**Electronic Supplementary File-1**

Schizophrenia phenomenology revisited: positive and negative symptoms are strongly related reflective manifestations of an underlying single trait indicating overall severity of schizophrenia.

Abbas F. Almulla a, Hussein Kadhem Al-Hakeim b, Michael Maes\* c, d, e

a Medical Laboratory Technology Department, College of Medical Technology, The Islamic University, Najaf, Iraq. E-mail: [abbass.chem.almulla1991@gmail.com](mailto:abbass.chem.almulla1991@gmail.com).

b Department of Chemistry, College of Science, University of Kufa, Iraq. E-mail: [headm2010@yahoo.com](mailto:headm2010@yahoo.com).

c\* Department of Psychiatry, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand;

d Department of Psychiatry, Medical University of Plovdiv, Plovdiv, Bulgaria;

e IMPACT Strategic Research Centre, Deakin University, PO Box 281, Geelong, VIC, 3220, Australia.

**Methods: Machine Learning Techniques**

Single joint PCA performed on the 8 symptom domains in MNP and controls was used to visualize the distribution of both groups in a 2D space (the PC plot) whereby MNP patients and controls are differentiated by marker colors and shapes (Unscrambler, CAMO, 2019). We employed a standard deviation weighting process and a 20-fold cross-validation scheme, while outlier limits are based on 0.05% F-residuals and Hotelling’s T2. Correlation loadings for each of symptom domains are displayed in a plot that comprises two ellipses, the outer ellipse indicating 100% explained variance and the inner one explaining 50% explained variance.

We used exploratory factor analysis (EFA) as a data-driven method to explore the factor structure of schizophrenia phenomenology in patients with MNP and all subjects combined (MNP and controls). EFA was performed using FACTOR, windows version 10.5.03 (Ferrando, 2013, 2017) to examine the factor structure of the dataset. Factors were extracted using the robust unweighted least squares (RULS) method and the number of bootstrap samples was 500 (Ferrando and Lorenzo-Seca, 2013; 2017; Lloret et al., 2017). The dispersion matrix was based on Pearson’s correlations, and a robust analysis was carried out with bias-corrected and accelerated (BCa) bootstraps. Before performing EFA, the Kaiser-Meyer-Olkin (KMO) test and Bartlett’s test of sphericity were applied to determine the matrix’ adequacy for factorization. Schwartz’s Bayesian Information Criterion (BIC), the Hull test and Parallel Analysis (Optimal Implementation) were employed as dimensionality tests and to estimate the number of factors to be retained. Model fit indices were computed in order to examine the goodness-of-fit of the model, namely the goodness-of-fit index (GFI) and the adjusted goodness-of-fit index (AGFI). The distribution of residuals is assessed with the root mean square of residuals (RMSR) with an expected mean value of RMSR for an acceptable model (Kelley’s criterion) and the weighted Root Mean Square Residual (WRMR), whereby values <1.0 represent a good fit. Closeness to unidimensionality was checked employing unidimensional congruence (UNICO), explained common variance (ECV) and mean of item residual absolute loadings (MIREAL). The data should be treated as essentially unidimensional when UNICO >0.95; ECV >0.85; and MIREAL <0.300. Construct replicability was assessed using the H index (values between 0 and 1) whereby values ≥0.80 indicate good replicability of the latent vector and stability across studies. The factor determinacy index (FDI) was used to estimate the quality of factor score estimates and values >0.80 are adequate.

If EFA suggested a unidimensional structure (one latent vector) underlying the 8 symptoms domains we planned to perform Partial Least Squares (PLS) analysis. PLS path analysis using PLS-structural equation modeling algorithms (SmartPLS) (Ringle et al., 2018) was employed in order to examine a) the contribution of the symptom domains to the latent vector extracted from all symptoms domains using a hierarchical component model (HCM) (reflective – reflective model) build using the repeated indicator approach (Garson, 2019), b) the convergent validity and reliability of the main construct, c) associations of the LV with known predictors including age and the MMSE (Kanchanatwan et al., 2018c), d) association between the main LV and a general index of severity comprising the BPRS (shBPRS), PANSS general (shPANSSg) and HAM-D (shHAM-D) scores (without all item used in constructing PHEM, FTD and PMR indices) and total HAM-A scores. The input variables were 2 single indicators (age and MMSE score) predicting the latent vector (reflective mode) extracted from the 8 symptom domains. The eight symptom domains are grouped into negative and PHEM symptoms considering reflective and formative models. We performed PLS analysis when the model fit and constructs complied with quality criteria including standardized root mean residual (SRMR) < 0.08; and adequate internal consistency reliability (construct validity) and convergent validity as indicated by composite reliability > 0.800; Cronbach alpha > 0.750; rho\_A >0.800 and average variance extracted (AVE) > 0.500. Indicators are only included in the LV when the factor loadings are > 0.650 with p < 0.01. Discriminant validity is examined using the Fornell-Larcker criterion and the Heterotrait–Monotrait (HTMT) ratio which should be <0.9 (and more conservative < 0.85). Subsequently, we performed complete consistent PLS bootstrapping (5000 bootstraps) and computed t-values and loadings on the LVs for the outer model, and path coefficients with exact p-values for the inner model. We also performed Confirmatory Tedrad Analysis (CTA) to check possible misspecification of our LV model, namely whether the LV is reflective (our hypothesis) or formative. We also performed blindfolding to examine predictive validity using construct cross-validated communality (Q2 statistic) whereby values of Q2 > 0 indicate that the model has predictive relevance and values > 0.35 indicate a large relevance.

Soft independent modeling of class analogy (SIMCA) is a supervised machine learning method which builds separate PCA models for all classes, thus one model for MNP and another model for controls (CAMO, 2018). A training set (comprising 50% of the MNP subjects and 50% of the controls) is used to construct the PCA models and a test (validation) set (the remaining 50%) is used to validate the models. The number of PCs used to build the models in the training set is determined by cross-validation after outliers are deleted as detected by sample residual vs samples and Hotelling’s T2 vs samples plots. Subjects from the test set are then classified into the group for which they display the best similarity based on critical limits for two relevant distances, namely Si, that is the subject to model distance (reflecting how far the subject is located from the target class) and Hi that is the leverage of one subject to the model center (reflecting how different the subject is from the other subjects). The test subjects are consequently projected into both PCA models whereby SIMCA allocates cases to the models by comparing the computed distances to the model subspaces at alpha=0.05. As such, subjects may be assigned to the target model (MNP class members) or the control model (alternative class members) or they can be allocated to both models (hybrids) or to none of the models (outsiders). Healthy controls that intrude into the MNP target class are identified as “aliens”. In this study we used a) the model-to-model distance indicating the degree of separation between both models with a distance > 3 indicating a good separation; b) the discrimination power plot showing the contribution of all features (the symptom domains) separating both models; and c) The Si/S0 (relative distance of the subjects to the class model) vs Hi plot with critical model membership limits allowing to classify cases into the target class (authentication), alternate class members, outsiders, hybrids or aliens.