used.

4.2 Individual analysis

The goal of the individual analysis is to observe the activity pattern of a typical depressed and non-depressed patient. To do this, I first randomly select a patient from each pool. This ensure that there are no bias in patient selection. After selecting the patients, I performed a preliminary analysis by plotting the distribution of the activity for both patient, the results for this can be seen in Fig. 1.

The next experiment is performed by grouping the activities data by weekday and averaging the activity measure. The average activity of both the patients are plotted side-by-side. The final experiment involved dividing the days into four right-inclusive ranges: Morning from 5a.m to 12p.m, Afternoon from 12p.m to 5p.m, Evening from 5p.m to 9p.m, and Night from 9p.m to 4a.m. The data of the patients are averaged by this time sections and the results is plotted side-by-side again. The idea of plotting both into the same plot is to see how the patients went with their day and week as well as to observe whether there are any discrepancy between the exercise pattern of the controlled and conditioned patient.

4.3 Group analysis

In this analysis, only the data of conditioned patients are available, so the main idea is not to compared between two types of patients, but to find out what characteristic do these patients shared. Thus, this section will mostly consists of distribution plotting. The analysis consists of plotting the distribution of gender, age, education, work, and marriage. I then do the same for the depression indicators, including bipolar, melancholy, inpatient, initial MADRS and final MADRS score. This way, we can see what are the dominant features of these patients.

4.4 Build prediciton model

The final experiment to be performed on the dataset is to build a prediction model, with the prediction target being the MADRS scores at the end of the data collection. This model could be useful in detecting possible depression patients or predict whether the condition of the patient becomes more severe. The features used to predict are age, bipolar type, melancholia type, inpatient condition, marrage, work, and the initial MADRS. Education was not used because some of the patients lack that information. One information to be noted is that aside from the initial MADRS, all of the features are categorical and cannot be used in the model. To extract a more informative features, one-hot encoding is used [8].

After the extraction, the feature matrix was of shape 20×21. Since the number of features is too big compared to the number of samples, the resulting model would be very susceptible to overfitting, i.e., it cannot generalized to unseen data [9]. To avoid this, Principle Component Analysis (PCA) is used [10]. I selected 3 principles components as it is a sufficiently small numbers for 20 samples and proceed to the model building.

The model of choice is Linear Regression model due to its simplicity. To get a more accurate evaluation model, I also incorporated Leave-One-Out cross validation (LOO). LOO is an extreme case of $k - fold$ cross validation where k is the number of samples. Since the number of samples is very small, it makes sense to preserve as much training data as possible. The final evaluation metric is then the average mean squared error of the folds.

5 Results

5.1 Individual analysis

Figure 1 shows the preliminary results of the individual analysis. We can see that for both of the patients, they spent most of the time of the experiment not active or have light activities that amounted to scores under 1000. This is not very surprising as apparently, moderate and vigorous activity only amounted for 39 minutes in a day [11]. The plot alone does not tell us much information, so we can move on to the results of the next experiments.

Figure 2 illustrates the results of the weekday analysis. Looking at the conditioned patient, they are the most active towards the end of the week. During these time, the activity points are around 250-300. On the other

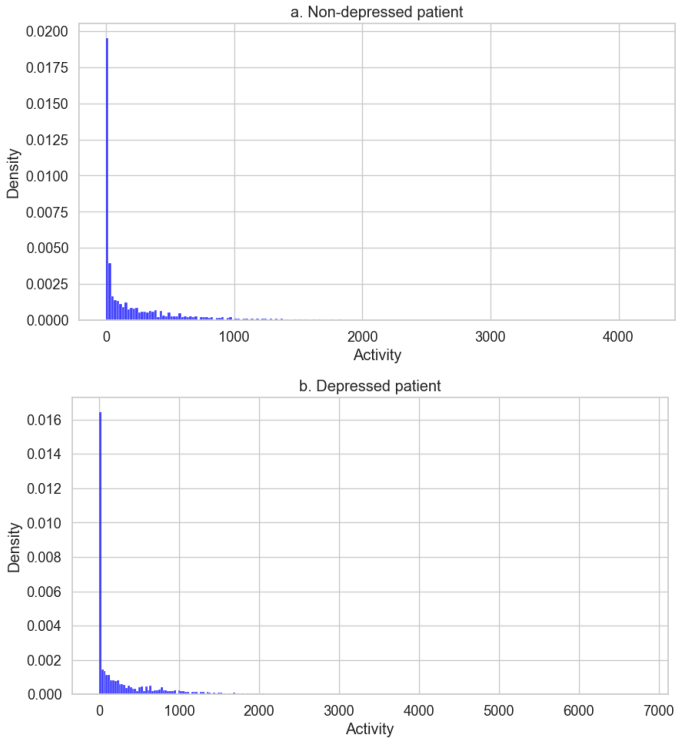

Figure 1: The activity distributions of chosen patients. a. A non-depressed patient. b. A depressed patient

hand, in the middle of the week (i.e., from Tuesday to Thursday) the patient is much less active, only ranging at around 200-250.

