Data and Code for
"Mis(sed) Diagnosis: Physician Decision Making and ADHD"

Kelli Marquardt
Federal Reserve Bank of Chicago
The University of Arizona

1. Overview

This replication package contains the code and data needed to reproduce the results in the paper **"Mis(sed) Diagnosis: Physician Decision Making and ADHD"** by Kelli Marquardt.

Due to the confidential and private nature of the primary data used in the analysis, the replication package does not include the original electronic health record derived dataset. Instead, it provides simulated datasets generated from aggregate statistics along with a structured prompt for generating synthetic clinical note text using a large language model. See README Section 5 for additional details.

The replication package is organized into three main directories:

- **`data/`**: Contains publicly available datasets, simulated datasets, and an `intermediate/` subfolder where intermediate outputs are saved during execution.
- **`code/`**: Contains all scripts required to reproduce the tables and figures in the paper.
- **`output/`**: Contains empty subfolders (`tables/, figures/, logs/`) where generated outputs will be saved.

In addition, the root directory includes the following:

- Three R scripts (`sim_patdat_fake.R, sim_docdat_fake.R, sim_notedat_fake.R`) used to generate the simulated datasets. See README Section 5.
- One text file (`prompt_generate_notes.txt`) also used to generate the simulated datasets. See README Section 5.
- One excel worksheet (`Fig5_Worksheet.xlsx`) required to generate Figure 5. See README Section 4.
- One Codebook (`Codebook.xlsx`) which describes the variables in the simulated datasets.

---

2. Data Availability Statement

The primary data used in this study is based on a derivative dataset constructed from **Electronic Health Records (EHRs)** in a limited form according to HIPPA definitions. These data were obtained following IRB Approval and data use agreements with the data provider- The University of Arizona Clinical Data Warehouse (now called the Clinical Data Archive). These data are confidential and cannot be shared publicly. Researchers interested in these data can find additional information (including application

process, purchasing, and institutional requirements) here: https://cb2.arizona.edu/services/health-data-science-services/clinical-data-warehouse.

This replication package includes simulated datasets generated from aggregate statistics as well as a structured prompt used to generate synthetic clinical note text with a large language model. These materials allow users to execute all code in the replication package and reproduce the full set of tables and figures, although numerical results will not exactly match those reported in the paper.

In addition to the confidential dataset, there are two publicly available data sources used in the analysis. These are included in the replication package or can be obtained directly from the original sources, as detailed below.

National Health Interview Survey (NHIS) via IPUMS
There are two extracts from the National Health Interview Survey (NHIS), obtained through IPUMS Health Surveys (https://nhis.ipums.org/nhis/). Access conditions: There is a nominal registration and data request process before IPUMS lets you download the data.

- `data/nhis_00008.dta`: Based on NHIS sample years 2000–2021, extracted and downloaded in November 2022 from IPUMS: https://doi.org/10.18128/D070.V7.1
- `data/nhis_00004.dat` and `data/nhis_00004.xml`: Based on NHIS sample years 2014–2017, extracted and downloaded in February 2025 from IPUMS: https://doi.org/10.18128/D070.V7.4

DSM-5 Diagnostic Criteria for ADHD
Diagnostic Criteria for ADHD are defined in the *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition (DSM-5).
Source link: https://www.ncbi.nlm.nih.gov/books/NBK571021/box/ch4.b35/

- `data/DSM_sx_combined.csv`: contains the diagnostic criteria for each of the ADHD subtypes. Symptom text has been edited to use complete sentences (e.g., beginning with "The patient has/is/often") rather than bullet-point fragments.

2.1 Statement about Rights

- I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.
- I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package.
- As per the JPE policy, the confidential data and respective code will be preserved for a period of five years following the paper publication date. I, the author, will provide reasonable assistance to requests for clarification and replication.

2.2 Datasets

The table below details the datasets required for replication. When provided, the file is located in the `data/` directory.

