Federal Reserve Bank of Chicago

Physician Practice Style for Mental Health Conditions: The Case of ADHD

Kelli Marquardt

August 2021

WP 2022-22

https://doi.org/10.21033/wp-2022-22

Physician Practice Style for Mental Health Conditions: The Case of ADHD

Kelli Marquardt*

Federal Reserve Bank of Chicago†
The University of Arizona
kmarquardt@frbchi.org

August 2021

Abstract

There is a robust literature documenting the importance of physician practice style (e.g., the propensity to perform certain operations) in explaining outcomes related to patients' physical health. Yet little is known about the role of physicians in explaining patients' mental health outcomes. This paper uses novel data on doctor note text together with natural language processing techniques to estimate and document heterogeneity in physician practice style for diagnosing Attention Deficit Hyperactivity Disorder (ADHD). I find significant variation in both *diagnostic intensity* (the mean propensity to diagnose) and *diagnostic compliance* (the weight that physicians place on medical guidelines). Physician characteristics can explain some of this heterogeneity. Specifically, both female physicians and recent graduates have higher diagnostic compliance and lower diagnostic intensity than their respective counterparts. Mental health diagnostic errors lead to excess medical and societal spending. Given the costs of such errors, the findings in this paper encourage a re-evaluation of the mental health identification process, though perhaps targeted at specific sub-groups of physicians.

Keywords: Physician Practice Style, Child Mental Health, Textual Analysis.
JEL classification: I10, J24, C8.

*This paper is based upon work supported by the University of Arizona Graduate and Professional Student Council, Research and Project (ReaP) Grant -2019. Data provided by the University of Arizona Center for Biomedical Informatics & Biostatistics, Department of Biomedical Informatics.

†The views here do not represent those of the Federal Reserve Bank of Chicago or the Federal Reserve System.

1 Introduction

Health care spending and patient outcomes vary dramatically both across and within markets (Kibria et al., 2013). Heterogeneity in physician practice style is a widely documented source of such variation. To reduce inefficiency in health spending, both policy-makers and researchers are interested in estimating physician practices style and understanding why physicians treat patients differently. The existing literature on this topic has typically used insurance claims or hospital discharge data to estimate physician practice style in physical health applications. For example, Epstein and Nicholson (2009) document heterogeneity across obstetricians in their propensity to preform C-sections. Currie et al. (2016) explore variation in the cardiologist's decision to preform invasive procedures for heart-attack patients. Gowrisankaran et al. (2018) examine the correlation of physician practice style across three conditions: Angina, Appendicitis, and Transient Ischemic Attacks.

While the extant literature shows the importance of physician practice style in physical health applications, little is known about the influence of the physician on patient mental health outcomes. In 2019, mental health sector spending was estimated at $225 billion, more than a 50% increase since 2009 (Open Minds, 2020). Early and accurate detection of mental health conditions could help reduce these cost burdens, both on an individual and national level. Therefore, it is important to understand the role of physicians in the mental health diagnostic decision making process. The goal of this paper is to quantify physician practice style as it relates to mental health diagnosis and to document heterogeneity across physicians as a potential source of variation in mental health care spending.

The challenge of quantifying practice style for mental health conditions stems from the process of diagnosis. The presence of a mental health condition cannot be determined via any blood test or medical imaging. Instead, a physician must conduct a behavioral interview and match subjective patient symptoms to those which define a diagnosis, outlined in *The Diagnostic and Statistical Manual of Mental Disorders*, which is currently in its fifth edition (DSM-V). Because these symptoms are expressed to the physician in a more conversational manner, they are not typically denoted in traditional health datasets, and therefore are not observable to the econometrician. In this paper, I propose using a new source of data, namely doctor note text from electronic health records, to overcome the partial observability problem and quantify physician practice style for a specific mental health condition, Attention Deficit Hyperactivity Disorder (ADHD).

ADHD is an ideal application for this study for two reasons. First, ADHD is the most prevalent child mental health condition, being diagnosed in nearly 10% of children worldwide. It is a costly condition, impacting individual children, families, and society. Doshi et al. (2012) estimate the annual economic impact of ADHD diagnosis in the range of $143-$266 billion US dollars. Second, despite the documented large costs associated with ADHD diagnostic errors, recent research suggests that this condition is often inaccurately diagnosed in practice (Merten et al., 2017).[1] Estimating the role that physicians play in the ADHD diagnosis decision can help influence medical and health care policy aimed at reducing ADHD diagnostic errors and their associated costs.

I obtain electronic health record data from a large healthcare system in Arizona, which importantly includes access to de-identified clinical doctor notes for over 12,311 pediatric patients. I use these data to estimate and document heterogeneity in physician practice style for 129 unique physicians. I first present a natural language processing (NLP) algorithm which I apply to clinical doctor note texts and derive how "appropriate" a patient is for an ADHD diagnosis based on symptom match. The NLP algorithm takes the patient record as an input and produces a single output metric, which can be interpreted as how closely the patient's expressed symptoms overlap with the ADHD specific symptoms defined by the DSM. I then use this patient metric as a control in a diagnosis decision-making model and estimate physician specific parameters. The intercept effect, which I call *diagnostic intensity*, measures the mean propensity to diagnose. The slope effect, which I call *diagnostic compliance*, measures how closely the physician follows national diagnostic guidelines when diagnosing a child with ADHD. Together, these two components define the physician practice style, providing a quantitative measure of mental health diagnosis quality.

[1] ADHD misdiagnosis has been both quantitatively and qualitatively examined across many disciplines. Within health economics, scholars have used birthdate to school entry cut-off date as a discontinuity predicting potential over diagnosis of ADHD (e.g., Elder, 2010; Schwandt and Wuppermann, 2016; Persson et al., 2021). Health services and psychology literature explore this topic using meta analyses (Sciutto and Eisenberg, 2007) and physician/patient surveys (Chan et al., 2005; Bruchmüller et al., 2012).

