Investigating the Association of Social and Familial Factors on the Incidence of Psychotic Episodes in Schizophrenic Patients

Amar Dholakia

Apr. 19, 2021

Introduction

Schizophrenia is a debilitating neuropsychiatric disorder in which afflicted individuals suffer from a plethora of negative (flattened affect, cognitive impairment) and positive (hallucinations, and delusions)[1]. Familial support however is well-known to ameliorate a patient's condition and reduce detrimental symptoms1. Immediate families emerged as the primary caregivers of schizophrenic patients after the advent of deinstitutionalization in the United States[2,3] in the late 1950's. Soon followed an abundance of research on the factors of family care and environment, that generally corroborated strong family aided patient's in their perception and management of symptoms.

A recent survey (2014) of risk factors of symptomatic relapse in Tanzania demonstrated that both patients' perceptions of risk and actual outcomes of relapse were reduced when they felt support from family and peers[4]. Duckworth and Halpern (2014) and Pitt et al. (2013) touch on the effectiveness of social interaction and community-based therapy in the reduction of crisis and emergency services used by schizophrenic patients[5,6]. While demonstrative, these studies discussed 'relapse' in broad terms, whereas we are specifically interested in patient psychosis, necessitating a more specific analysis.

The effect of specifically of familial factors quantified as expressed emotion (EE) is a well-established risk factor of relapse and psychosis. Kavanagh[7] (1992) conducted a meta-analysis on studies investigating the effect of familial expressed emotion (EE) on symptomatic relapse, noting that on average, patients with high-EE family members (hostile, critical) were at higher risk of relapse than those with low-EE family members (positive, supportive). It was however noted that many of these studies did not use cross-sectional data, that is, measures of EE were mostly measured from retrospective patient reports.

Furthermore, Brown et al. produced a series of landmark papers (1962; 1972) that directly linked a greater risk of reoccurrence of severe symptoms in recently deinstitutionalized patients with 'highly emotionally involved' household members even after adjusting for symptom severity[8,9]. Brown et al. however measured only high EE (negative) criteria in his original study (1962), and while they measured 'warmth' as low EE criteria in the replicate study (1972), it was ultimately not included in the calculation of the aggregate EE index.

Despite low-EE data being available for half a century, Amaresha et al.[10] commented that it is under-leveraged in the literature, despite evidence that low-EE caregivers promote better patient outcomes. Further, Amaresha et al. critique that the EE scale historically has not considered cultural differences in family dynamic. Particularly, it is noted that percentage of high-EE families tends to be higher in Australia, Egypt, Asian families in the UK (particularly, Pakistani families) relative to the States, yet, high EE does not reliably predict patient relapse[11]. In Israel, to express anger is to show strength, and so high-EE percentage is overestimated compared to North American families[12]. There are multiple scales to measure EE, such as the traditional CFI and FAS, which exclude patient responses in favour of caregivers' only. Hence, we are motivated to use more cultural and context-independent factors, as well as integrate patient responses and sentiments. Social factors are also of interest to us for two reasons: modern technology allows us constant and convenient contact with otherwise distant family and friends; in our current COVID crisis, many people have reported feeling more isolated, and less connected than pre-pandemic[13], which may induce or exacerbate psychotic episodes in schizophrenic patients.

Considering social determinants in addition as a risk factor for psychotic episode, Leff and Vaughn (1976) replicated the results in Brown et al[14]. While social factors are defined slightly differently (reduced communication and awareness of external world), social withdrawal was also mentioned and noted to increase risk of symptomatic relapse. Leff and Vaughn however considered the influence of social factors in the framework of an EE-based caregiver-centric approach, further justifying investigation in a novel context-independent and patient-based exploration.

We would thus like to investigate the context-independent social and familial factors affecting occurrence of psychotic episode in past year (**EP**) in tandem. Our dataset contains survey data from the patients, emphasizing patients' beliefs and perceptions, instead of excluding it, like in past literature. We intend on performing multiple correspondence analysis (**MCA**) to identify appropriate groupings of variables, investigate the variance structure of the data, and correct for correlation between our variables of interest. We will regress these dimensions on EP, to attribute associations of risk of EP with groupings of variables identified in MCA, with the expectation of verifying previous findings in literature.

Data and Methods

The dataset is entitled Collaborative Psychiatric Epidemiology Surveys (CPES), 2001-2003 [United States] (ICPSR 20240); DS1[15]. The dataset contains a variety of demographic, psychological, and health data for over 20,000 respondents, all patients with a clinical diagnosis of schizophrenia. The majority of non-demographic data was self-reported, and collected in the form of questionnaires, and so is categorical. Our variables of interest include our binary response, "Did you have a psychotic episode in the past year?" (Yes/No), demographics (age, sex, number of members in household, race), and familial and social factors as outlined below. Outside of age, and number of household members, all factors were categorical, either binary (EP, sex), multi-levelled (race), or multi-levelled (>2 levels), on an ordered scale.

For all ordered multi-levelled social and familial factors, a score of 1 indicated a positive covariate outcome, whereas a score of 4 indicated a negative outcome. For visualization and interpretative purposes, we grouped our non-demographic variables into 'positive' and 'negative' factors, which correspond to the nature of the relationship/interaction being queried. For example, the question 'Does the respondent feel that the family works well together?' is categorized as a positive familial factor (**PFF**), whereas 'Does the respondent argue with their family?' would be a negative familial factor (**NFF**). Similarly, we denote positive and negative social factors as **PSF** and **NSF** respectively. Note that the scale interpretation holds between positive and negative factors, that is, the higher-valued the response, the worse the outcome.

A full description of our selected variables follow: • How well does the individual think family works together? (PFF) • How much does the individual trust/confide in family? (PFF) • How often does the individual express feelings to family? (PFF) • How close does the individual feel to family? (PFF) • How isolated does the individual feel to the family? (NFF) • How often does the individual argue with family over customs/traditions? (NFF) • How often does the individual talks to relatives/friends? (PSF) • How often does the individual rely on relatives/friends for serious problems? (PSF) • How often does the individual discuss worries to relatives/friends? (PSF) • How often does the individual argue with relatives/friends? (NSF)

Finally, we decided to create a new variable 'minority', which mapped all values corresponding to non-white races to 1 (indicating minority), else 0 (indicating white/non-minority).

Two datasets were created to perform analysis in parallel. In our first dataset, any respondents who did not answer for EP (missing values) were filtered out, but missing values for our selected covariates were retained (**DS1**). DS1 contained 1232 observations in total. Our second dataset (**DS2**) was created by filtering out any observations with missing values in any of the covariates, and consisted of 394 observations. Interestingly, there was a lot of missing data for non-minority respondents, leading our DS2 to contain no non-minorities. As such, we decided to drop the 'minority' variable for analysis on DS2. The impacts of this decision on analysis and modelling is explained in Results and Discussion.

Our data cleaning demonstrated that using only high-fidelity data (DS2) resulted in loss of over three quarters of our original dataset, DS1. This motivated us to perform and compare our analysis on both datasets. Our missing data was handled via an EM-based imputation whilst constructing our principal components (or 'dimensions' in MCA), using the missMDA package. The method is elaborated further in the MCA subsection of methods.