File Name	Source	Provided
`nhis_00008.dta`	https://nhis.ipums.org/nhis/	Yes
`nhis_00004.xml` and `nhis_00004.dat`	https://nhis.ipums.org/nhis/	Yes
`DSM_sx_combined.csv`	https://www.ncbi.nlm.nih.gov/books/NBK571021/box/ch4.b35/	Yes
`est_dat_fake.csv`	Produced in `sim_patdat_fake.R` (see README Section 5)	Yes
`doc_dat_fake.csv`	Produced in `sim_docdat_fake.R` (see README Section 5)	Yes
`note_dat_fake_base.csv`	Produced in `sim_notedat_fake.R` (see README Section 5)	Yes
`note_dat_fake_gptnotes.csv`	Produced using `prompt_generate_notes.txt` (see README Section 5)	No

---

3. Replication Instructions

This section starts with the software, package, and other requirements needed to run the replication scripts. Followed is a detailed list of the scripts located in the **`code/`** directory, in the order in which they should be executed. Paths are written in a platform-independent way using R's file.path() where applicable. Expected inputs, outputs, and approximate run times are noted.

3.1 Computational Requirements (Hardware, Software, and External Dependencies)

- The code was run on a Lenovo 82T1 laptop with a 12th Gen Intel(R) Core(TM) i7-1280P processor (1800 MHz, 14 cores, 20 logical processors), 32 GB RAM, running Microsoft Windows 11 Pro (64-bit, version 10.0.26200, build 26200).

- **Stata/MP 18.0**
- **R 4.4.0**
    - The code requires the following R libraries (versions in parentheses): **dplyr** (1.1.4), **ggplot2** (3.5.1), **textclean** (0.9.3), **data.table** (1.15.4), **tidytext** (0.4.2), **SnowballC** (0.7.1), **tidyr** (1.3.1), **mlr** (2.19.2), **randomForest** (4.7-1.2), **parallel** (4.4.0), **parallelMap** (1.5.1), **pbapply** (1.7-2), **stringr** (1.5.1), **hunspell** (3.0.6), **ggwordcloud** (0.6.2), **textstem** (0.1.4), **tm** (0.7-16), **MASS** (7.3-60.2), **fastDummies** (1.7.3), **sandwich** (3.1-0), **lmtest** (0.9-40), **cowplot** (1.1.3), **scales** (1.3.0), **ipumsr** (0.9.0), **openNLP** (0.2-7), **NLP** (0.3-2), **rJavaEnv** (0.3.0), and **wordnet** (0.1-18).
    - **Note**: all required packages are installed in `step0_InstallPackages.R`.

- **Other External Dependencies**
    - **WordNet Dictionary**
        - Download WordNet 2.1 from https://wordnet.princeton.edu/download/current-version
        - Note the local installation path (required in `step2a_PrepNLPtokens.R`).

- **GloVe Word Embeddings**
  - Download `glove.6B.zip` from https://nlp.stanford.edu/projects/glove/ under "Download pre-trained word vectors"
  - Unzip into `data/Glove/`.
- **Java Environment for R**
  - Install and configure Java for use with R, following RStudio instructions.
- **Medical Dictionary**
  - Download and extract from https://github.com/glutanimate/hunspell-en-med-glut-workaround.git
  - Save to `data/hunspell-en-med-glut-workaround-master/`.

3.2 Replication Script Order

- **`step0_InstallPackages.R`**

Purpose: Install all required R packages and external dependencies.
Users should follow step by step instructions in the code rather than running the entire script at once. Details on required pre-replication downloads are provided here as well.

- **`step1_PredictQ.R`**

Purpose: Train and apply a machine-learning algorithm to predict Qij for all visits; generate word cloud and validation figures.

- **Inputs**:
  - `data/note_dat_fake_base.csv`
  - `data/note_dat_fake_gptnotes.csv`
- **Outputs**:
  - `data/intermediate/cleaned_appt_notes_fake.csv`
  - `data/intermediate/note_dat_fake_wQ.csv`
  - `output/logs/step1_predictQ_log.txt`
  - `output/figures/fig_c1.png`, `fig_2a.png`, `fig_2b.png`
- **Run time**: < 1 minute (simulated data); ~ 3 hours (real data)

- **`step2a_PrepNLPtokens.R`**

Purpose: Prepare the DSM text and the patient note text for NLP algorithm match.