In an ideal setting, physicians would have sufficient time and skill in their behavioral assessments, resulting in high *diagnostic compliance* and low *diagnostic intensity*. In other words, the physician would follow the DSM-V guidelines and diagnose a patient with ADHD if and only if they meet the necessary requirements for diagnosis. However, due to a variety of factors including time and information constraints, physicians do not diagnose in an ideal setting. I find significant variation in physician practice style for ADHD, implying that physicians choose different diagnosis codes even for patients with identical set of behavioral symptoms. The median physician in my sample has a diagnostic intensity of -0.17 and diagnostic compliance of 0.44. The negative intensity estimate suggests that the median physician is, on average, risk-averse when choosing to diagnose their patient with ADHD. The positive (but small) diagnostic compliance estimate suggests that physicians put some weight on official DSM-V guidelines, but also likely rely on prior beliefs and/or other patient signals when making the diagnosis decision.

I then use hand-collected information on physician background to estimate how much of the variation in physician practice style can be explained by physician characteristics and training. I find that physician gender and experience are the strongest predictors of physician practice style for ADHD with both female physicians and recent graduates having higher diagnostic compliance and lower diagnostic intensity than their respective counterparts. Because deviation from medical guidelines can result in diagnostic errors and unnecessary spending, these results suggest a re-evaluation of how mental health conditions are identified with a focus on the heterogeneity in physician guideline adherence.

The rest of the paper is outlined as follows. In the next section, I describe the data and provide background medical information on the ADHD diagnostic process. Section 3 details the methodology needed for mental health practice style estimation. This includes a first stage natural language processing algorithm and second stage diagnosis decision-making model. Section 4 presents the main results and heterogeneity analysis. Finally, Section 5 concludes with a discussion and extensions for future work.

2 Background and Data

Attention Deficit Hyperactivity Disorder is a chronic mental health condition recognized by symptoms of inattention, hyperactivity, and impulsivity. These symptoms are often first recognized in childhood (average age of diagnosis is 7 years old), but can persist into, or even develop in, adulthood (Holland and Riley, 2017). In addition to genetics and a strong heritability component to developing ADHD, the condition can also manifest due to less understood environmental, social, and other structural factors (Hinshaw, 2018).

While medical research is trying to find ways to identify ADHD via genetic testing and/or brain imaging, the current gold standard for ADHD identification is defined by *The Diagnostic and Statistical Manual of Mental Disorders* (DSM). The DSM provides a list of behavioral symptoms that define a particular mental health condition and guidelines for when a patient appropriately meets criteria for clinical diagnosis.

Table 1 presents the DSM-V criteria for ADHD. There are three types of ADHD: inattentive type, hyperactive-impulsive type, and combined type. According to the DSM, a patient has ADHD of a specific type if and only if they experience 6 or more of the defined symptoms in two or more settings (e.g., home and school). In order to be clinically diagnosed with ADHD, a patient must first receive a behavioral assessment from a trained physician. This can be done by a pediatrician, primary care physician, psychologist, or psychiatrist. The physician will conduct an interview with the child (and most times the parent and another caregiver) to determine which symptoms the child experiences in daily life. Physicians may use published ADHD rating-scales along with open-ended questions, but should consult the DSM and document the presence of relevant symptoms in their doctor note (American Academy of Pediatrics, 2011). The physician will then use the information obtained via behavioral assessment to decide whether or not to diagnose the patient with ADHD, ideally following the DSM guidelines, though potentially deviating due to subjectivity of symptom definition or other factors such as appointment time constraints.[2]

[2] Aside from the results documented later in this paper, other literature confirms the finding that physicians vary in their diagnostic compliance. Chan et al. (2005) present results from physician surveys noting differences in how many and which physicians follow national ADHD diagnostic guidelines. Additionally, a list of papers use vignette surveys to show the ADHD diagnosis decision is influenced by non-behavioral

Table 1: DSM-V Symptoms for ADHD

Type I- Inattention
1. Often fails to give close attention to details or makes careless mistakes.
2. Often has difficulty sustaining attention in tasks or play activities.
3. Often does not seem to listen when spoken to directly.
4. Often does not follow through on instructions.
5. Often has difficulty organizing tasks and activities.
6. Often is reluctant to engage in tasks that require sustained mental effort.
7. Often loses things necessary for tasks or activities.
8. Is often easily distracted by extraneous stimuli.
9. Is often forgetful in daily activities.
Type II- Hyperactive/Impulsive
1. Often fidgets with or taps hands or feet or squirms in seat.
2. Often leaves seat in situations when remaining seated is expected.
3. Often runs about or climbs in situations where it is inappropriate.
4. Often unable to play or engage in leisure activities quietly.
5. Is often "on the go," acting as if "driven by a motor."
6. Often talks excessively.
7. Often blurts out an answer before a question has been completed.
8. Often has difficulty waiting his or her turn.
9. Often interrupts or intrudes on others.

Note: This table reflects abbreviated list of DSM-V symptoms by ADHD type. The full version is published in American Psychiatric Association (2013).

The traditional datasets used to estimate physician practice style (e.g., insurance claims or hospital discharge files) do not contain information on behavioral symptoms. Therefore, I choose to utilize a more detailed source of information, namely electronic health records (EHR) from a large healthcare system in Arizona. While the generalizability of the analysis is limited due to the specific location, the micro-level detail of these data is extremely beneficial for this study. In particular, access to the clinical doctor notes allows me to observe which mental health symptoms are discussed with a physician during behavioral assessment. This information is missing from standard health datasets, yet is key for both identification and estimation of physician practice style in mental health applications.

The EHR dataset includes 56,505 distinct pediatric appointments from January 2014 to September 2017. This covers 12,311 patients (ages 5-16) and 218 physicians. I have access to patient identifiers, physician names, associated diagnoses (if any), and de-identified clinical

patient characteristics such as gender, race, and income (Lowe et al., 2007; Morley, 2010; Bruchmüller et al., 2012; Marquardt, 2021).

doctor notes for each patient encounter. Physician reported diagnoses are translated into ICD-10 codes. I label a child as receiving an ADHD diagnosis if he/she has an encounter with an associated ICD-10 code of F90.0, F90.1, or F90.2.