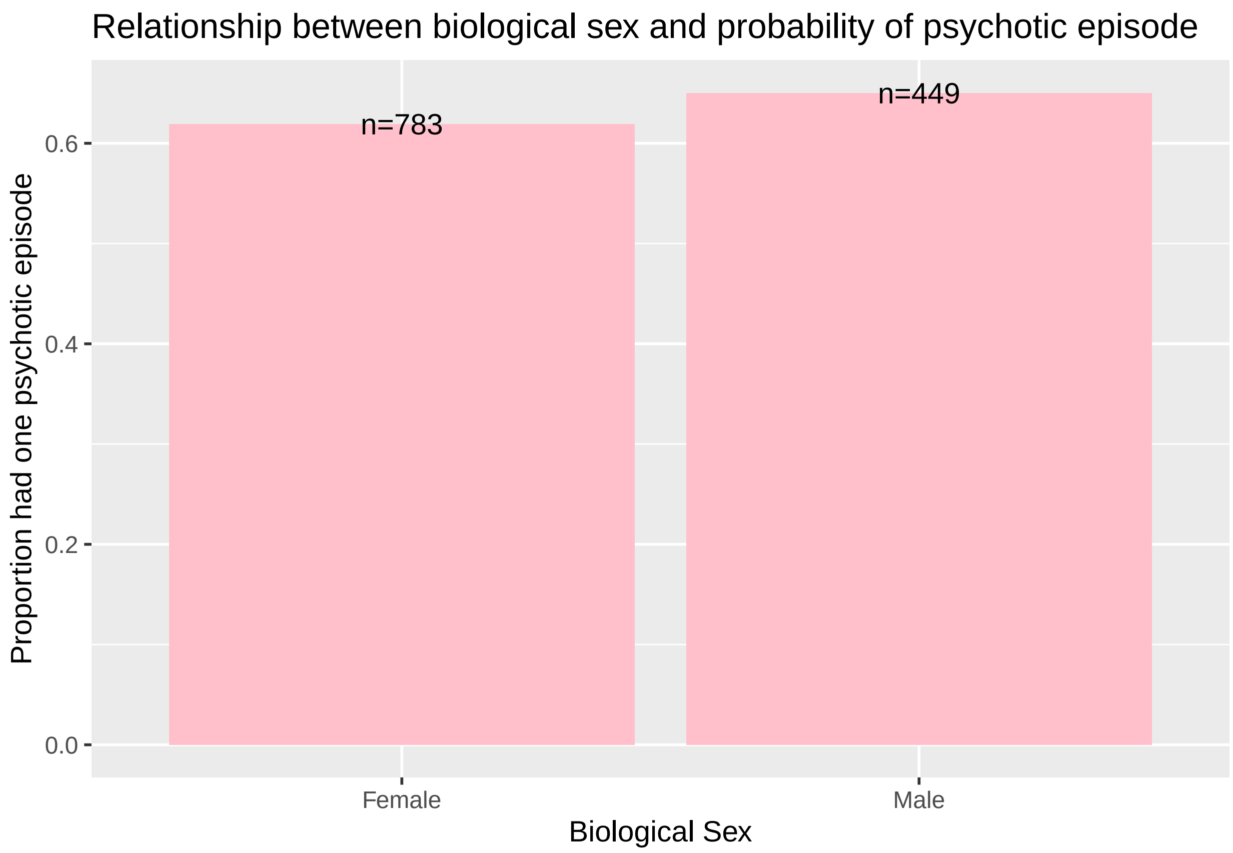
Relationship between biological sex and probability of psychotic episode
Proportion had one psychotic episode
0.0
0.2
0.4
0.6
n=783
n=449
Female
Male
Biological Sex

Fig 1.1 -

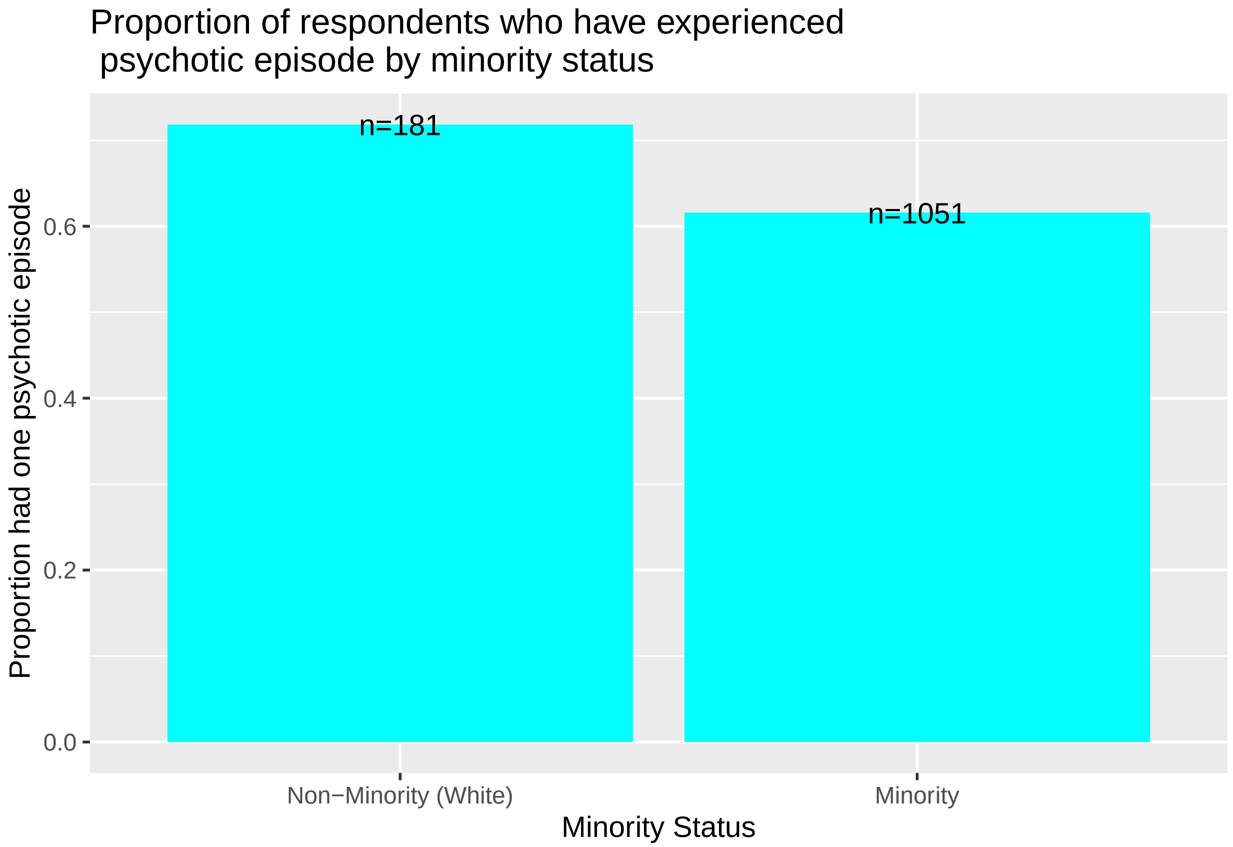

Fig. 1.1 - Comparing Prevelance Rates of EP by Sex and Minority Status

When comparing proportion of males who have experienced a psychotic episode, to females, we don't see much of a difference. We do note that a considerably greater proportion of females reported an episode (~63.6%) compared to males (~36.7%), which is slightly biased in terms of females. In comparison, the distribution of sex in the entire database is approximately 55:45 in favour of females.

Interestingly, we see that a smaller proportion of minorities have experienced a psychotic episode in the past 12 months (slightly above ~60%), compared to slightly above ~70% of non-minorities. This small discrepancy can be attributed to a variety of complex and cultural differences in familial and social interaction, stigmatization of mental health, type/quality/involvement of care given by caregiver, and quality of care received by medical professionals. Furthermore, there is an evident class imbalance between minorities and non-minorities in our dataset, with a strong underrepresentation of non-minorities. As we filtered out missing data based on our response variable, there is no way to internally fix this imbalance, and will unavoidably impact our results, as we intend to control for demographics.

```
## ‘summarise()‘ has grouped output by ’age_bins’. You can override using the ‘.groups‘ argument.
```

