**Short Proposal**

**Title:** Exploring Associations Between Affective Polarity Extracted from Life Chart Text by Automated Analysis and Mood/Anxiety Symptom-based Self Reports in T1000 Database

**Data extraction using natural language processing**: Self-report questionnaires only capture some elements of clinical presentation in mental health disorders, whereas there are a lot more incremental benefits in incorporating narrative notes obtained from structured interviews such as life chart [[1](#_ENREF_1)]. The “life chart” is a summary of events in patient’s life in different periods of time, 0-5, 5-10, 11-14, 15-18, 18-25, 25-35, and then every ten years, and is written by an interviewer based on patient’s narratives. McCoy et al. [[2](#_ENREF_2)] have demonstrated a natural language processing method that aggregates words conveying positive or negative emotion (i.e., valence) in a patient’s narrative discharge summaries to predict all-cause mortality and hospital readmission. Likewise, we want to use this natural language processing tool to capture information from life chart summaries. In particular, we will use the Pattern[[3](#_ENREF_3)], an open source implementation of lexical opinion mining. In brief, this method finds every matched term and phrases in lexicon including around 3,000 annotated words for polarity (continuous value from-1 to 1), and subjectivity (subjective 0, not subjective1). For each narrative note, we compute three scores, subjective positivity, subjective negativity, and mean of all recognized adjective words after multiplying inverted polarity by the intensity modifier (1/2-2x), for each narrative note. We consider all periods in the life chart to calculate the scores. We will call these three scores as two affected scores, and one adjective score. There are \*\* number of different interviewers that they wrote patients’ life charts. We will determine the author of the life chart and use it as a covariate.

**Further statistical analysis**: We start analysis by identifying the populations from T-1000 to extract mood/anxiety group and healthy comparison subjects group. We will determine the distributional characteristics of our measured scores for the populations. We explore association between means and standard deviation for the dependent measured scores together and separately. We use a correlation coefficient to determine the strength of the linear relationship between dependent measured scores and symptom based self-reports (PANAS-X, BIF, Symptom scales (PHQ-9, OASIS), PROMIS (anxiety, depression, anger)) and we will follow up characteristics of PROMIS outcome measures periodically. Logistic regression will be used to ascertain the effects of age, gender, BMI, socioeconomic status, and education as confounders in the populations. Also, we will test the mediating factors (CTQ, TEC) influences to dependent measured scores. In addition to that in each part we examine a multivariable model with inclusion of all symptoms. In addition to multivariable model analysis, we combine symptoms using principle component analysis (PCA), which allows us to take into account the inter-relatedness of symptoms. Consequently, to test how measured scores are associated with self-reported symptoms, a one-way multivariate analysis of variance (MANOVA) will be used to test overall differences. Moreover, we will use LASSO and random forest algorithm in order to test module features with all clinical data plus measured scores. We will also use glmnet cross validation to perform classification with and without adjusting some features.

Suggested by: Saeid Parvandeh, PhD Student at University of Tulsa

Estimated Time to Implement Codes, Run the Analysis and Provide Reports: 50 hours

Reference:

1. McCoy, T.H., Jr., et al., *Improving Prediction of Suicide and Accidental Death After Discharge From General Hospitals With Natural Language Processing.* JAMA Psychiatry, 2016. **73**(10): p. 1064-1071.

2. McCoy, T.H., et al., *Sentiment Measured in Hospital Discharge Notes Is Associated with Readmission and Mortality Risk: An Electronic Health Record Study.* PLoS One, 2015. **10**(8): p. e0136341.

3. Smedt, T.D. and W. Daelemans, *Pattern for python.* J. Mach. Learn. Res., 2012. **13**(1): p. 2063-2067.

4. Knudsen, K.B., et al., *Associations between adherence, depressive symptoms and health-related quality of life in young adults with cystic fibrosis.* SpringerPlus, 2016. **5**(1): p. 1216.

5. Uzuner, O., et al., *Identifying patient smoking status from medical discharge records.* J Am Med Inform Assoc, 2008. **15**(1): p. 14-24.

6. Zhou, L., et al., *Identifying Patients with Depression Using Free-text Clinical Documents.* Stud Health Technol Inform, 2015. **216**: p. 629-33.

7. Carrell, D.S., et al., *Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence.* Am J Epidemiol, 2014. **179**(6): p. 749-58.

8. Sohn, S., et al., *Drug side effect extraction from clinical narratives of psychiatry and psychology patients.* J Am Med Inform Assoc, 2011. **18 Suppl 1**: p. i144-9.

9. Savova, G.K., et al., *Discovering peripheral arterial disease cases from radiology notes using natural language processing.* AMIA Annu Symp Proc, 2010. **2010**: p. 722-6.

This is a good start, I would rephrase this slightly differently:

Your main dependent measures are:

1. subjective positivity,
2. subjective negativity,
3. mean of all recognized adjective words

We need to determine which source of text we will use.

Need to determine who the author of the note is and use it as a covariate.

In terms of the analyses I would suggest the following:

1. Identify the populations from the T-1000 you want to study:
   1. Mood and Anxiety Problem Group
   2. Healthy Comparison Subjects
2. Determine the distributional characteristics of your main dependent measures
3. Compare the groups (mood / anx versus comparison group) in terms of the dependent measures
4. Determine the degree to which socio-demographic characteristics affect the dependent measures:
   1. Age
   2. Gender
   3. Income
   4. Education
5. Determine the relationship with respect to the following trait variables:
   1. PANAS-X subscales
   2. BIF
   3. Symptom scales
      1. PHQ-9
      2. OASIS
   4. PROMIS (anxiety, depression, anger)
6. Determine mediating influences:
   1. Childhood Trauma
   2. Life Events
7. Follow up characteristics:
   1. PROMIS outcome measures