Patient Set

The patient set contains all patients (ages 5-16) that have an encounter with a pediatrician and/or child psychologist during the sample period. Patient summary statistics are presented in Table 2.

Table 2: Patient Summary Statistics

	Mean	Std. Dev.
ADHD Dx.	0.07	0.25
Total Encounters	4.59	4.05
Private Ins.	0.43	0.50
Public Ins.	0.52	0.50
Age	10.24	3.44
Male	0.51	0.50
White	0.29	0.45
Hispanic	0.43	0.50
PEDS Doctor	0.74	0.44
PSYCH Doctor	0.09	0.29
N (patients)	12,311	

Note: Table presents summary statistics for the set of pediatric patients who had an encounter between 2014 and 2017 with a diagnosing physician. ADHD Dx. constructed using icd9 and icd10 codes. Patient age averaged across all encounters in sample period.

Approximately 7% of patients are diagnosed with ADHD during the sample period. This is slightly smaller than the national average during the time period, though not surprising as nearly half of the patients in the sample are Hispanic, and the documented diagnosis rate for this ethnicity is lower than others (Morgan et al., 2013).[3] A majority of the patients have an appointment with a PEDS doctor (indicated in the EHR as "pediatrics" or "family medicine") while only 9% see a PSYCH doctor (indicated in the EHR as "psychiatrist" or "behavioral specialist").

[3]It is suggested that this low diagnosis rate for Hispanic population is driven mainly by patient access to and/or preference towards mental health care, rather than bias in the physician decision making process itself.

Physician Set

The original dataset includes 218 unique physicians. After removing physicians that never diagnose ADHD and/or have less than 30 patients over the sample period, I am left with 129 physicians.

From the electronic health records directly, I note the name of the physician and associated specialty. To obtain additional physician information, I search the following sources: docinfo.org, doctor.webmd, and healthgrades.com. Because I do not have national physician identifiers, I must search these sources with full names only. I ensure the search results correspond to the appropriate physician by matching on full name and location of practice. I am able to obtain information on physician gender, where they attended medical school, and what year they graduated. I determine medical school rank based on US News 2018 rankings.[4] I code a physician as attending a top Medical School if their associated school is ranked in the top 100.

Table 3: Physician Summary Statistics

	Mean	S.D.	Min.	Max.
Female	0.667	0.473	0	1
TopMedSchool	0.388	0.489	0	1
PEDS	0.930	0.256	0	1
PSYCH	0.070	0.256	0	1
Grad. Year	2007	11	1959	2016
Total Patient	194.845	264.236	30	1327
Tot. ADHD Pat.	18.062	20.027	1	98
N(doctor)	129			

Note: Table presents summary statistics for all physicians in health center with >= 30 patients and at least 1 ADHD diagnosis. TopMedSchool based on US News ranking of medical school. Total Patient (Tot. ADHD Pat.) calculated as total number of pediatric patients (with ADHD diagnosis) seen during sample period.

Table 3 presents physician summary statistics based on the above data collection procedure. Approximately 67% of physicians are female. A little over a third of the physicians attended a top ranked medical school, with graduation years ranging from 1959 to 2016. While some mental health conditions can only be diagnosed by psychiatrists, ADHD is often

[4]While none of the doctors actually attended medical school in 2018, the 2018 ranking is likely similar to the ranking of their school in the years enrolled. Schnell and Currie (2018) use the same medical school ranking source and calculate "pairwise correlation coefficient all greater than .96 across annual rankings from 2010-2017".

diagnosed by a primary care physician (Visser et al., 2015). This is consistent in my sample where the vast majority of physicians have a pediatric specialty, with only 7% specializing is psychiatry.

3 Methods

I estimate two dimensions of physician practice style: *diagnosis intensity*, the average propensity to diagnose, and *diagnostic compliance*, the weight that physicians place on national diagnostic guidelines. In theory, identification of this practice style would come from comparing physician diagnosis decisions for a set of patients with the same number of behavioral symptoms. In practice, however, empirical estimation is complicated because there are no obvious health variables in traditional datasets that can be used to control for mental health symptom severity. In econometric terms, patient symptoms are traditionally unobservable to the econometrician. In this section, I show how access to clinical doctor note data can alleviate some of these identification concerns and are an important source of information needed to estimate physician practices style for mental health applications.

I propose a 2-step estimation approach. First, I construct a measure of patient ADHD symptom match using a natural language processing algorithm applied to clinical doctor note text. This measure does not rely on the physician diagnosis decision and thus can be thought of as an unbiased and observable proxy for the “unobserved” symptom measure needed for identification. Second, I estimate physician practice style by comparing physician diagnosis decisions *conditional* on the constructed patient ADHD symptom match from step 1.

3.1 Patient ADHD Symptom Match

In order to quantitatively measure how physicians make decisions in the mental health context, it is first important to determine if a patient is suffering from behavioral symptoms, i.e., measure the patient’s “appropriateness” for diagnosis.

Even detailed electronic health records do not report readily observable patient behavioral symptoms. Instead, this information is found in the clinical doctor note. The existing approach for extracting behavioral symptoms from clinical text is documented and applied in Leroy et al. (2018). The methods involve recruitment of psychiatric experts to hand-label a random set of doctor notes to determine which words/phrases match with DSM-V criteria based on content similarity. I propose an alternative method of extracting behavioral symptoms that does not rely on funding or time needed to recruit experts for hand-labeling. Additionally, my proposed procedure does not rely on physician opinion or the final physician diagnosis decision. Thus, the text processing algorithm produces an unbiased yet noisy proxy for patient symptom match based solely on the official DSM-V text as reference.