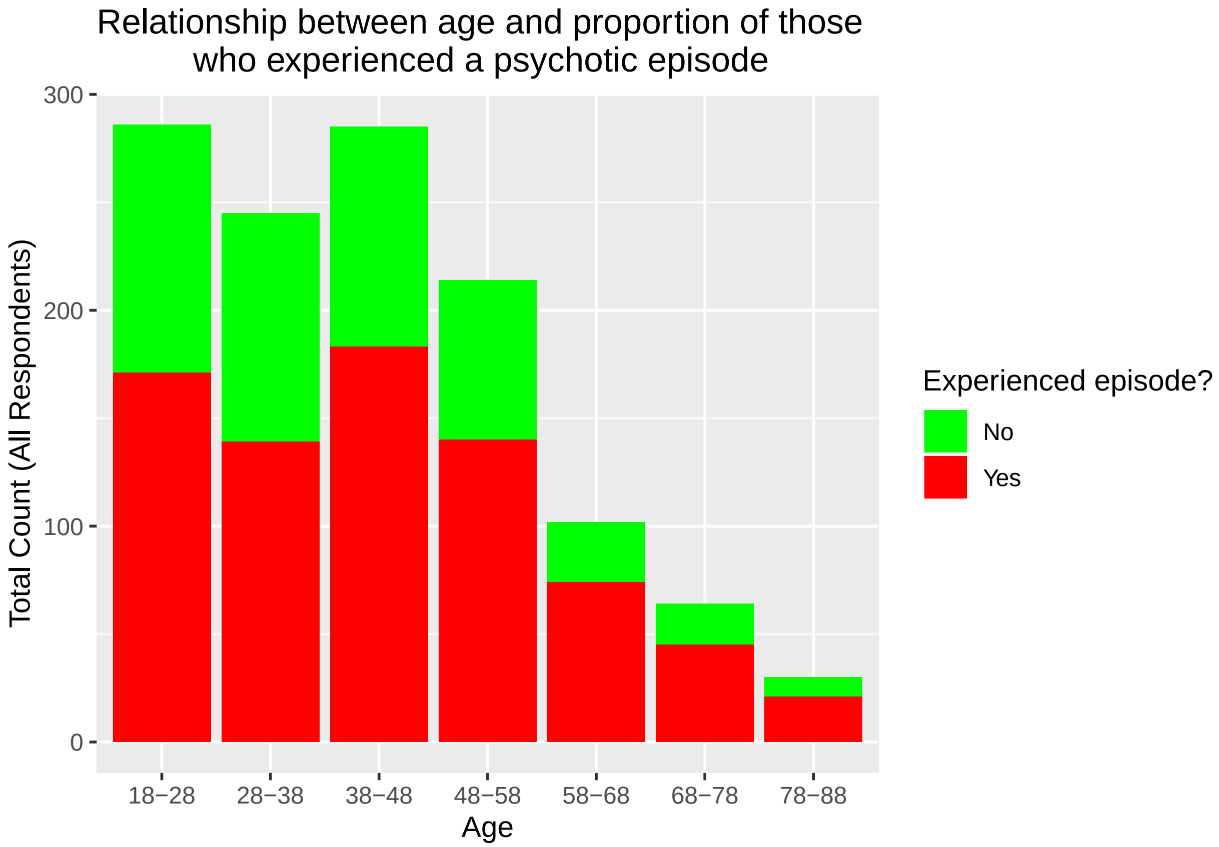

Fig 1.2 - Total respondents and Prevalence of EP by Age

Here, we see that the probability of incurring a psychotic episode tends to increase slightly with age. We do note that the elderly are underrepresented, and may reflect in our model estimates as we will adjust for demographics.

```
## ‘summarise()‘ has grouped output by ’fam’. You can override using the ‘.groups‘ argument.
```

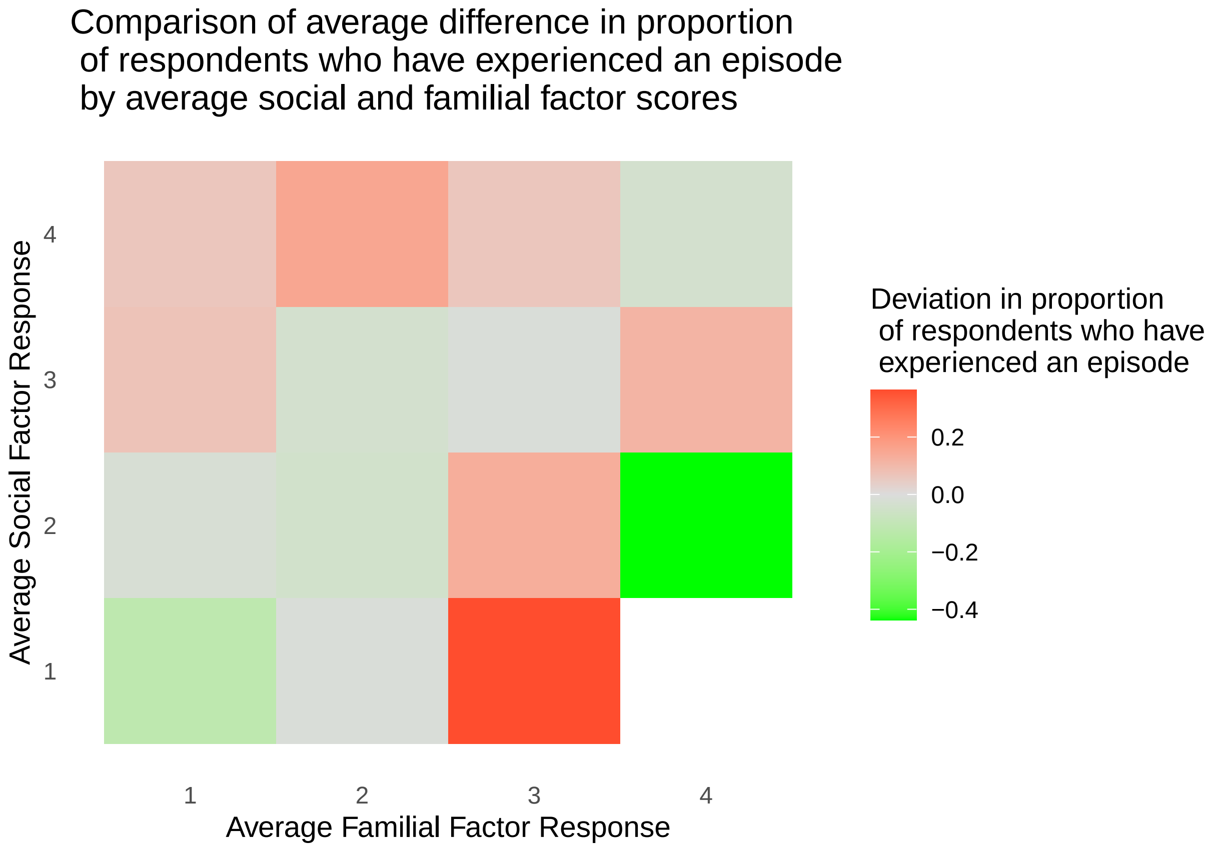

Figure 1.3 - Average deviation in prevalence of EP from mean prevalence for joint bin of average social and familial scores

Fig. 1.3 summarizes differences in proportions of respondents who have experienced a psychotic episode, grouped jointly by their average social and familial factor scores. When conditioning on the familial factor response, we generally see that the greater the average social factor response (worse outcome), the greater the deviation in proportion of respondents who have suffered a psychotic episode. This hold for the first 3 familial factor bins, but not so much for the fourth one, however, this can be explained by not many respondents belonging to the highest social or familial factor bins.

Furthermore, when conditioning on social factor bins of 1 and 2, we see that with increases in average familial factors, there is a greater rate of psychotic episodes by group. Interestingly, we do note however that at higher social scores, these trends do not hold; particularly when social factors are conditioned on bins 3 and 4, the incidence of psychotic episodes does not always increase with familial factor score. This might be again explained by smaller sample sizes producing more biased proportion estimates.

However, we also note that higher social and familial factor scores may also indicate that the respondent is very independent and/or 'high-functioning', perhaps due to manageable or mild symptoms. This is in contrast to the converse extreme, where patients are much more dependent on caregivers, as their symptoms are much less manageable. Although we do not do so in our analysis, it would be interesting to try to tease apart and compare incidence of episodes between independent from isolated individuals at the high ends of our scale.

Methods

MCA and Bayesian logistic regression were the primary methods of analysis in our investigation. MCA is an PCA analogue for categorical/scale data, and excellent for survey data[16]. MCA was used to analyze correlations and similarities within the familial and social factors, independent of the modelled response. MCA also produced independent dimensions, and dimension scores for each individual, controlling for expected collinearity between factors. MCA was performed on DS1 and DS2 to investigate the effect of missing data on intervariable relationships and modelling.

The MCA function from the FactoMineR package[17] was used to construct dimensions. For DS2 (no missing data), dimensions were constructed using the all familial and social factors as active variables. Demographic variables (excluding minority) were set as supplementary variables and did not contribute to the axes construction. For DS1, it was necessary to impute missing values prior to dimension construction. Imputation was handled automatically using the *imputeMCA* function from the missMDA package in R[18]. The *imputeMCA* function uses an EM-based algorithm to impute missing categorical data with the probabilities that the individual would score a certain level in that class. First the membership probabilities are estimated, then MCA is performed on the imputed data set. This repeats iteratively until convergence. Using the *estim_ncpMCA* function from the MissMDA package, an optimal number of 4 dimensions were estimated to be used for imputation.

MCA provided an estimate of dimension eigenvalue and proportion of variance explained (%EVR). Per Abdi[19], MCA with the indicator matrix 'severely underestimates eigenvalues', hence the proposed correction (see below) was applied to produce more robust %EVR estimates. Dimensions 2 and 4 were selected for regression, as explained in results.

Two logistic regression models were fit using STAN, and both were fit on DS1 and DS2 once.

Model 1:

$$logit(Y) = \beta_0 + \beta_1 \cdot D4_c + \beta_2 \cdot a_c + \beta_3 \cdot h_c + \beta_4 \cdot s + \beta_5 \cdot m$$

where Y indicates whether or not the respondent had a psychotic episode in the past 12 months, $D4_c$ is the centered and scaled scores for dimension 4, a_c is the centered and scaled age, h_c is the centered and scaled number of household members, and s and m are the sex and minority status of the respondent, respectively.