For the controlled patient, they are usually 20-50 points higher than the conditioned patient. The controlled patient also shares a somewhat similar trend with the conditioned patient where they are more active in the weekend than during the mid week. However, the distribution is more uniformed compared to the conditioned patient, remains at around 300 points. The only noticeable difference is on Saturday, where the patient scored a point of over 350. Curiously, on Friday and Sunday, the conditioned patient has more activity in general than the controlled patient.

Moving to the final individual analysis. Figure 3 shows the results of the activity scores during the days of the patients. The conditioned patient is very active during the afternoon, scoring an average score of over 500. During other time of the day, they are less active, especially during the morning and

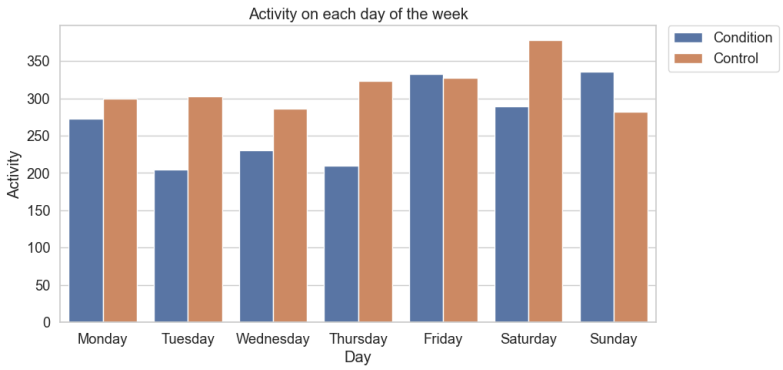

Figure 2: The activity distributions of the patients, grouped by weekday

night. The average scores of evening, morning, and night are 350, 200, and 110, respectively. For the controlled patient, they also very active in the afternoon with an activity score of 500. Morning and evening is also the time where this patient have some activities, with the average scores of 350 and over 300, respectively. Finally, night is the time where the patient is the least active, scoring over 110, similar to the controlled patient.

An interesting comparison can be drawn between these two patients from the two previous plots. In Figure 2, it is rather self-evidence that the controlled patient exercises more than the conditioned patient. However, looking at Figure 3, we can see that aside from morning and night, it is the conditioned patient that exercise more than the controlled patient. In both afternoon and evening, the conditioned patient is about 50 points higher than the controlled patient. But during the morning, the controlled patient is more than doubly active than the conditioned patient, and in the night the controlled patient is about 5-10 points more active.

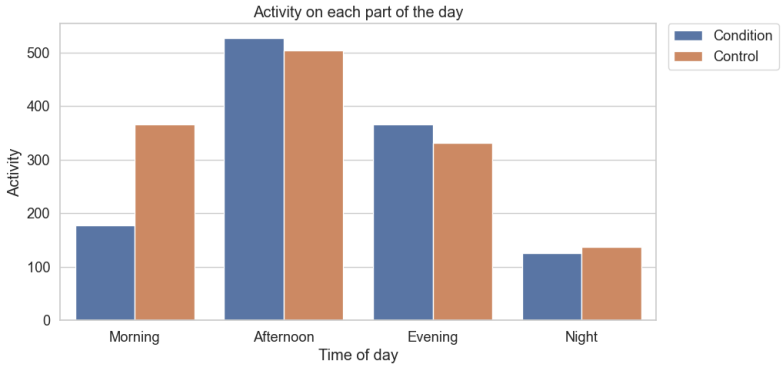

Figure 3: The activity distributions of the patients, grouped by time of the day

5.2 Group analysis

Figure 4 shows the distributions of the demographic information for the depressed patients. For gender, we can see that there is a rather equal numbers of males and females patients, most of the patients are 35-39 or 45-49. For education, 11 of the patients only study up to their highschool (indicate by 6-10 years of education). 6 patients have 11-15 years, which ranging from highschool to 3 years of college. Only 2 patients have 16-20 education years, i.e., likely to finish college. In terms of employment status, only 2 of the patients are employed or studying at the time of data collection, the rest of them are unemployed. Also among the patients, 9 of them are married or coinhabiting, and 11 are single.