According to the Natural Language Processing (NLP) literature, the best method for measuring document similarity with a limited training dataset is a Bag of Words Model (BOW) with cosine similarity measures.[5] However, this traditional model measures similarity based on *word-match* as opposed to *content-match*. Because physicians use natural language during behavioral assessments, it is unlikely that they will document the DSM-V words exactly. I adjust the traditional BOW framework to account for semantic similarity. While the algorithm is presented more generally first, the proceeding section outlines the choices I make for the specific ADHD context of interest. The Appendix provides an overly simplified example to demonstrate each algorithm step and documents external resources used.

3.1.1 NLP Algorithm: General

Let i denote doctor note (or full patient record) and s denote the official DSM-V text. The index s can reference a single symptom (one line from Table 1), a group of symptoms (1-9 of Type I or Type 2 in Table 1), or the entire DSM-V text for a given mental health condition (all of text in Table 1).

[5] The training dataset in this case is 'limited' in the sense that it contains only the DSM-V text for a given mental health condition.

Step 1: Text Cleaning & Pre-Processing

Traditional text is messy. This first step cleans the text and prepares it for mathematical analysis, making sure that words that mean the same thing are represented by the exact same grouping of characters. I break this part into two sub-steps because medical text requires special medical dictionaries for cleaning.

1a: Medical

- Spell check and replace words using a medical dictionary.
- Replace typical medical abbreviations with full meaning.

1b: Traditional

- Fix Contractions (e.g., "doesn't" → "does not")
- Remove Special Characters (e.g., #*%@)
- Lowercase every word
- Replace each word with its *stem* (e.g., "studies"/"studying" →"study")

Step 2: Obtain Word Groupings and Reduce Size

While step 1 ensures that same words are represented by the same characters, step 2 ensures that *similar* words are represented by the same characters. The idea here is to preserve the content and meaning of the text. It is important to note that some words have different meanings depending on the part of speech (e.g., "offer" as a noun ≠ "offer" as a verb). Therefore, this step requires a part of speech (POS) tagging algorithm. This step also mentions some word reduction options which can be implemented without large content loss in order to save computational time/space. This is especially important in text analyses as the number of words (and therefore the BOW matrix dimension) becomes quite large and synonym search is computationally expensive.

- Determine the part of speech (POS) for each word in each document.
- If needed for computational purposes, reduce corpus size:
  - keep only common adjectives, nouns, verbs, and negation words ("not", "non", etc.)
  - remove *stop words* which are common English words like "and", "or", "have"
  - remove words less than 3 characters or greater than 16

- Use WordNet to replace each word with its most common synonym. WordNet is a lexical English database which groups words according to general meaning based on word-POS pair. For example, this step will replace the words "best" and "well" with the word "good", while keeping the word "good" as is. (e.g., "good","best","well" $\rightarrow$ "good").

To further allow for variation in natural language, one can additionally group using "close" words for each DSM-V symptom word using pre-trained word embeddings from GloVe (Global Vectors for Word Embeddings). GloVe is a machine learning algorithm used to classify words as multi-dimensional vectors of real numbers (word embeddings) based on their context in a document. "Closeness" is determined by cosine distance in R^{300} space. As an example, the 10 closest words for the term "inaccurate" are: inaccurate, erroneous, mislead, incorrect, untrue, incomplete, accurate, unreliable, bias, factually.

Step 3: Tokenize

This step requires converting each patient or DSM-V symptom document into a vector of n-grams. Uni-grams ($n = 1$) are single words, and bi-grams ($n = 2$) are adjacent word pairs. Although more computationally expensive, I use at least $n = 2$ to allow for negation. For example, using both uni-grams and bi-grams, the phrase "patient is not sad" becomes the vector [patient, is, not, sad, patient is, is not, not sad]. Including bi-grams ensures that "not sad" does not match with "sad" when measuring document similarity.

Step 4: Build the Adjusted BOW Model Matrix

Each document vector can now be combined to create the BOW Model Matrix. In the natural language processing literature, this matrix is also often referred to as the *Document Term Matrix*. Here, the matrix columns represent unique bi-grams or uni-grams and the rows represent each document. The matrix elements are binary {0,1} indicating if the column bigram/unigram appears in patient document or DSM-V symptom row. Appendix Table A1 provides a visual example.

Step 5: Measuring Content Overlap: x^*_{is} and x^*_i

The final step is to calculate patient symptom overlap using the BOW matrix and to create the control needed for physician practice style estimation.

- The patient document symptom overlap measure (x^*_{is}) is calculated via cosine similarity between the patient document vector i and the DSM symptom vector s from the Adjusted BOW Model Matrix in Step 4. Mathematically, letting $\hat{k_i}$ denote the BOW vector for patient document i, and $\hat{d_s}$ denote the BOW vector for DSM-V symptom text s, I define $x^*_{is} \equiv \frac{\hat{k_i} \cdot \hat{d_s}}{||\hat{k_i}||||\hat{d_s}||}$. This essentially measures the overlap between words in the DSM-V criteria and words in the patient clinical doctor note, adjusting for note length.
- Because diagnosis depends on the *entire* set of DSM-V symptoms for the condition of interest and not just the subset of symptoms denoted by s, it is important to collapse x^*_{is} to x^*_i. The most obvious method would be to take the average (or a weighted average) across all symptoms s for each patient i. However, the most logical collapse process depends on the specific application and how symptom subsets are defined.

It is important to note that the definition of s and choice of aggregation will affect the *levels* of the patient control measure. Regardless of this choice, however, the relative *differences* in x^*_i are informative of symptom match rank. Specifically, higher values of x^*_i denote a closer overlap between patient note and DSM-V diagnostic criteria.