Model 2:

$$logit(Y) = \beta_0 + \beta_1 \cdot D4_c + \beta_2 \cdot D2_c + \beta_3 \cdot a_c + \beta_4 \cdot h_c + \beta_5 \cdot s + \beta_6 \cdot m$$

where $D2_c$ is the centered and scaled scores for dimension 2. The rest is specified as in Model 1.

As all numeric covariates were centered and standardized, we investigated effects in the framework of standard deviations. Thus, we felt that priors of $N(0, 1)$ on all β_i were suitable. For DS1, we specified the full model. For DS2, we removed minority, as there were no remaining non-minorities after filtering out missing values.

1000 samples were drawn from the posterior density, from which the mean point estimate, and 95% credible interval for β_i were calculated on a log-odds scale. Expected log-predictive densities were computed for both models for comparative purposes, and discussed in results.

Results

MCA

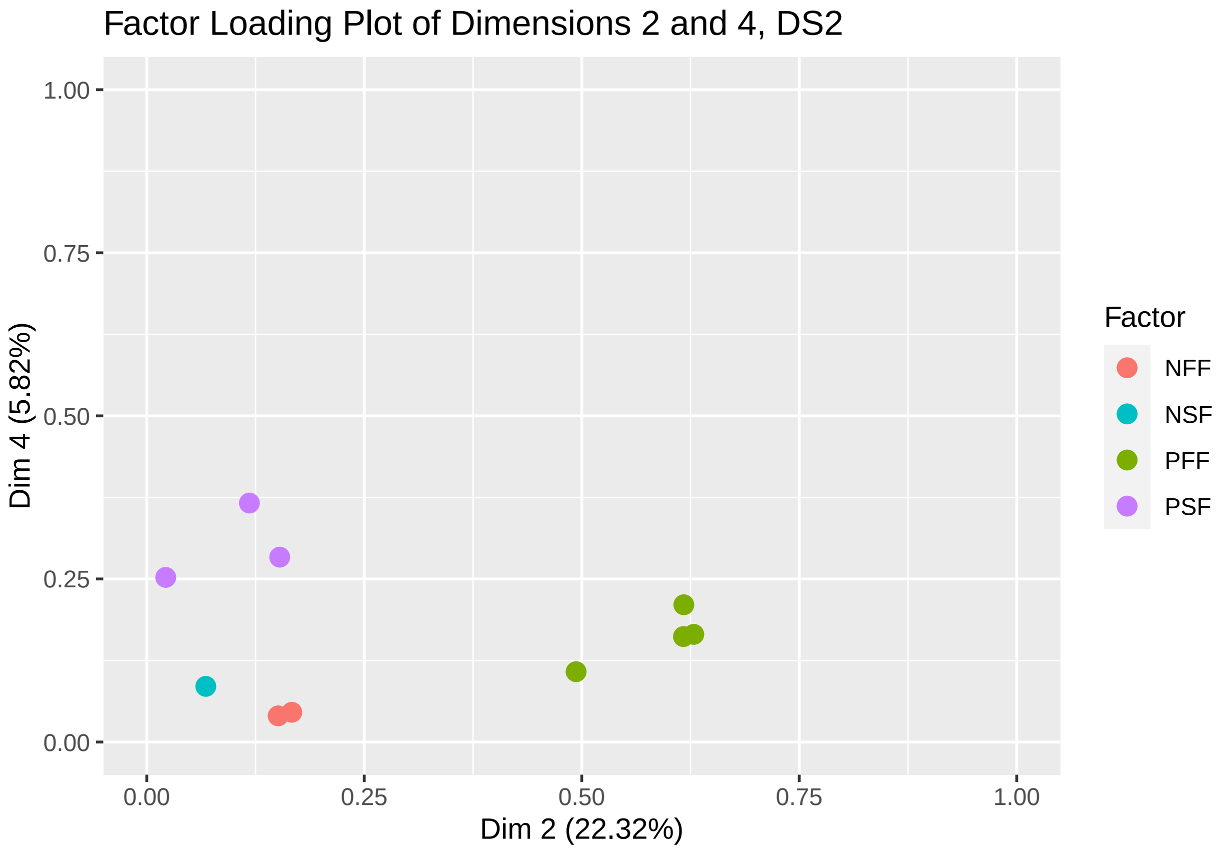

Figure 2.1 - Factor loading plot for Dimensions 2 and 4 for DS2. Familial and social factors were included as active variables, demographics as supplemental variables (not shown).

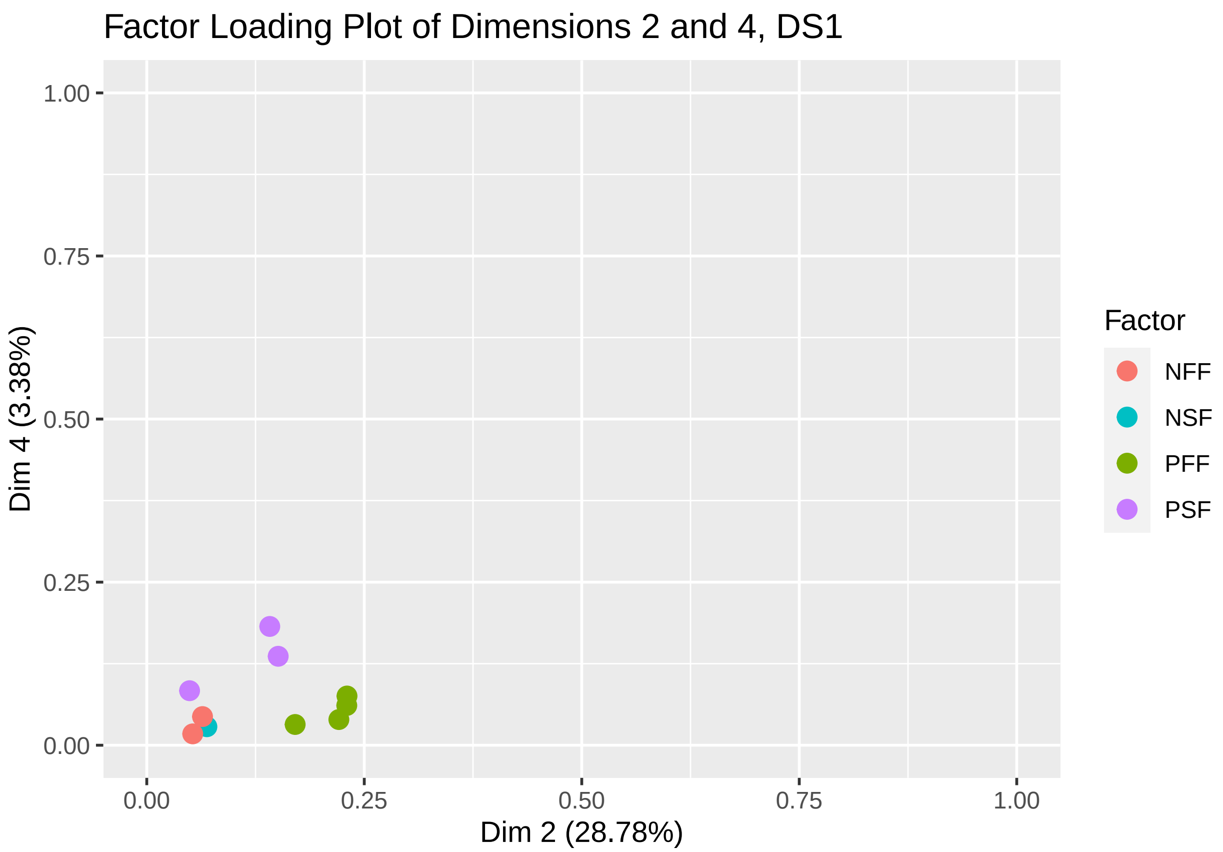

Fig. 2.1 and 2.2 correspond to the loading plots for Dim 2 against Dim 4 for DS1 (full) and DS2 (lossy) datasets, respectively. After correcting for eigenvalue underestimation, we see that Dim 2 and Dim 4 explains 22.32% and 5.82% of the total variance between active variables respectively in DS1, compared to 28.78% and 3.38% of the total variance in DS2. Furthermore, our plots suggest that PFFs and PSFs help explain the majority of the variation in dimensions 2 and 4, respectively, relative to other factors. PFF focused on familial interaction, feelings of trust, closeness and cohesiveness of family contribute greatly to the construction of both the first and second dimension, whereas PSFs describe support, reliance and interaction with peers. Interestingly, both NFFs and NSFs do not seem to contribute to the construction of either dimension, and thus had no interpretation in our subsequent regression. As demographic variables were included as supplementary variables (excluded from the MCA), they (age, gender, and number of household members) are not correlated with the dimensions, nor the other dominant variables of interest, i.e. the PFFs. Hence, we would not be introducing multicollinearity into a model that adjusts for demographic information.