User input required: Specify the local WordNet path on line 51.

Java / openNLP troubleshooting: openNLP depends on rJava, which requires a working Java installation that R can locate correctly. If openNLP fails to load, the issue is typically Java configuration for rJava.

- **Windows:** Re-run the Java setup in `step0_InstallPackages.R` using rJavaEnv. This replication package was developed and tested using that setup.
- **Linux/macOS:** Ensure JAVA_HOME points to a valid JDK installation.
  - E.g. before loading openNLP, try
    ```
    jvm_path <- list.files(Sys.getenv("JAVA_HOME"), pattern = "libjvm.so",
    ```

```
recursive = TRUE, full.names = TRUE)
dyn.load(jvm_path)
```

- **Inputs**:
  - `data/DSM_sx_combined.csv`
  - `data/intermediate/cleaned_appt_notes_fake.csv`
- **Outputs**:
  - `data/intermediate/DSM_Vectors.csv`
  - `data/intermediate/pat_notes_fake_Vectors.csv`
- **Run time**: ~ 5 minutes (simulated data); ~ 3 hours (real data)

- **`step2b_NLPMatch_tfidf.R`**

Purpose: Compute ADHD match values based on patient note text.

- **Inputs**:
  - `data/intermediate/DSM_Vectors.csv`
  - `data/intermediate/pat_notes_fake_Vectors.csv`
  - `data/intermediate/note_dat_fake_wQ.csv`
- **Outputs**:
  - `data/intermediate/nlp_match_dat_fake.csv`
  - `output/tables/tab_c1.txt`
- **Run time**: < 1 minute (simulated data); ~ 3 hours (real data)

- **`step3_EstimateModel.R`**

Purpose: Estimate Model (stage1 & stage2) and produce related tables/figures

- **Inputs**:
  - `data/est_dat_fake.csv`
- **Outputs**:
  - `data/intermediate/pcp_fe_dat.csv`
  - `data/intermediate/main_est.csv`
  - `output/figures/fig_4.png, fig_a1.png`
  - `output/tables/tab_5.txt, tab_7.txt, tab_8.txt, tab_a2.txt, tab_a4.txt, tab_a6.txt, tab_a7.txt`
- **Run time**: < 1 minute

- **`step4_Bootstrap.R`**

Purpose: Re-estimate main model parameters on bootstrapped samples and produce the main estimate table.

- **Inputs**:
  - `data/est_dat_fake.csv`
  - `data/intermediate/main_est.csv`
- **Outputs**:
  - `output/tables/tab_6.txt`
- **Run time**: ~ 1 hour.
  - To reduce run time, user can choose to decrease the number of bootstrap samples (`num_bs`) on line 32.

- **step5_VaryTau.R**

Purpose: Estimate model with \tau for each ADHD subtype and produce table 9.

- **Inputs**:
  - `data/est_dat_fake.csv`
- **Outputs**:
  - `output/tables/tab_9.txt`
- **Run time**: ~ 3-5 minutes.

- **step6a_OtherTabFigs_pat.R**

Purpose: Produce various tables/figures (using patient data).

- **Inputs**:
  - `data/est_dat_fake.csv`
- **Outputs**:
  - `output/figures/ fig_3.png, fig_c2.png`
  - `output/tables/ tab_2.txt, tab_3.txt, tab_4.txt, tab_a1.txt, tab_a8.txt, tab_c2.txt, tab_c3.txt`
- **Run time**: <1 minute

- **step6b_OtherTabFigs_doc.R**

Purpose: Produce tables based on physician type dataset.

- **Inputs**:
  - `data/doc_dat_fake.csv`
  - `data/intermediate/pcp_fe_dat.csv`
- **Outputs**:
  - `output/tables/tab_a3.txt, tab_b1.txt`
- **Run time**: <1 minute

- **step6c_Estimate_Robust.R**

Purpose: Determine robustness by estimating model on select subset of patients.