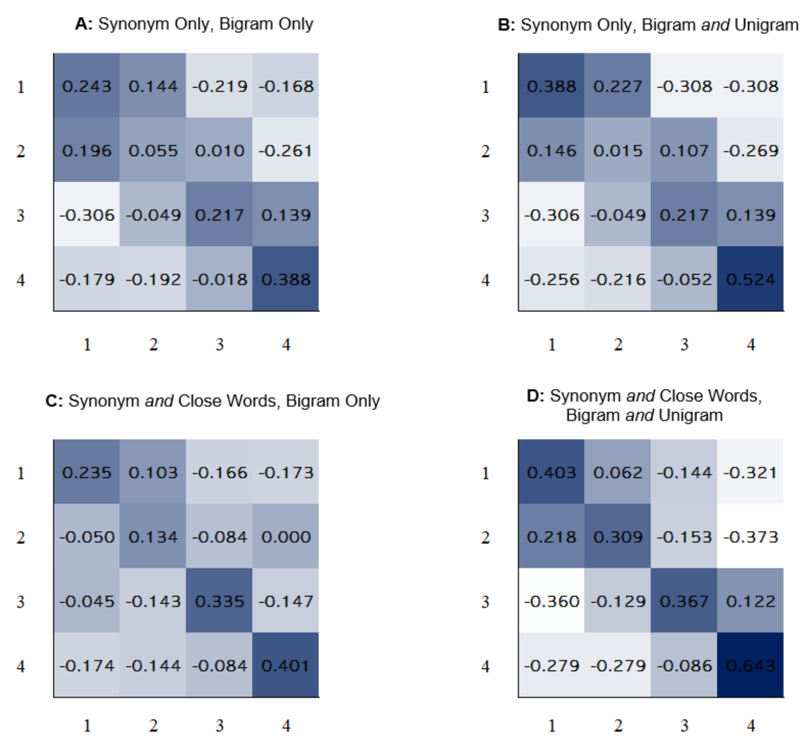
3.1.2 NLP Algorithm: ADHD Application

Due to sample size constraints, I conduct the analysis at the patient level. Therefore, each patient document i is the set of patient notes combined across all patient appointments. I define the DSM-V document s to be the entire DSM-V text for ADHD (i.e., the entire text from Table 1).

I test the validity of the algorithm by comparing the outcome measure to a hand-coded sub-sample of patient records. Specifically, before I estimate practice style for the entire sample, I collect a random subset of doctor note records for 100 patients and read each note carefully, documenting the number of ADHD symptoms that appear, using the official DSM-V symptom list as reference. I then compare the NLP algorithm rank to the hand-coded rank of patient, analyzing four alternative algorithm predictions. The adjustments I consider are to algorithm Step 2 (synonym only vs. synonym *and* 10 close words) and Step 3 (bigrams only vs. unigrams *and* bigrams).

Figure 1 presents the correlation plots, comparing the hand-coded rank to the algorithm predicted rank. Each plot represents the alternative algorithms. Each square within the plot represents the level of overlap between hand-coded quartile (x-axis) and algorithm-predicted quartile (y-axis). High values correspond to higher levels of overlap. Comparing A to B and C to D suggests that the algorithms that include both bigrams *and* unigrams in Step 3 out-perform those using bigrams alone. Additionally comparing B to D suggests the need for both synonym *and* close word match for algorithm Step 2. Therefore, when calculating ADHD symptom match for the entire sample, I use the plot D algorithm which includes grouping by synonyms and close words and tokenizing using both uni-grams and bi-grams.

Figure 1: NLP Algorithm Correlation Plots



Note: Correlation plots from 4 alternative NLP algorithms. Each plot displays overlap of algorithm-determined quartile and hand-coded quartile for a random subset of 100 patient records. High values correspond to high overlap in quartile ranking.

3.2 Estimating Physician Practice Style

After obtaining the proxy for patient ADHD symptom match in stage 1, the second stage estimates practice style for each physician in the sample. The two components of physician practice style (compliance and intensity) can be estimated via the following regression where i denotes patient, j denotes physician, and $D_{ij} \in \{0, 1\}$ indicates whether physician j diagnoses patient i with ADHD.

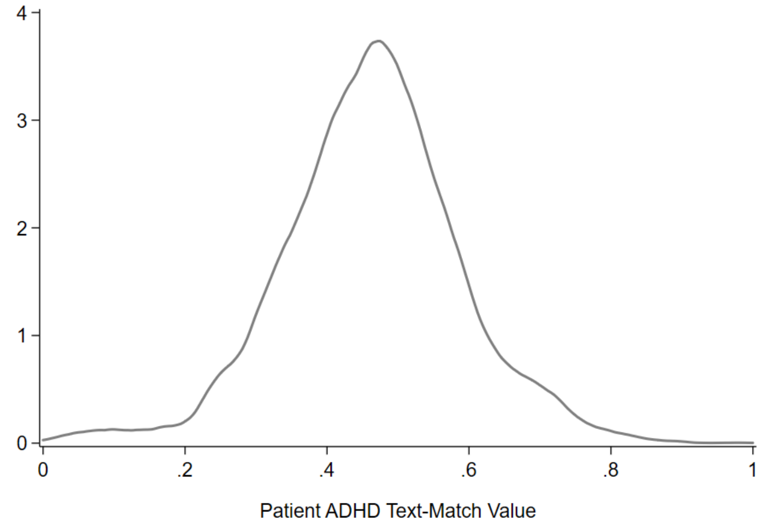
$$D_{ij} = \alpha_j + \beta_j x_i^* + \varepsilon_{ij} \tag{1}$$

Estimating equation (1) determines physician practice style for each physician j: $(\hat{\alpha}_j, \hat{\beta}_j)$. The first component, α_j, is the average propensity to diagnose ADHD. It can be thought of as the physician's *intensity* of diagnosis for any given patient. The second component, β_j, represents the additional weight that physicians place on the patient symptoms when they make their diagnosis decision. Because x_i^* measures overlap with national diagnostic guidelines, this term is appropriately labeled *diagnostic compliance.* I compare the estimates across physicians and interpret relatively high values of compliance (β) and low values of intensity (α) as practice style estimates associated with more accurate ADHD diagnoses on average.

4 Results

It takes approximately 48 hours to run the first stage text analysis procedure outlined above on an individual laptop with 4 processors. Because x_i^* is measured with relativity, I standardize to the range $[0, 1]$. For reference, $x_i^* = 0$ implies patient i had the least overlap with ADHD specific symptoms and $x_i^* = 1$ implies patient i had the most overlap. The ADHD symptom match mean is 0.462 with a standard deviation of 0.125. Figure 2 presents the distribution across the full set of patients.