Interestingly, after missing data imputation in DS2, PFFs and PSFs are still the dominant loading factors for Dim 2, and Dim 4 respectively, as seen in Fig. 2.2. However, PFFs and PSFs incur the most shrinkage in their estimated contribution to dimension, with PFFs shrinking from ~60% to about ~20% (on average) in Dim 2, and PSF shrinking from 30% to 12.5% (on average). This is perhaps expected, as familial factors contain more missing data (n=838/1232), compared to social factors (n=608/1232), whereas demographic data is replete (n=0). Despite shrinkage, we still see that the PFFs still dominate the loadings of both Dim 1 and Dim 2, relative to other factors. Missing data imputation does not seem to greatly affect the structure of the factor loading plot, suggesting that the MCA still identifies similar covariance relationships in our data after imputation. The comparisons of models fit on DS1 and DS2 are valid, after adjusting for sample size.

Modelling

Table 1: Table 2.1 - Model 1, DS1, Parameter and 95% Credible Intervals Estimates

Parameter	Mean	Lower	Upper
Dim 4	0.11500876	0.005028168	0.22735861
Age	0.18573996	0.058603857	0.30828554
Sex	0.18086893	-0.057763764	0.42564485
# House Members	0.02313832	-0.047072855	0.09390161
Minority	-0.36948387	-0.715829881	-0.06843375

Table 2: Table 2.2 - Model 2, DS1, Parameter and 95% Credible Intervals Estimates

Parameter	Mean	Lower	Upper
Dim 2	0.06966789	-0.050026262	0.18600039
Dim 4	0.11355739	0.003895853	0.24341989
Age	0.19121470	0.068974825	0.31903988
Sex	0.17759925	-0.102596493	0.41254631
# House Members	0.01937666	-0.056923416	0.09485925
Minority	-0.36987134	-0.695965447	-0.04748644

Table 3: Table 2.3 - Model 1, DS2, Parameter and 95% Credible Intervals Estimates

Parameter	Mean	Lower	Upper
Dim 4	0.240953937	0.01050562	0.4759507
Age	0.060423132	-0.14705524	0.2747535
Sex	-0.026795229	-0.42300389	0.3745994
# House Members	0.008526467	-0.13402623	0.1482570

Table 4: Table 2.4 - Model 2, DS2, Parameter and 95% Credible Intervals Estimates

Parameter	Mean	Lower	Upper
Dim 2	0.138576211	-0.06959723	0.3447660
Dim 4	0.236382939	0.01644251	0.4560011
Age	0.065555241	-0.13216913	0.2797229
Sex	-0.049393704	-0.43932555	0.3450340
# House Members	0.005949601	-0.13097137	0.1378162

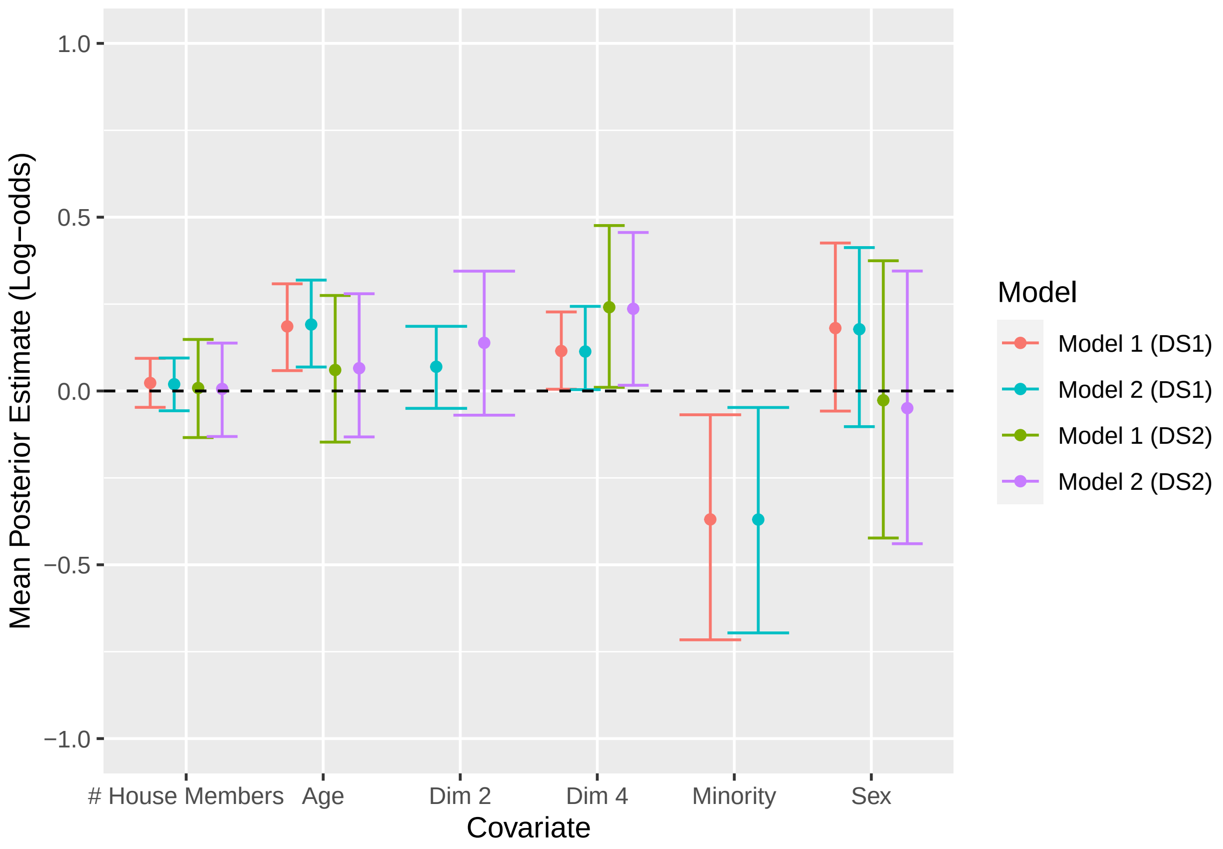

Figure 2.3 – Mean Posterior Estimate of Log-odds ratio with 95% credible intervals for demographic variables (age; gender; number of household members, and minority status) for both candidate models.

After adjusting for demographics, in both models, we a significant or near-significant positive effect in Dim 4. Unfortunately, in model 2, while Dim 2 is positively associated with an increased risk of EP as expected, this effect does not seem to be significant. Comparing on the same dataset, coefficient estimates between models 1 and 2 are generally consistent, and the inclusion of Dim 2 in the second model does not (at first glance) negatively impact interpretation of other covariates.

While not detecting a significant effect from the PFF in Dim 2 was disappointing, we were pleasantly surprised by the association detected between PSF in Dim 4 and EP. A one standard deviation increase in the scores for Dim 4 is associated with an on average increase in 0.240 and 0.236 in log-odds of EP (DS2), and 0.116 and 0.113 in log-odds of EP (DS1), for Models 1 and 2, respectively. This suggests that in aggregate, if a patient reports worse or scarcer PSFs (higher respondent score), then they were at a slightly higher risk of experiencing an EP in the past year. A similar interpretation can be made for the PFFs described by Dim 2, however, again, no significant association was found. While both effects are positive, as expected, we do however note that the magnitude of the estimated effects for Dim 2 and 4 are both quite small, and that the practical association with increased risk of EP is not much greater than 0.

Interestingly, the inclusion of demographics elucidated intriguing risk factors not obviously indicated by our EDA. Particularly, when Models 1 and 2 (respectively) are fit on DS1 (missing data), we see that with one standard deviation of age, we see small positive increases of 0.181 and 0.191 in log-odds of EP. We note that while this is significant, there magnitude of this association is small, and thus there is practically no effect. This effect is not significant when fit on DS2. Gender and household members are not a significant factor in any model or dataset. Perhaps most interesting is the strong negative association between minority (non-white) status and log-odds of EP. Minority status reported the strongest significant effect of all variables considered with an average decrease of 0.369 and 0.370 in log-odds of EP. This suggests that minority status a strong mitigating factor in the risk of EP.

Model Comparison

Finally, we compared model performance and validity by estimating the leave-one-out expected log predictive densities (LOO-ELPD or ELPD). Results of LOO-ELPD are summarized in the tables below.

