DATA COMPLETENESS AND PRELIMINARY ANALYSIS FOR SOMATIC SYMPTOM DISORDER (SSD) RESEARCH

Report Date: February 28, 2025

Ryhan Suny

1. IN THIS DOCUMENT,

I lay out most of my data analytics findings with the goal of checking whether the extracted CPCSSN dataset is “good enough” or rather complete enough to run causal/ML algorithms for at least experiment 1 out of the 5 I planned in a separate document. The jupyter notebook (python) analysis turned out to be 145+ pages in pdf and therefore required me to produce this summary report in writing and in tabular format for better organization.

The primary objective of this analysis is to evaluate the completeness of the CPCSSN Care4Mind Dataset (February 2025 extraction), align EHR data with DSM-5 criteria for Somatic Symptom Disorder (SSD), and identify refined patient cohorts for in-depth analysis. The aim is to ensure that apparent somatic symptom presentations are not primarily explained by legitimate medical necessity or comorbidity.

Highlights and Power:

> *After refining our cohort to exclude patients with high medical necessity and serious conditions, approximately **7,910 patients (2.7% of the total)** fully meet the combined DSM-5 criteria for SSD—indicating a robust sample for our causal mediation study. This subgroup is derived from 56,035 patients with multi-system somatic symptoms (15.91%), 44,208 with persistent anxiety (12.55%), and 49,081 with excessive utilization (13.94%), providing **sufficient power** for the analysis. While structural equation modeling literature occasionally recommends 3,000–5,000 participants for complex models, Fritz and MacKinnon (2007) demonstrate that for single-mediator designs, a sample size of 462 achieves 80% power to detect small effects (standardized path coefficients around 0.14). Our sample size far exceeds this threshold, even under conservative assumptions.*

2. DATA COMPLETENESS CHECKS

2.1 Lab Results Frequency Analysis

Metric	Count	Percentage	Notes / Justification
Total lab records	8,528,807	100.00%	Direct count from **Lab_prepared.csv**.
Numeric **TestResult_calc**	4,304,804	50.47%	Converted using `pd.to_numeric()`.
Non-numeric or missing **TestResult_calc**	4,224,003	49.53%	Non-convertible or missing entries.
Complete normal range data (Upper & Lower Normal)	1,219,701	14.30%	Only rows with both `UpperNormal` and `LowerNormal` populated.
Top lab tests in **Name_calc**	–	–	TOTAL CHOLESTEROL, HDL, LDL, TRIGLYCERIDES, FASTING GLUCOSE.

About **14%** of lab rows have upper/lower normal limits, enabling automatic “normal vs. abnormal” classification. A personalized **12-month window** was defined per patient (index date = most recent lab), within which **37.60%** of those with valid range data had ≥3 normal labs, suggesting “excessive” testing for some.

2.2 Anxiety-Related Prescription Analysis

Metric	Value	Notes / Justification
Medication table dimensions	7,706,628 rows, 36 columns	Loaded from **Medication_prepared.csv**.
Missing `Name_calc` entries	315,680	4.10% of all rows lack a medication name.
Total anxiety medications prescribed	949,107	12.3% of all prescriptions (SSRIs, SNRIs, benzodiazepines, other anxiolytics, etc.).

Patients prescribed anxiety meds	122,486 of 292,050	41.9% of those with any medication record.
Persistent anxiety medication use (≥6 mo)	77,896	26.7% of total medication patients.
Patients meeting DSM-5 B2 and B3*	7,910	2.7% of the total population (persistent anxiety + excessive test usage).

“Persistent health anxiety” for Criterion B2 is operationalized as ≥6 months of continuous medication.

2.3 Somatic Symptom Patterns Analysis (Criterion A)

Metric	Count	Percentage
Encounter diagnosis records	12,471,764	100.00%
Somatic symptom diagnoses	789,181	6.33% of all diagnoses
Patients with ≥1 somatic symptom diagnosis	197,154	55.98% of 352,161
Multiple body systems (≥2 somatic systems)	93,176	26.46% of 352,161
Patients after refinement**	56,035	15.91% of 352,161

After excluding cases with high medical necessity (e.g., advanced comorbidity), ~15.91% remain as potential Criterion A.

2.4 Healthcare Utilization & Doctor Shopping Analysis (Criterion B3)

Utilization Measure	Count	Percentage	Definition / Threshold
Total encounters	11,577,739	100.00%	All rows in **Encounter_prepared.csv**.

High utilizers (≥95th percentile of total)	17,612	5.00%	≥100 total encounters.
Average ≥2 visits/month	51,676	14.68%	Calculated from monthly grouping in prior 12 months.
Doctor shoppers (≥90th percentile, ≥5 providers)	47,646	13.54%	Based on distinct `Provider_ID`.
≥2 B3 indicators	75,362	22.41%	E.g., doctor shopping + high visits, or ED usage + repeated complaint.
Strict B3 (≥3 indicators)	8,877	2.52%	Must have three different B3 flags simultaneously.