Figure 2: Density of ADHD Symptom Match x_i^*



Note: Figure presents density of constructed patient symptom-match (x_i^*). High match values correspond to patient records with closest match to DSM-V criteria for ADHD, presented in Table 1.

Table 4 presents the eventual ADHD diagnosis rate for all patients, for those in the bottom 25% of ADHD symptom match rank, and for those in the top 25% of ADHD symptom match rank. Overall, 6.89% of children are diagnosed with ADHD. The diagnosis rate for the lowest symptom ranked patients is 1.11%, and the highest ranked patients are diagnosed at 21.16%. This table provides suggestive evidence that physicians are recognizing when patients present with ADHD specific symptoms and ranking/diagnosing them accordingly. However, the standard deviation of diagnosis is much larger in the high-symptom group compared to low-ranked patients, suggesting significant variation in diagnosis decision across children and likely their physicians.

Table 4: Patient Outcomes by Symptom Match Rank

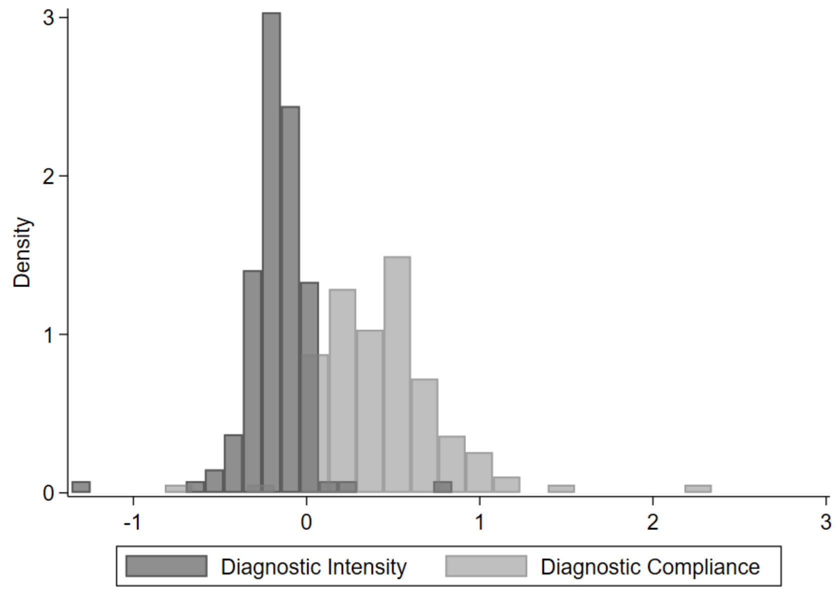
ADHD Symptom Match Rank	Diagnosis Rate	Diagnosis S.D.
ALL	0.069	0.253
LOW	0.011	0.105
HIGH	0.212	0.408

Note: Table presents eventual ADHD diagnosis rate and ADHD diagnosis standard deviation for all patients, patients with low ADHD symptom match rank and patients with high ADHD symptom match rank. LOW includes patients in the bottom 25% of the ADHD symptom match distribution and HIGH includes those in the top 25%.

4.1 Practice Style Estimates

Next, I estimate physician practice style for each of the 129 physicians in the sample according to equation (1). Figure 3 presents the distribution of practice style estimates across physicians.

Figure 3: Practice Style Variation



Note: Figure presents histogram of physician practice style estimates across all physicians, j. Diagnostic Intensity corresponds to $\hat{\alpha_j}$ and Diagnostic Compliance corresponds to $\hat{\beta_j}$ from estimating equation 1.

The physician practice style estimates suggest that the median physician prefers *not* to diagnose ADHD ($\hat{\alpha}_{med} = -0.17$). They place a positive, but less than 1, weight on the national diagnostic guidelines ($\hat{\beta}_{med} = 0.44$) implying that while physicians make some reference to the DSM-V in the diagnostic process, they are also likely to rely on prior beliefs and/or other patient signals when making the official diagnosis decision.

Because ADHD symptom match, x_i^*, is measured with relativity, point interpretation of the estimates are not necessarily straightforward. Rather, one can analyze differences in diagnostic probabilities for the median patient. For example, if the median patient (at 0.465 symptom match value) were to see the median physician, the probability of ADHD diagnosis would be 3.46%.[6] If the median patient were to see a doctor with a diagnostic in-

[6]This value comes from plugging in $\hat{\alpha}_{med} = -0.17$ and $\hat{\beta}_{med} = 0.44$ to equation 1.

tensity one standard deviation above the median, the probability of diagnosis would increase significantly to 22.45%. If the patient were to see a doctor with a diagnostic compliance one standard deviation above the median, the probability of diagnosis would increase to 20.0%.[7]

4.2 Correlates of Practice Style

Why do practice styles differ across physicians? I empirically explore this question by analyzing the correlation between practice style estimates and physician characteristics. In other words, I ask what (if anything) can explain the differences in how physicians choose to diagnose ADHD. Similar analyses have been considered in alternative applications. These include Currie et al. (2016) for heart attacks, Van Parys (2016) for minor injuries, and Chan (2016) for internal medicine resource use. This literature finds that while physician gender, age, and training are correlated with physician practice style, the majority of the variation cannot be explained by such simple demographics, encouraging future explorations in this field.

To conduct this analysis, I estimate the following two equations, using the physician practice style estimates as LHS variables and physician characteristics on the right:

$$\hat{\alpha_j} = \gamma_0 + \gamma_1 Male_j + \gamma_2 TopMedSchool_j + \gamma_3 Specialty_j + \sum_k [\gamma_k experience_{jk}] + \eta_j^a \quad (2)$$

$$\hat{\beta}_j = \delta_0 + \delta_1 Male_j + \delta_2 TopMedSchool_j + \delta_3 Specialty_j + \sum_k [\delta_k experience_{jk}] + \eta_j^b \quad (3)$$

Here, k identifies different bins based on years since medical school graduation (less than 5 years, 5-15 years, 15-25 years, 25-35 years, and more than 35 years). $TopMedSchool_j$ indicates if physician j went to a top ranked medical school, and $Specialty_j$ indicates the physician's specialty: PEDS or PSYCH.