Table 5: Table 2.5 - Comparison of LOO-ELPDs between Models 1 and 2, fit on DS1 and DS2. Valid comparisons between models can only be made on the same dataset

Dataset	Model 1	Model 2
DS1	-1.07e+06	-1.08e+06
DS2	-1.15e+05	-1.17e+05

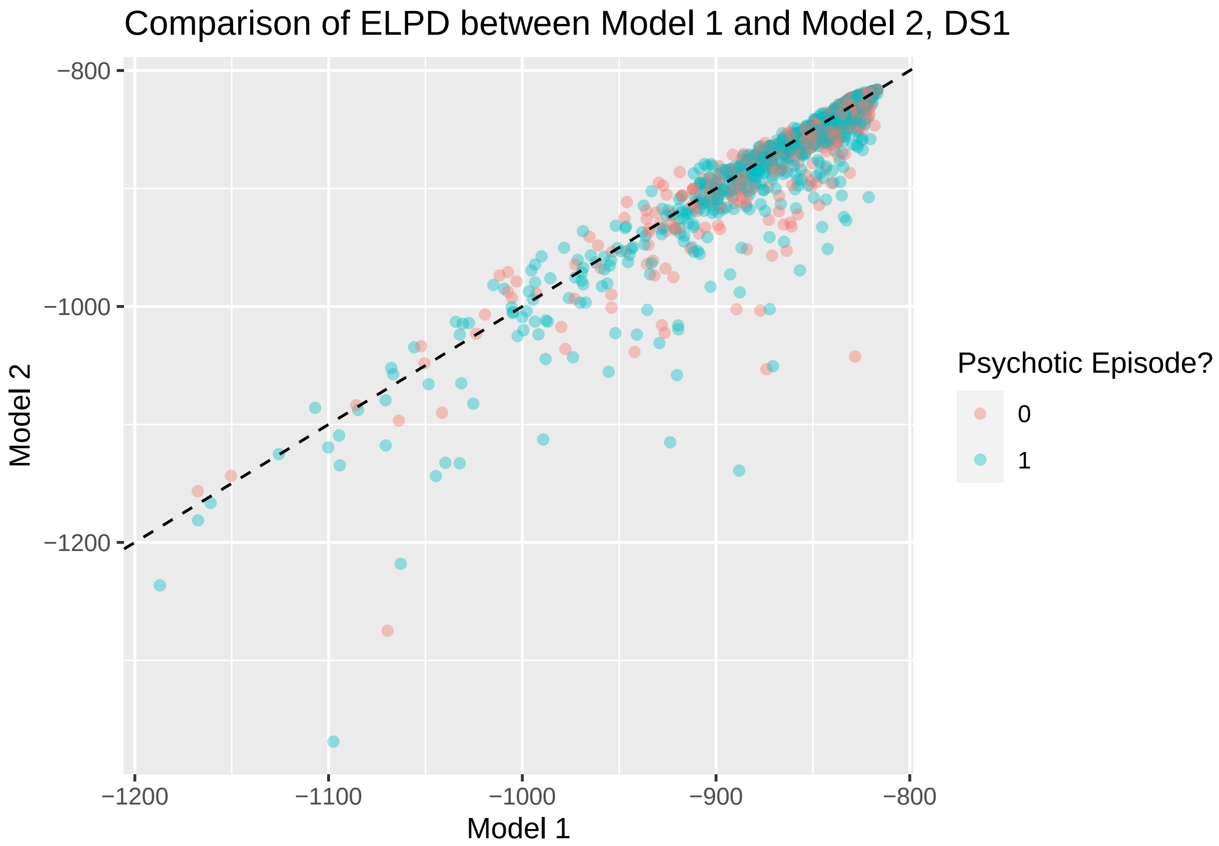

Figure 2.4 - Comparison of ELPD between Model 1 and 2 fit on DS2 (missing data), with log pointwise prediction grouped by whether respondent experienced a psychotic episode in the last 12 months.

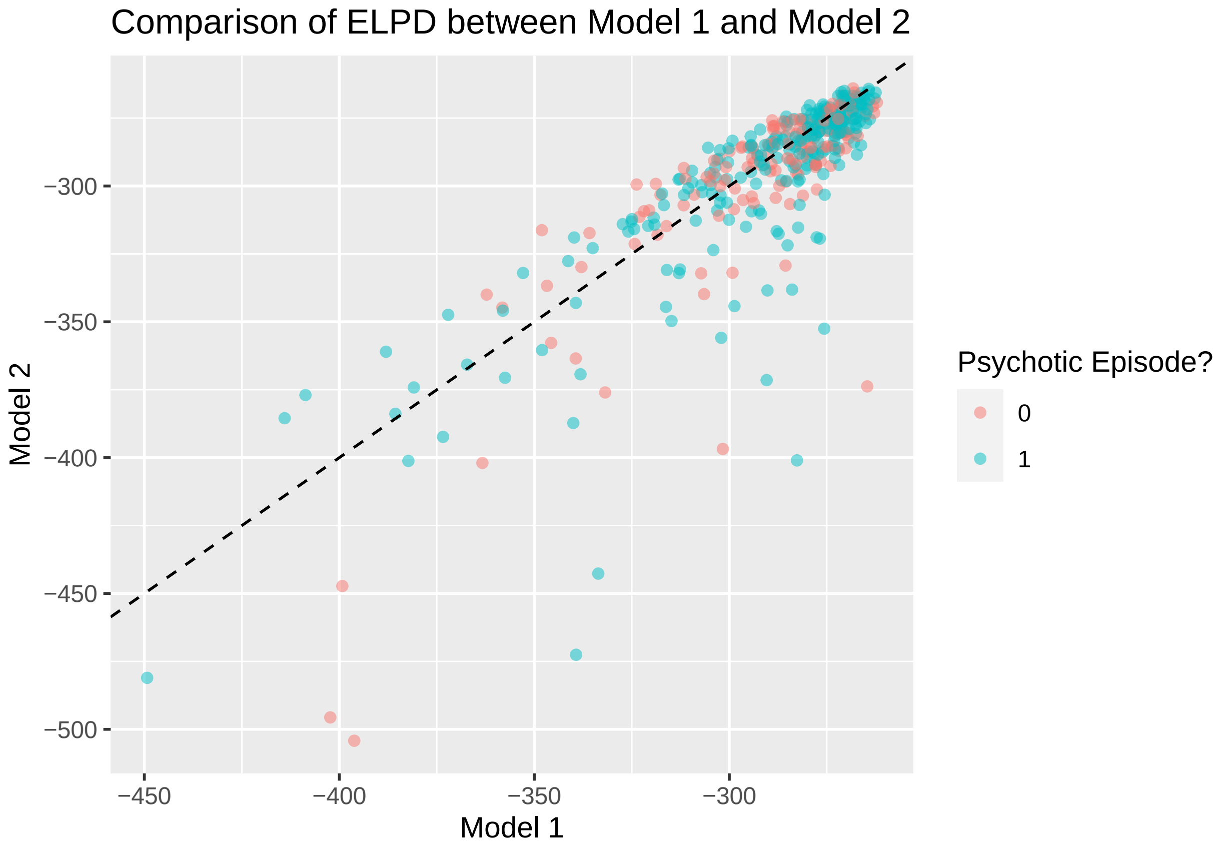

Figure 2.5 - Comparison of ELPD between Model 1 and 2 fit on DS1 (no missing data), with log pointwise prediction grouped by whether respondent experienced a psychotic episode in the last 12 months.

It is quite clear that Model 1 (demographics + PSF) performs significantly better than Model 2 (demographics + PSF + PFF) for both datasets. Thus, we elect to reject the addition of Dim 2 (PFF) in our final models. We also note that excluding the missing data greatly improves the performance of both models, which is to say, despite imputation and a larger sample size, we do not arrive with as strong a model compared to if we had a complete dataset.

It is worth noting that the k-values for PSIS-LOO w ere reported to be uniformly greater than 1, which calls into question the reliability of the metric when comparing our models. However, from the plots above, we see that for quite a few data (bottom right), Model 1 has a much greater ELPD, whereas Model 2 shows comparable ELPD for the rest of the data. We noted no discernable differences in EP to explain this discrepancy. This suggests that Model 1 generally fits better to the data than Model 2.