Figure 5 shows the distributions of depression-related information of the patients. We can see that a lot of indicators are very imbalanced. Starting with Bipolar indicator, 14 of the patients are unipolar, 5 of the patients are bipolar type II, and the rest are type I. Among the patients, 18 of them did not exhibit melancholy, and only 2 of them have melancholy. Also, 15 of the patients do not stay at the hospital, and 5 of the patients have to stay at the hospital. For the MADRS scores, the initial score have a multimodal distribution, with the two peaks at around 17 and 25, whereas the final MADRS scores have a shape more resembling of a bell-curve.

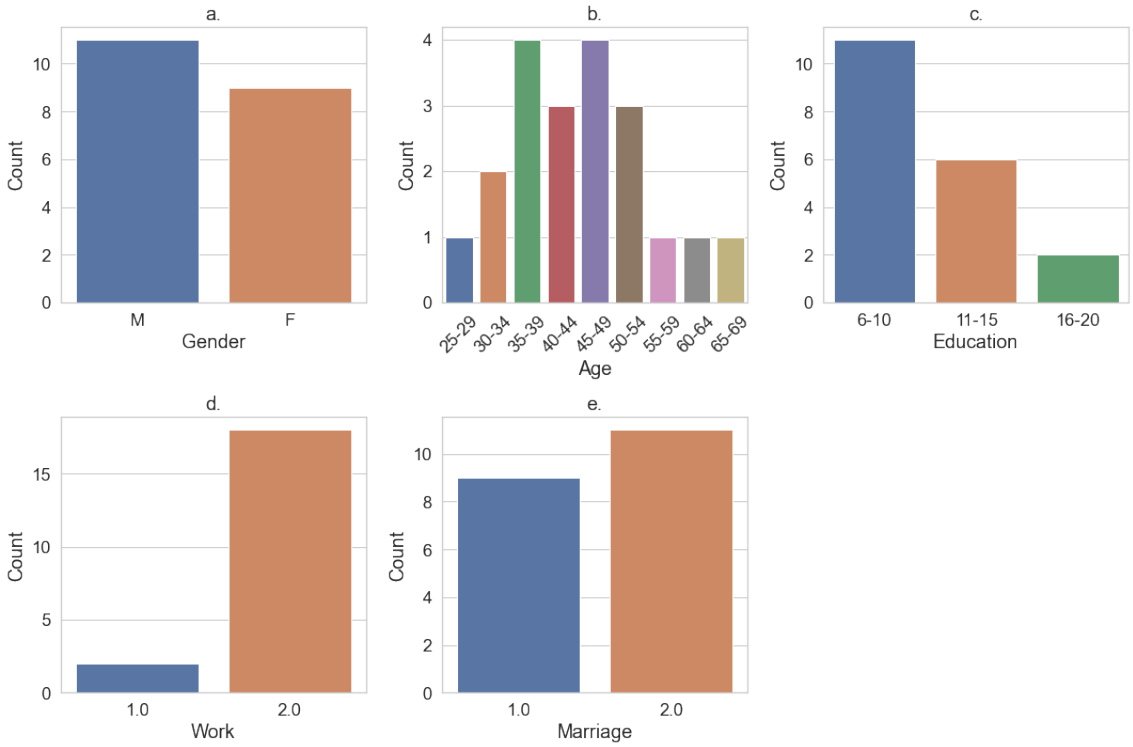

Figure 4: The distribution of demographic information among the patients. a. Gender, b. Age in group, c. Education, group by years of education, d. Work, 1 is employed or currently studying, 2 is unemployed, e. Marriage, 1 is married or coinhabiting, 2 is single

5.3 Prediction model

After running the model with leave-one-out cross validation, the average training error of the model is 8.079, with standard deviation 0.606. During testing, the average testing error is 12.423 with standard deviation of 14.196.