[7] These estimates come from adding diagnostic intensity standard deviation of 0.195 and adding diagnostic compliance standard deviation of 0.358 to the median estimate, respectively.

Tables 5 and 6 present the estimates from equation 2 and equation 3 respectively. Columns (1) and (2) use the full physician sample. Because my sample includes only 9 psychiatric doctors, in columns (3) and (4) I present the results using the subset of pediatric doctors only. This controls for the possible unobserved mental health severity for patients that are referred to or seek out a psychiatric specialist. Importantly, columns (2) and (4) include the patient symptom match value x_i^* from stage 1 to control for the patient mix across physicians. My preferred specification uses the PEDS sample only and includes the control for symptom match, corresponding to column (4) in each table.

Table 5: Correlates of Diagnostic Intensity ($\hat{\alpha_j}$)

	(1)	(2)	(3)	(4)
Male	0.055*	0.057*	0.049	0.051*
	(0.032)	(0.032)	(0.031)	(0.030)
TopMedSchool	-0.052*	-0.050*	-0.031	-0.030
	(0.029)	(0.028)	(0.031)	(0.030)
PSYCH	-0.117	-0.089		
	(0.119)	(0.120)		
5-15 yrs	0.041	0.041	0.046	0.046
	(0.031)	(0.031)	(0.030)	(0.030)
15-25 yrs	0.083**	0.079**	0.080**	0.075**
	(0.032)	(0.031)	(0.031)	(0.030)
25-35 yrs	0.052	0.045	0.041	0.034
	(0.055)	(0.054)	(0.055)	(0.054)
35+ yrs	0.036	0.028	0.029	0.020
	(0.025)	(0.024)	(0.025)	(0.023)
x_i^*		-0.154***		-0.163***
		(0.040)		(0.038)
Sample	ALL	ALL	PEDS	PEDS
Observations	23007	23007	21951	21951

Note: Table presents estimates from equation 2. Experience reference group: 0-5 years. Standard errors in parentheses, clustered at physician level. (* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$)

Table 6: Correlates of Diagnostic Compliance ($\hat{\beta}_j$)

	(1)	(2)	(3)	(4)
Male	-0.121*	-0.125*	-0.126*	-0.130*
	(0.068)	(0.067)	(0.069)	(0.068)
TopMedSchool	0.085	0.082	0.047	0.045
	(0.063)	(0.062)	(0.068)	(0.067)
PSYCH	0.284	0.235		
	(0.234)	(0.237)		
5-15 yrs	-0.061	-0.061	-0.076	-0.076
	(0.067)	(0.067)	(0.066)	(0.066)
15-25 yrs	-0.151**	-0.142**	-0.156**	-0.148**
	(0.065)	(0.064)	(0.065)	(0.064)
25-35 yrs	-0.086	-0.072	-0.042	-0.030
	(0.116)	(0.115)	(0.117)	(0.116)
35+ yrs	0.027	0.041	0.046	0.060
	(0.064)	(0.066)	(0.055)	(0.057)
x_i^*		0.274***		0.270***
		(0.093)		(0.088)
Sample	ALL	ALL	PEDS	PEDS
Observations	23007	23007	21951	21951

Note: Table presents estimates from equation 3. Experience reference group: 0-5 years. Standard errors in parentheses, clustered at physician level. (* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$)

I estimate that male physicians have higher diagnostic intensity and lower diagnostic compliance than female physicians, though this difference is only significant at the 10% level. There is also slight evidence that physicians from top-ranked medical schools have lower diagnostic intensity, but no significant difference in compliance. Physician experience seems to be correlated with practice style, though the magnitude and strength depends on the experience bin. In general, compared to the omitted group (physicians will less than 5 years of experience), more experienced physicians tend to have higher diagnostic intensity and lower diagnostic compliance. Finally, the coefficient on the patient symptom match (x_i^*) is significantly negative in Table 5 but positive in Table 6. This implies that patient selection/mix may influence physician practice style. In other words, physicians who see more ADHD-severe patients on average (as indicated by higher x_i^*) are more likely to have high diagnostic compliance and low diagnostic intensity compared to physicians with less severe patients on average.

The substantial variation in physician practice style indicates that physicians will make different diagnosis decisions even for patients with similar symptoms. This finding suggest that children may be inaccurately diagnosed with ADHD, leading to excess spending both at the individual and societal level.[8] Because ADHD misdiagnoses will decrease with diagnostic compliance and (likely) increase with diagnostic intensity, the results from Tables 5 and 6 can speak to some potential policies aimed at reducing such misdiagnosis concerns and the associated costs to society.

The difference in practice style based on physician gender suggests that male and female physicians may differ in how they learn/interpret best-practice guidelines for mental health conditions. It would be beneficial for medical schools to recognize this and potentially re-visit the behavioral health curriculum with this disparity in mind. Additionally, I show physician experience is a significant predictor of practice style, with recent graduates having higher compliance and lower diagnostic intensity. This could be due to a variety of factors such as medical school education changes, residency reform, learning (or forgetting) over time. While determining the actual mechanism is outside the scope of this paper, one policy suggestion would be to require mid-career physicians to take part in re-education programs for best mental health practices. Finally, the results also show that physicians who see more severe cases on average (measured by the symptom match value x_i^*) have higher compliance and lower intensity for diagnosis. A possible response to this finding would be to encourage physician specialization within the broad pediatric practice.

5 Conclusion

Understanding the role that the physician plays in the mental health diagnostic process is important as it can provide insights into the sources of mental health heterogeneity and excess expenditures. This paper contributes to the extant literature on physician practice style by providing a framework and application to estimate practice style in the context of *mental health*. I utilize novel doctor note records and natural language processing tech-

[8]Marquardt (2021) provides an estimate for the excess costs associated with ADHD diagnostic errors in the range of $60-$117 billion US dollars.

niques to overcome the unobservability problem with patient symptom controls and estimate physician practice style for diagnosing Attention Deficit Hyperactivity Disorder in children.