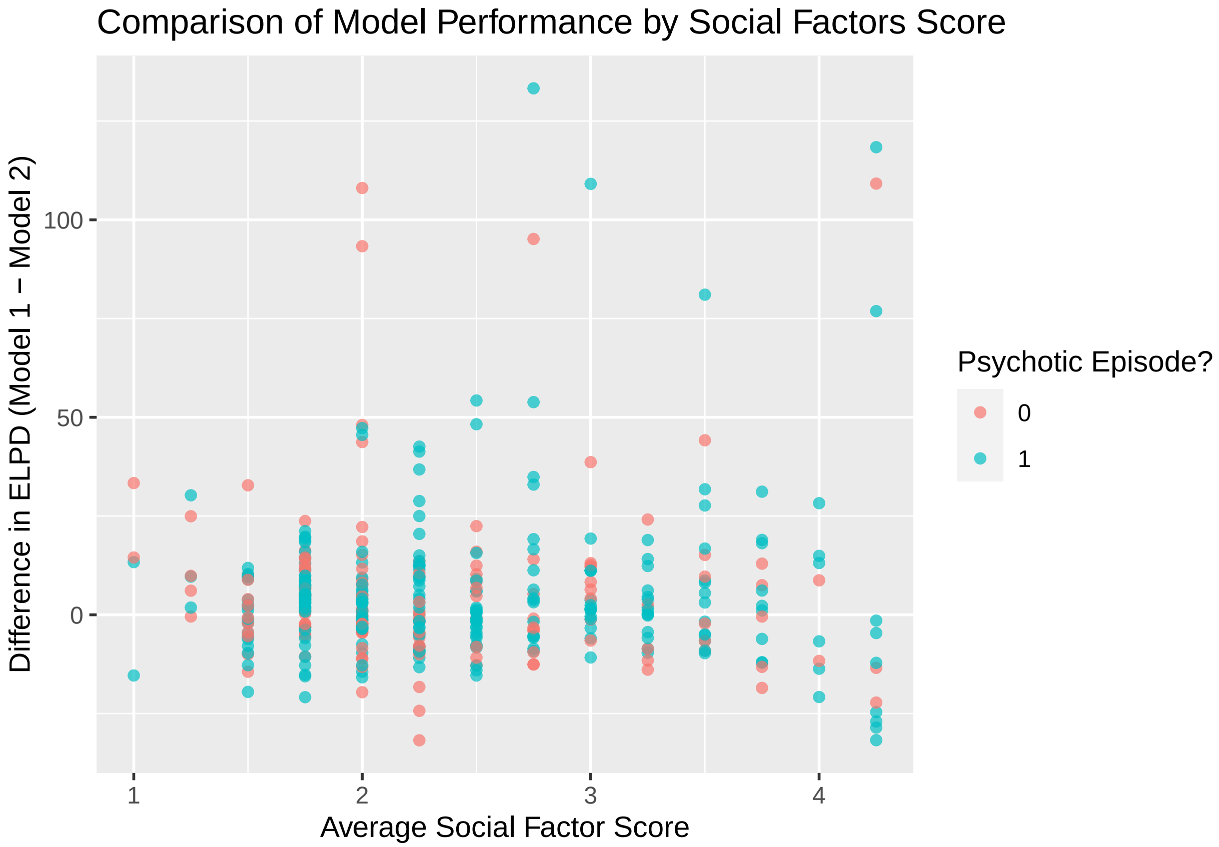

Figure 2.6 - Comparison of ELPD between Model 1 and 2 fit on DS2 (no missing data), as a function of average social factor score. Log pointwise prediction grouped by whether respondent experienced a psychotic episode in the last 12 months.

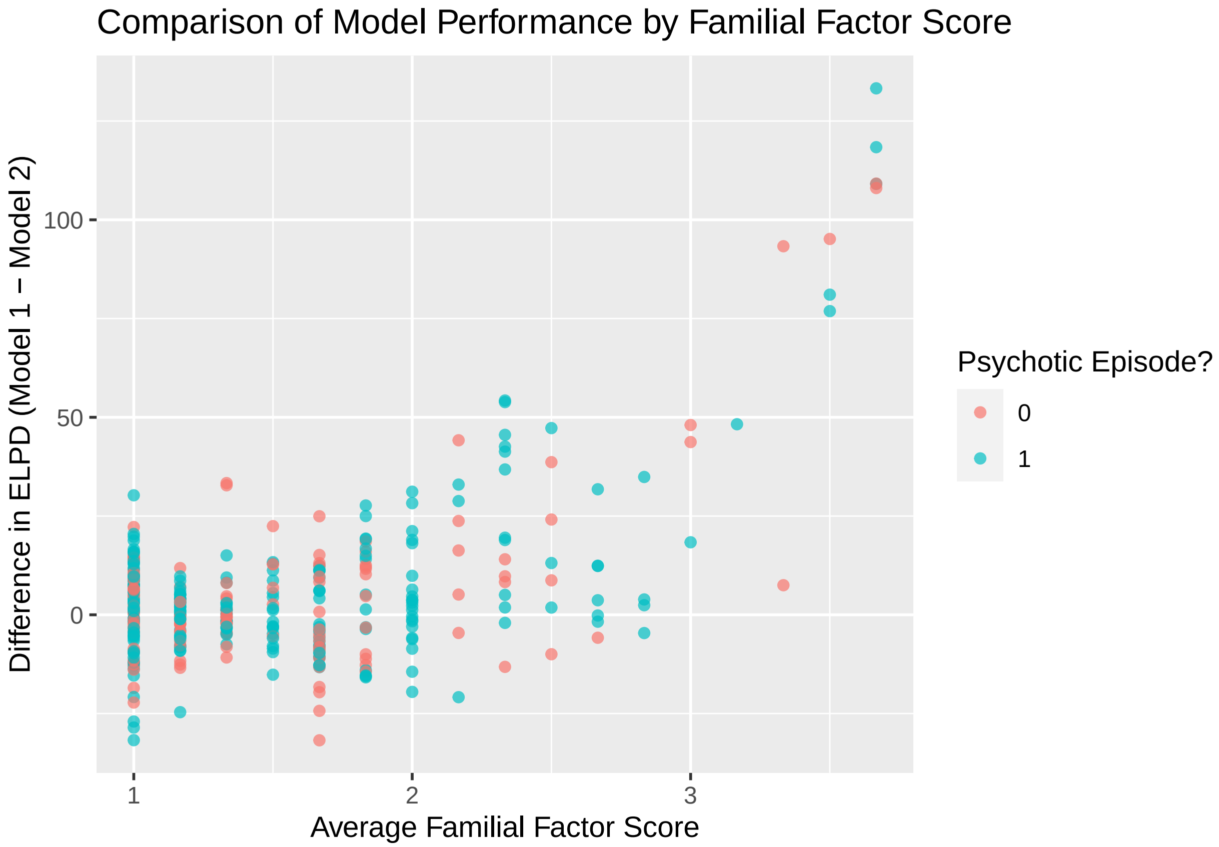

Figure 2.7 - Comparison of ELPD between Model 1 and 2 fit on DS2 (no missing data), as a function of average social factor score. Log pointwise prediction grouped by whether respondent experienced a psychotic episode in the last 12 months.

We also investigated differences in model performance as a function of average social and familial factor scores. In Fig 2.6, as average social score increases (less favourable social outcomes), Models 1 and 2 are comparable in performance. Intriguingly, as the average familial score increases as in Fig 2.7, Models 1 strongly begins to outperform Model 2, despite the latter including PFF as a covariate and the former excluding it, while we might have expected the converse to occur. This suggests that adding Dim 2/PFF adds unnecessary complexity to our model, causing overfitting, and that Dim 4/PSF and demographics are better in modelling association with EP, whereas Dim 2/PFF do not add any important information. The same plots were not produced for DS1 due to the bulk of missing data obscuring the visual.

Discussion

We propose a final model to measure the association of familial and social factors and risk of psychotic episode:

$$logit(Y) = \beta_0 + \beta_1 \cdot D4_c + \beta_2 \cdot a_c + \beta_3 \cdot h_c + \beta_4 \cdot s + \beta_5 \cdot m$$

which was proposed as Model 1, and excluded the second dimension (PFF).

We used the response “Have you experienced a psychotic episode in the past 12 months?” as proxy, assuming that the respondents’ perception of familial and social factors did not vary in that time period. In summary, we find that a favourable score for questions regarding positive social interactions is significantly associated with a decreased risk of psychotic episode in respondents, after adjusting for age, sex, number of household members, and minority status.

Our finding that favourable positive social factors (PSF) were associated with decreased risk of psychotic episode was slightly surprising, but not totally unexpected, based on our previous discussion of the literature[5,6,14]. We had strong expectations that PFFs would be a strong mitigating factor of psychotic episodes, but regrettably, we were not able to attribute any significant association. Potential reasons for this result is explained in the limitations below. Encouragingly, we did note a potential reduction in risk, and given more complete data, we still expect there to be a positive association between familial factors and EP.