6 Discussion

6.1 Individual analysis

From the individual analysis results, we can see that the controlled patient is more active than the conditioned patient, particularly during the morning

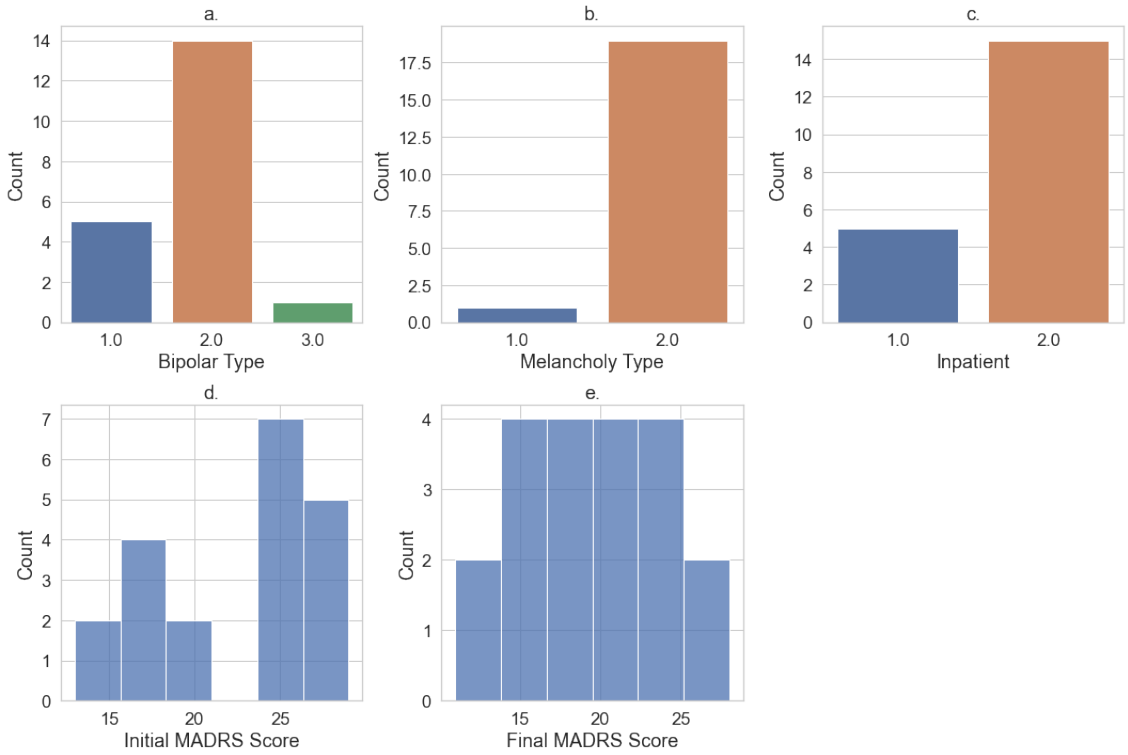

Figure 5: The distribution of depression-related information among the patients. a. Bipolar type, 1 for bipolar 2, 2 for unipolar, 3 for bipolar I, b. Melancholia indicator, 1 for having melancholia, 2 for not having it, c. Inpatient, 1 for inpatient, 2 for outpatient, d. MADRS score at the start of data collection, e. MADRS score at the end of data collection

from 5a.m to 12p.m. But it is important to note that these are just two patients and they do not represent the whole group. Simply select two other patients and rerun the code, I obtained another results where the conditioned patient is the more active patient. This analysis here is just to provide a glimpse into the activity pattern of a patient with depression and a patient without it.

6.2 Group analysis

From the demographic analysis, we can see that the two most outstanding features of the patients with depression are that they have rather low edu-

cation year, typically 6-10 years of education and unemployed as of the time of data collection. This results are largely consistent with what presented in the literature [12], but also goes against some literature suggesting that it is people with high income, high education, and married are at risk of suicidal or depression [13, 14]. It should be noted that this experiment is done on 20 patients, which is extremely small.

For the depression indicators analysis, we can see that most of the indicators are very imbalanced. This could negatively affect the performance of the prediction model. Another observation we can made is that the MADRS scores of the patients have decreased during the course of the information collection process. This is because the initial scores is rather left skewed, indicating that the initial scores were high, and the mean of the distribution was at around 22-23. But at the end of the collection process, the scores distribution have a bell curve shape with the mean of 20. This shows that the scores of the patients have decreased.

6.3 Prediction model

In the context of the current application, the training error is an acceptable error. The low standard deviation also shows that leaving a sample out of the model does not alter its performance too much. However, the mean and standard deviation of the testing error indicates that this model is very overfitted.

Considering the model as a whole, the samples were very limited, and some of the features are very imbalanced or skewed. This definitely played a role in the poor performance of the model. The simplicity of the model is another factor that could affect its performance. Linear Regression assumes that the independent variables exhibit a linear relation with the dependent model. This does not hold in all situations. Thus, to improve the model further, more patients data should be collected or we could use more complex model such as MLP.