A **12-month anchor** around each patient's final encounter date was used to count encounters and providers. About 2.52% reached three distinct B3 indicators.

2.5 Comorbidity & Medical Necessity Analysis

Comorbidity	Count	Percentage
Hypertension	69,301	19.68%
Cancer	53,180	14.97%
Asthma/COPD	43,330	12.31%
Multiple chronic conditions (≥2)	76,072	21.60%
High Charlson Index (≥3)	14,415	4.09%

A **Charlson Comorbidity Index** was calculated based on ICD-9/ICD-10 mappings. An additional **Medical Necessity Score** was computed by combining the Charlson Index with other chronic conditions, using weighted sums. A simple linear regression model predicted expected encounters:

$$\hat{y} = \beta_0 + \beta_1 \cdot M$$

where MMM is the Medical Necessity Score. The **Utilization Ratio** is defined as:

$$Utilization\ Ratio\ = \frac{Actual\ Encounters}{\hat{y}+1}$$

Patients above the 90th percentile of the ratio (approximately 2.02) were flagged as “excessive users.” **High ratio patients:** 35,613 (10.12%).

---

3. DSM-5 ALIGNMENT: REFINEMENT & RESULTS

Criterion	Before Refinement	After Refinement	Rationale
A: Somatic symptoms (≥2 systems)	93,176 (26.46%)	56,035 (15.91%)	Excluded serious conditions (cancer, Charlson ≥3).
B2: Persistent anxiety ≥6 mo	77,896 (22.12%)	44,208 (12.55%)	Removed those with legitimate medical reasons for anxiety.
B3: Excessive utilization	75,362 (21.40%)	49,081 (13.94%)	Excluded medically justified high usage (normal ratio).

Refined cohorts are notably smaller once legitimate comorbidity-driven utilization is excluded.

---

4. POSSIBLE NEXT STEPS

Possible Steps	Description
Establish domain-specific lab cutoffs	Recapture more than 14% of labs using recognized cutpoints for HbA1c, TSH, etc.
Use derived metrics in causal mediation	Investigate how “excessive lab testing” mediates the link between health anxiety and health outcomes.
Consult clinical experts	Confirm that “excessive” labs or provider visits are genuinely unwarranted and refine ≥5 providers threshold as needed.
Explore nonlinear or Advanced ML models	A more flexible approach to predicting expected encounters may better capture real-world utilization than a simple linear regression.

5. METHODOLOGICAL SUMMARY

Step	Method
Data Completeness Checks	Assessed missingness in numeric columns and coverage of normal ranges using pandas.
Criterion A (Somatic Symptoms)	Grouped ICD codes into body systems; defined persistence as ≥180 days between first and last encounter.
Criterion B2 (Persistent Anxiety)	Identified anxiety medications (SSRIs, SNRIs, benzodiazepines, etc.) and measured duration (≥6 months).
Criterion B3 (Excessive Utilization)	Calculated encounter percentiles and flagged high utilizers, doctor shoppers, and repeated visits.
Comorbidity & Medical Necessity	Computed the Charlson Comorbidity Index; built a Medical Necessity Score and derived a Utilization Ratio.
Regression Modeling	Fitted a simple linear regression to predict expected encounters from the Medical Necessity Score.
12-Month Windows	Applied a retrospective 12-month window for both labs and encounters using each patient's most recent date.

6. SUMMARY

Data completeness strongly affects SSD research reliability. Only ~14% of lab rows readily provide normal/abnormal flags, and a significant fraction of high encounter counts appear justified by serious conditions. By excluding medical-necessity-driven utilizers, we narrow potential SSD cohorts. This refined alignment with DSM-5 criteria forms a solid foundation for further causal modeling, prospective validation, and eventual clinical integration.

APPENDIX: DETAILED METHODOLOGICAL EXPLANATIONS

The following sections provide **more thorough detail** on key analytical steps and short code snippets that were part of the original notebook.

A1. Lab Classification: Normal vs. Abnormal

1. **Parsing Normal Ranges**
   - We used columns `LowerNormal` and `UpperNormal` when **both** were available (~14% of lab rows).
   - A row's **is_normal** was set to `True` if

     $$LowerNormal \leq TestResult\_calc \leq UpperNormal$$

2. **Personalized 12-Month Window**
   - For each patient, define an **index date = most recent lab date**.
   - Include only labs within 365 days **before** that index date.
   - Count how many are flagged **is_normal**.
   - "Excessive testing" typically required **≥3** normal labs in that personalized window.