- **Inputs**:
  - `data/est_dat_fake.csv`
- **Outputs**:
  - `output/tables/tab_b2.txt`
- **Run time**: ~ 3 minutes

- **step6d_NHIS_SampleCompare.R**

Purpose: Generates NHIS vs. empirical sample comparison (Table B3).

- **Inputs**:
  - `data/est_dat_fake.csv`
  - `data/nhis_00004.dat`
  - `data/nhis_00004.xml`
- **Outputs**:
  - `output/tables/tab_b3.txt`

- **Run time**: <1 minute

- `step6e_NHIS_DxTrends.do (Stata)`

Purpose: Produce ADHD diagnosis rate trends by sex (Figure 1).

- **Inputs**:
  - `data/nhis_00008.dta`
- **Outputs**:
  - `output/figures/fig_1.png`
- **Run time**: <1 minute

---

4. List of Paper Exhibits and Programs

This section maps each table and figure in the paper and appendices to the corresponding program (and line) in the `code/` directory. All output files are saved in the `output/tables/` or `output/figures/` directories. Tables are named `tab_X.txt` and figures are named `fig_X.png`. Appendix tables and figures use prefixes aX, bX, or cX corresponding to Appendices A, B, and C.

Table #	Program	Line #	Notes
1	N/A	N/A	Abbreviated list of DSM-V symptoms by ADHD type. Source: https://www.ncbi.nlm.nih.gov/books/NBK571021/box/ch4.b35/
2	`step6a_OtherTabFigs_pat.R`	418	
3	`step6a_OtherTabFigs_pat.R`	316	
4	`step6a_OtherTabFigs_pat.R`	205	
5	`step3_EstimateModel.R`	381	
6	`step4_Bootstrap.R`	459	
7	`step3_EstimateModel.R`	782	
8	`step3_EstimateModel.R`	1010	
9	`step5_VaryTau.R`	461	
A1	`step6a_OtherTabFigs_pat.R`	534	
A2	`step3_EstimateModel.R`	587	
A3	`step6b_OtherTabFigs_doc.R`	286	
A4	`step3_EstimateModel.R`	351	
A5	N/A	N/A	not based on data input
A6	`step3_EstimateModel.R`	957	
A7	`step3_EstimateModel.R`	886	
A8	`step6a_OtherTabFigs_pat.R`	625	
B1	`step6b_OtherTabFigs_doc.R`	153	
B2	`step6c_Estimate_Robust.R`	456	
B3	`step6d_NHIS_SampleCompare.R`	203	
C1	step2b_NLPMatch_tfidf.R	580	
C2	step6a_OtherTabFigs_pat.R	720	
C3	step6a_OtherTabFigs_pat.R	746	

Figure #	Program	Line #	Notes
1	step6e_NHIS_DxTrends.do	end	
2a	step1_PredictQ.R	649	
2b	step1_PredictQ.R	627	
3	step6a_OtherTabFigs_pat.R	97	
4	step3_EstimateModel.R	426	
5	Fig5_Worksheet.xlsx	N/A	User should input values in cells B3-B6 based on the "at Female estimates" in Table 7. Text boxes and observed disparity line will need to be adjusted manually.
A1	step3_EstimateModel.R	503	
C1	step1_PredictQ.R	539	
C2	step6a_OtherTabFigs_pat.R	681	

5. Simulated Data Generation

Because the original derived EHR based data are confidential, the replication package includes simulated tabular datasets and materials for generating synthetic clinical note text that together facilitate replication. While the key estimation variables are simulated from model-based DGP and actual model parameter estimates (Table 6), control variables and IPCP cost-shifters are randomized. Therefore, numerical results obtained using these simulated datasets will not exactly match those based on the original EHR-derived data.

Simulated Tabular Data

Two scripts generate simulated tabular datasets based on aggregate statistics from various tables/figures in the paper. These simulated datasets replace analysis data derived from the confidential EHRs and are sufficient to run the estimation and exhibit-generation scripts noted above.