Interestingly, our model suggests that minority status is a strong mitigating factor of psychotic episode. Literature suggests that some minority groups do in fact report much lower incidences of psychotic episodes due to differences in how families and caregivers in various cultural groups interact with schizophrenic patients[11]. Furthermore, this dataset was collected in the United States, where mental health prevalence is generally lower minority subgroups, particularly African Americans[20], and may also reflect in our findings. Alternatively, mental health in general is often very stigmatized in both Western and some non-Western Asian ethnic groups, especially by family members[21,22]. In some Asian cultures, stigmatization from one's own family members is so deep that victims often suffer real, tangible declines in quality of life, independence, and ability to function[23,24]. We suggest that such cultural stigma may be associated with a greater risk of episode. Ultimately, however, further analysis on ethnic subgroups will need to be re-done, as provided by this dataset.

Our goal to incorporate MCA to eliminate multicollinearity, whilst identifying risk factors was incredibly fruitful. We identified two main dimensions (Dim 2 and Dim 4) that cleanly separated PFF and PSF, respectively, allowing us to investigate aggregate associations with these variable 'bundles' and the risk of EP. It was noted that percentage variation explained by the chosen dimensions is also quite low, but their inclusion in regression was justified as they carry meaningful and significant information about our covariates of interest. Admittedly, we would have liked to select our first two dimensions, but the first dimension was not associated with any change in risk of EP. Despite our success with PFF and PSF, we were not able to incorporate NSF or NFF in a meaningful or interpretable way in our models, as they did not contribute strongly to the construction of any of the regressed dimensions.

Limitations and Conclusions

The primary limitation of our models is interpretability. We regressed our response, patient incidence of psychotic episode in past year, on dimension scores, which are real-valued and practically arbitrary. However, as dimensions are linear combinations of our survey variables, we can (very) generally claim that mean increase in dimension unit is positively associated with an overall mean increase in all survey categories. From our MCA factor loading plots, we were able to verify groups of correlated variables by dimension, that were in turn grossly associated with incidence of EP. However, our models are not able to attribute this effect to any one factor within that bundle, and covariate interpretability remains an open question in the field of factor analysis regression.

Some inconsistencies in our regression analysis (insignificant associations) and MCA (poor NFF and NSF performance) may be attributed to flaws in our variable selection. The dataset contains 15 family factor variables, of which 6 were subsetted (4 positive; 2 negative), and 16 social factor variables, of which 4 were subsetted (3 positive; 1 negative). Clearly, there was an unintentional bias towards the selection of positive over negative variables. Variable selection was deliberate, but selection bias unintended, as we decided to focus more on variables that quantified interactions with family, such as arguing and confiding. An immediate future consideration would re-running this analysis on the entire set of familial and social factors to avoid biased results.

Despite encouraging interpretative results, ELPD demonstrates that our models generally fit very poorly on the data. We believe this is principally due to our model being too simple/underfitting, both by feature design and by variable selection. Furthermore, underfitting is a likely culprit, as previous attempts to model risk factors of other characteristics of schizophrenia (etiology) were challenging due to the complex nature of the disease22. More complex models featuring polynomial or interaction terms might better explain associations between EP and covariates. LASSO regression could aid in variable selection, whilst also controlling multicollinearity, thus is an alternative to MCA.

Furthermore, our results demonstrated that missing data non-negligibly impacted our model estimates. The minority variable was excluded from our non-missing data models (justified in Methods), but strikingly was the strongest predictor of EP. Perhaps counterintuitively, the credible intervals are for the non-missing data models are wider than those of the missing data models, while we might expect the converse. Note, however, that in both cases, our MCA scores are fully observed all individuals, despite imputation methods used in the non-missing case. Furthermore, by not excluding missing data, we incorporate many more observations (scores) into our model fit, greatly reducing credible interval width for our non-missing data models26. As such, we suspect that credible intervals for the non-missing data cases may not be reliable, and are in fact under-estimated, relative to the known data models.

Despite these drawbacks, we are still optimistic that our methods have demonstrated that better and more hospitable familial and social relationships are associated with a reduction in risk of psychotic episode in schizophrenic patients. Particularly, we found that better PSFs, factors focused on positive aspects of social relationships, were the strongest indicator of such a reduction. One advantage of our method is its ability to harness 'big data' and build complex and scalable models (even in face of multicollinearity) that help better understand structural interplay of patient outcomes and complex epidemiological covariates. By addressing some of the limitations, we hope that we can build a more complex and valid model that verifies the role of other familial and social factors in patient outcomes.

Works Referenced

[1] Flaum M, Schultz SK. The core symptoms of schizophrenia. Annals of medicine. 1996 Jan 1;28(6):525-31.

[2] Lamb HR, Bachrach LL. Some perspectives on deinstitutionalization. Psychiatric services. 2001 Aug;52(8):1039-45.

[3] Caqueo-Urízar A, Rus-Calafell M, Urzúa A, Escudero J, Gutiérrez-Maldonado J. The role of family therapy in the management of schizophrenia: challenges and solutions. Neuropsychiatric disease and treatment. 2015;11:145.

[4] Sariah AE, Outwater AH, Malima KI. Risk and protective factors for relapse among individuals with schizophrenia: a qualitative study in Dar es Salaam, Tanzania. BMC psychiatry. 2014 Dec;14(1):1-2.

[5] Duckworth K, Halpern L. Peer support and peer-led family support for persons living with schizophrenia. Current opinion in psychiatry. 2014 May 1;27(3):216-21.

[6] Pitt V, Lowe D, Hill S, Prictor M, Hetrick SE, Ryan R, Berends L. Consumer-providers of care for adult clients of statutory mental health services. Cochrane Database of Systematic Reviews. 2013(3).

[7] Kavanagh DJ. Recent developments in expressed emotion and schizophrenia. The British journal of psychiatry. 1992 May 1;160(5):601-20.

[8] Brown GW, Monck EM, Carstairs GM, Wing JK. Influence of family life on the course of schizophrenic illness. British journal of preventive & social medicine. 1962 Apr;16(2):55.

[9] Brown GW, Birley JL, Wing JK. Influence of family life on the course of schizophrenic disorders: A replication. The British Journal of Psychiatry. 1972 Sep;121(562):241-58.

[10] Amaresha AC, Venkatasubramanian G. Expressed emotion in schizophrenia: an overview. Indian journal of psychological medicine. 2012 Jan;34(1):12-20.

[11] Bhugra D, McKenzie K. Expressed emotion across cultures. Advances in Psychiatric Treatment. 2003 Sep;9(5):342-8.

[12] Heresco-Levy U, Greenberg D, Dasberg H. Family expressed emotion: Concepts, dilemmas and the Israeli perspective. Israel journal of psychiatry and related sciences. 1990.

[13] Hwang TJ, Rabheru K, Peisah C, Reichman W, Ikeda M. Loneliness and social isolation during the COVID-19 pandemic. International Psychogeriatrics. 2020 Oct;32(10):1217-20.

[14] Vaughn C, Leff J. The measurement of expressed emotion in the families of psychiatric patients. British journal of social and clinical psychology. 1976 Jun;15(2):157-65.

[15] Margarita A, Jackson JS, Kessler, RC, Takeuchi D. Collaborative Psychiatric Epidemiology Surveys (CPES), 2001-2003. United States. Ann Arbor, MI: Inter-university Consortium for Political and Social Research. 2016-03-23. Retrieved from: https://doi.org/10.3886/ICPSR20240.v8

[16] Ayele D, Zewotir T, Mwambi H. Multiple correspondence analysis as a tool for analysis of large health surveys in African settings. African health sciences. 2014;14(4):1036-45.

[17] Lê S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. Journal of statistical software. 2008 Mar 18;25(1):1-8.

[18] Josse J, Husson F. missMDA: a package for handling missing values in multivariate data analysis. Journal of Statistical Software. 2016 Apr 4;70(1):1-31.

[19] Abdi H, Valentin D. Multiple correspondence analysis. Encyclopedia of measurement and statistics. 2007;2(4):651-7.

[20] McGuire TG, Miranda J. New evidence regarding racial and ethnic disparities in mental health: Policy implications. Health Affairs. 2008 Mar;27(2):393-403.

[21] US Department of Health and Human Services. Chapter 2: Culture counts: The influence of culture and society on mental health. Mental health: Culture, race, and ethnicity—A supplement to mental health: A report of the surgeon general. Rockville, MD: US Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Mental Health Services. 2001.

[22] Ng CH. The stigma of mental illness in Asian cultures. Australian & New Zealand Journal of Psychiatry. 1997 Jun;31(3):382-90.

[23] Sue S, Morishima JK. The mental health of Asian Americans: Contemporary issues in identifying and treating mental problems. Jossey-Bass; 1982 Oct 13.

[24] Wahl OF, Harman CR. Family views of stigma. Schizophrenia Bulletin. 1989 Jan 1;15(1):131-9.