A2. Anxiety Medication: Persistent Use & B2

1. **Defining Anxiety Med Classes**
   - SSRIs (e.g., SERTRALINE, ESCITALOPRAM)
   - SNRIs (VENLAFAXINE, DULOXETINE)
   - Benzodiazepines (DIAZEPAM, LORAZEPAM), etc.
2. **Medication Duration**
   - We used columns like `StartDate` and `StopDate` to compute "days on medication." $duration\_days \geq 180\ days$
   - **Persistent** anxiety medication use = **≥180 days**.
3. **B2 Intersection**
   - This set of patients was matched with those having ≥3 normal labs (excessive tests), forming a subset that meets both **B2** and **B3**.

A3. Somatic Symptom Detection (Criterion A)

1. **ICD-Based Body Systems**
    - For each row in `EncounterDiagnosis`, we matched `DiagnosisCode_calc` to regular expressions for general, GI, GU, musculoskeletal, etc.
    - A new column `is_somatic_symptom` flagged if any body-system symptom was found.
2. **Temporal Persistence**
    - For each patient's somatic-symptom-coded encounters, we checked the earliest vs. latest **EncounterDate**.
    - If $latest - earliest \geq 180\ days$, the patient had "persistent" symptoms.

A4. Doctor Shopping & B3 Indicators

1. **Encounter Window**
    - Similar to labs, we anchored each patient to the final `EncounterDate` and looked at prior 12 months.
2. **High Utilization**
    - We computed each patient's total encounters in that window.
    - 90th or 95th percentile cutoffs flagged "high utilizers."
3. **Provider Diversity** ("Doctor Shopping")
    - `provider_count` = the number of distinct `Provider_ID` per patient.
    - 90th percentile (≥5 providers) was used as the threshold.
4. **Strict B3**
    - We enumerated how many indicators each patient had (e.g., *doctor_shopping=1, repeated_symptom_visits=1, frequent_ED_use=1*, etc.)
    - B3 strict required ≥3 different indicators.

A5. Comorbidity: Charlson Index & Medical Necessity

1. **Charlson Mapping**
    - ICD-9/10 patterns (e.g., `^250` for diabetes, `^I50` for heart failure) were used.
    - Each patient got a "charlson_index" equal to the sum of relevant condition weights.
2. **Medical Necessity Score**
    - Additional chronic conditions (e.g., cancer, autoimmune) contributed partial weights.
    - The sum gave a "medical_necessity_score."

      $$Medical\ Necessity\ Score = \sum_{i=1}^{n} \omega_i \cdot C_i$$

      where C_i is an indicator for the i'th condition and ω_i its weight.

3. **Linear Model for Expected Encounters**

    - Expected encounters were predicted using:

$$\widehat{y} = \beta_0 + \beta_1 \cdot M$$

    - We fit `encounter_count` vs. `medical_necessity_score` in a simple linear regression:

```
model = LinearRegression()
X = combined_df[['medical_necessity_score']].values
y = combined_df['encounter_count'].values
model.fit(X, y)
combined_df['expected_encounters'] = model.predict(X)
```

4. **Utilization Ratio**
    - The Utilization Ratio was then computed as:

$$Utilization\ Ratio = \frac{Actual\ Encounters}{\widehat{y}+1}$$

    Patients above the 90th percentile (~2.02) are flagged as “excessive beyond comorbid justification.”

A6. Refined Exclusions & Final SSD Cohort

1. **Exclude Serious Medical Conditions**
    - Cancer, advanced heart disease, high Charlson (≥3).
    - Anxiety explained by major illnesses was also removed.
2. **Exclude Those with Normal Utilization Ratio**
    - We wanted only those whose usage was “excessive” after controlling for comorbidity.
3. **Intersection of A + B2 + B3**
    - Recompute final subsets with medical necessity removed.
    - This refined approach drastically shrinks the cohorts.
4. **Unexplained Symptoms**
    - In some expansions, we also tested how many ICD-coded symptoms had no corresponding condition to explain them (e.g., GI symptoms in someone with no GI disease).
    - “Unexplained” or “symptom-condition mismatch” could further refine Tier 1 or Tier 2 SSD.

A7. Example Code Snippet for Intersection of Criteria

```
criterion_a_patients = set(... )  # A: persistent multi-system
criterion_b2_patients = set(... ) # B2: persistent anxiety
criterion_b3_patients = set(... ) # B3: excessive usage
full_dsm5_pattern =
criterion_a_patients.intersection(criterion_b2_patients).intersection(cr
iterion_b3_patients)
```

Then we removed any with `has_serious_condition == 1` or `high_utilization_ratio == 0` if that disqualified them.

Note: This Appendix summarizes additional logic and short code segments that guided the thorough analysis. Actual scripts, including data-loading commands and intermediate data merges, appear in the original Jupyter notebook.

References:

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science, 18*(3), 233–239.

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods, 16*(2), 93–115.