7 Conclusion

In this project, I have conducted some analyses on the data of patients with and without depression. I have look at how the activity patterns of a typical controlled and conditioned patient look like, grouped by day of the week and

hour of the day. I then analyse the demographic and the distribution of the mental illness among the depressed patients. Finally, I built a regresison model with the aim to predict future depression state of the patient using available demographic data and depression state. Through the analysis, I have provided some insights on the behaviour pattern of a patient, as well as the type of demographic information that can be expected from the patient.

One of the limitations of the analysis is the lack of data. The extremely low generalization power of the model is suffered from this limitation. Another limitations could be the choice of model. Future iterations of the analysis should aim to gather more data and use a more complex prediction model. This could lead to an improvement in the prediction model, as well as the analysis. Then we can have an additional tool that can help in the treatment of depression among the patients.

References

[1] R. C. Kessler, P. R. Barker, L. J. Colpe, J. F. Epstein, J. C. Gfroerer, E. Hiripi, M. J. Howes, S.-L. T. Normand, R. W. Manderscheid, E. E. Walters *et al.*, "Screening for serious mental illness in the general population," *Archives of general psychiatry*, vol. 60, no. 2, pp. 184–189, 2003.

[2] P. W. Corrigan, A. Green, R. Lundin, M. A. Kubiak, and D. L. Penn, "Familiarity with and social distance from people who have serious mental illness," *Psychiatric services*, vol. 52, no. 7, pp. 953–958, 2001.

[3] M. B. Keller, "Depression: a long-term illness," *The British Journal of Psychiatry*, vol. 165, no. S26, pp. 9–15, 1994.

[4] M. A. Silver, M. Bohnert, A. T. Beck, and D. Marcus, "Relation of depression of attempted suicide and seriousness of intent," *Archives of General Psychiatry*, vol. 25, no. 6, pp. 573–576, 1971.

[5] K. Zhang, Y. Yao, and K. Hashimoto, "Ketamine and its metabolites: Potential as novel treatments for depression," *Neuropharmacology*, vol. 222, p. 109305, 2023.

[6] A. Abd-Alrazaq, E. Al-Jafar, M. Alajlani, C. Toro, D. Alhuwail, A. Ahmed, S. M. Reagu, N. Al-Shorbaji, M. Househ *et al.*, "The effectiveness of serious games for alleviating depression: systematic review and meta-analysis," *JMIR Serious Games*, vol. 10, no. 1, p. e32331, 2022.

[7] "Aripiprazole (abilify): Depression, major depressive disorder (mdd) appendix 5, validity of outcome measures." https://www.ncbi.nlm.nih.gov/books/NBK409740/, 2016, [Accessed 09-12-2023].

[8] S. Okada, M. Ohzeki, and S. Taguchi, "Efficient partition of integer optimization problems with one-hot encoding," *Scientific reports*, vol. 9, no. 1, p. 13036, 2019.

[9] X. Ying, "An overview of overfitting and its solutions," in *Journal of physics: Conference series*, vol. 1168. IOP Publishing, 2019, p. 022022.

[10] A. Zare, A. Ozdemir, M. A. Iwen, and S. Aviyente, "Extension of pca to higher order data structures: An introduction to tensors, tensor decompositions, and tensor pca," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1341–1358, 2018.

[11] C. Tudor-Locke, C. Leonardi, W. D. Johnson, and P. T. Katzmarzyk, "Time spent in physical activity and sedentary behaviors on the working day: the american time use survey," *Journal of occupational and environmental medicine*, pp. 1382–1387, 2011.

[12] K. A. McMillan, M. W. Enns, G. J. Asmundson, and J. Sareen, "The association between income and distress, mental disorders, and suicidal ideation and attempts: findings from the collaborative psychiatric epidemiology surveys," *The Journal of clinical psychiatry*, vol. 71, no. 9, p. 1497, 2010.

[13] T. Madsen, E. Agerbo, P. B. Mortensen, and M. Nordentoft, "Predictors of psychiatric inpatient suicide: a national prospective register-based study," *The Journal of clinical psychiatry*, vol. 72, no. 2, p. 15139, 2011.

[14] N. Akhtar-Danesh and J. Landeen, "Relation between depression and sociodemographic factors," *International journal of mental health systems*, vol. 1, pp. 1–9, 2007.