- `sim_patdat_fake.R` produces `data/est_dat_fake.csv`
- `sim_docdat_fake.R` produces `data/doc_dat_fake.csv`

Synthetic Clinical Note Text

Due to sensitivity concerns, the replication package does **not** include the generated synthetic clinical notes themselves. Instead, it includes a simulated tabular data base and a structured prompt used to generate the notes as a separate file.

- `sim_notedat_fake.R` produces `data/note_dat_fake_base.csv`
- `prompt_generate_notes.txt` is the structed prompt used to generate synthetic doctor note text.

The prompt, along with `data/note_dat_fake_base.csv`, can be used with a large language model (ChatGPT, version 5.2) to generate synthetic clinical note text suitable for executing the NLP components of the analysis (codes `step1_PredictQ.R` through `step2b_NLPMatch_tfidf.R`). The output from the large language model should be saved as `data/note_dat_fake_gptnotes.csv`

Important Note:

The simulated note base data and subsequent filled note data (`note_dat_fake_base.csv` and `note_dat_fake_gptntoes.csv`) include `pat_id` values from 1 to 100, which do **not** correspond to the `pat_id` values in `est_dat_fake.csv` (1 to 10,000). The note data represent a small illustrative subset included solely to facilitate replication of the NLP components of the analysis (codes `step1_PredictQ.R` through `step2b_NLPMatch_tfidf.R`).

In an application using the original EHR-based data, the outputs generated by `step2b_NLPMatch_tfidf.R` (specifically `Qi, type2_max`, and `xi_*`) would be merged with patient-level analysis data using `pat_id`. Because full-scale linked note data are not included in the replication package, this merge step is omitted. Instead, patient-level values for `Qi, type2_max`, and `xi_*` are generated directly within `sim_patdat_fake.R` to be included for all patients in `est_dat_fake.csv`, thus facilitating replication of the remaining non-NLP based scripts.

---

6. References

American Psychiatric Association (2013a). *Diagnostic and Statistical Manual of Mental Disorders.* Washington, DC, 5 edition.

American Psychiatric Association (2013b). Diagnostic Criteria for ADHD (Exhibit 4.12). In *Substance Use Disorder Treatment for People with Co-Occurring Disorders*, chapter 4: Mental and Substance-Related Disorders. Substance Abuse and Mental Health Services Administration (US), Rockville, MD. Updated 2020. National Center for Biotechnology Information (US). https://www.ncbi.nlm.nih.gov/books/NBK571021/box/ch4.b35/.

Blewett, L. A., Rivera Drew, J. A., Fenelon, A., King, M. L., Williams, K. C. W., Backman, D., Breton, E., Cooper, G., and Richards, S. (2025). IPUMS Health Surveys: National Health Interview Survey [NHIS], Version 7.1 and 7.4. Minneapolis, MN: IPUMS, 2025. https://doi.org/10.18128/D070.V7.1 and https://doi.org/10.18128/D070.V7.4

Glutanimate. (n.d.). hunspell-en-med-glut-workaround. GitHub repository. (Based on work by the Hunspell Project; Robinson, R.; & Rajasekharan, N.) https://github.com/glutanimate/hunspell-en-med-glut-workaround.

Kaplan, J. & Schlegel, B. (2023). fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables. Version 1.7.1. URL: https://github.com/jacobkap/fastDummies, https://jacobkap.github.io/fastDummies/.

Pennington, J., Socher, R., and Manning, C.D. (2014). GloVe: Global Vectors for Word Representations. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532-1543. Accessed via Stanford NLP, https://nlp.stanford.edu/projects/glove/

Princeton University (2010). About WordNet. *Wordnet, Princeton University.* https://wordnet.princeton.edu/

Data: The data were purchased using funds awarded via the University of Arizona Graduate and Professional Student Council Research and Project Grant 2019. Data provided by The University of Arizona Center for Biomedical Informatics & Biostatistics- Department of Biomedical Information Services, Clinical Data Archive (formerly Clinical Data Warehouse).
https://cb2.arizona.edu/services/health-data-science-services/clinical-data-warehouse