After determining the patient ADHD symptom match via a novel text processing algorithm, I then estimate two dimensions of practice style for each physician: diagnostic intensity and diagnostic compliance. The former measures the mean propensity to diagnose ADHD, while the latter measures how closely the physician follows medical guidelines during the diagnostic process. I compare these estimates across physicians, finding significant variation in both intensity and compliance, likely resulting in heterogeneous rates of ADHD diagnostic errors across patient sets.

The baseline probability of ADHD diagnosis is 6.9%; however, I show that this rate can nearly triple if seen by a physician with practice style estimates one standard deviation above the median. I further explore how physician attributes are correlated with their practice style, finding that female physicians and recent medical school graduates have higher diagnostic compliance and lower diagnostic intensity than their respective counterparts. Because accurate diagnosis is essential for reducing excess medical and social spending, these results have potential policy implications, encouraging a re-examination of the identification process for mental health conditions taught in both in medical schools and through continuing education requirements.

There is a large list of potential applications and extensions that can make use of the proposed methods in this paper. First, it would be interesting to apply the methods to different mental health conditions. Is physician practice style for one mental health condition associated with their practice style for another? This question is similar to those asked by Gowrisankaran et al. (2018) who show that physician practice style is positively correlated across three different (physical) health conditions. It is also important to explore how physician practice style changes based on patient demographics. Are physicians more/less compliant based on patient gender? This is particularly interesting in the ADHD context where the diagnostic gap between male and female patients is quite large (14.8% vs 6.7%).

Additionally, these methods could be used to explore heterogeneity in mental health outcomes across geographies and over time. As physicians gain more experience with a particular condition, do they become better at following guidelines (higher compliance) or do they tend to forget the requirements learned in medical school and practice with lower compliance? How much of the spatial and/or time variation in diagnosis rates can be attributed to physician practice style for mental health conditions? This paper provides a framework and methodology to answer such questions, and demonstrates that, in the context of diagnosing children with Attention Deficit Hyperactivity Disorder, physician practice style is an important mechanism explaining variation in mental health outcomes. Understanding how this heterogeneity leads to diagnostic errors and subsequent excess medical spending is an important goal for future research.

References

American Academy of Pediatrics (2011). ADHD: clinical practice guideline for the diagnosis, evaluation, and treatment of attention-deficit/hyperactivity disorder in children and adolescents. Subcommittee on Attention-Deficit/Hyperactivity Disorder, Steering Committee on Quality Improvement and Management.

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

Bird, S., Loper, E., and Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Bruchmüller, K., Margraf, J., and Schneider, S. (2012). Is ADHD diagnosed in accord with diagnostic criteria? Overdiagnosis and influence of client gender on diagnosis. *Journal of consulting and clinical psychology*, 80(1):128.

Chan, D. (2016). Informational Frictions and Practice Variation: Evidence from Physicians in Training. National Bureau of Economic Research, w21855.

Chan, D. C., Gentzkow, M., and Yu, C. (2019). Selection with variation in diagnostic skill: Evidence from radiologists. National Bureau of Economic Research, w26467.

Chan, E., Hopkins, M. R., Perrin, J. M., Herrerias, C., and Homer, C. J. (2005). Diagnostic practices for attention deficit hyperactivity disorder: A national survey of primary care physicians. *Ambulatory Pediatrics*, 5(4):201–208.

Chorniy, A., Currie, J., and Sonchak, L. (2018). Exploding asthma and ADHD caseloads: The role of medicaid managed care. *Journal of health economics*, 60:1–15.

Currie, J. and MacLeod, W. (2020). Understanding doctor decision making: The case of depression treatment. *Econometrica*, 88(3):847–878.

Currie, J., MacLeod, W. B., and Van Parys, J. (2016). Provider practice style and patient health outcomes: The case of heart attacks. *Journal of health economics*, 47:64–80.

Doshi, J. A., Hodgkins, P., Kahle, J., Sikirica, V., Cangelosi, M. J., Setyawan, J., Erder, M. H., and Neumann, P. J. (2012). Economic impact of childhood and adult attention-deficit/hyperactivity disorder in the United States. *Journal of the American Academy of Child & Adolescent Psychiatry*, 51(10):990–1002.

Elder, T. E. (2010). The importance of relative standards in ADHD diagnoses: Evidence based on exact birth dates. *Journal of health economics*, 29(5):641–656.

Epstein, A. J. and Nicholson, S. (2009). The formation and evolution of physician treatment styles: An application to cesarean sections. *Journal of health economics*, 28(6):1126–1140.

Gowrisankaran, G., Joiner, K. A., and Léger, P.-T. (2018). Physician practice style and healthcare costs: Evidence from emergency departments. National Bureau of Economic Research, w24155.

Hinshaw, S. P. (2018). Attention deficit hyperactivity disorder (ADHD): controversy, developmental mechanisms, and multiple levels of analysis. *Annual review of clinical psychology*, 14.

Holland, K. and Riley, E. (2017). ADHD Numbers: Facts, Statistics, and You. https://www.addrc.org/adhd-numbers-facts-statistics-and-you/.

Kibria, A., Mancher, M., McCoy, M. A., Graham, R. P., Garber, A. M., Newhouse, J. P., et al. (2013). Variation in health care spending: target decision making, not geography.

Kronenberg, C., Doran, T., Goddard, M., Kendrick, T., Gilbody, S., Dare, C. R., Aylott, L., and Jacobs, R. (2017). Identifying primary care quality indicators for people with serious mental illness: A systematic review. *British Journal of General Practice*, 67(661):e519–e530.

Leroy, G., Gu, Y., Pettygrove, S., Galindo, M. K., Arora, A., and Kurzius-Spencer, M. (2018). Automated extraction of diagnostic criteria from electronic health records for autism spectrum disorders: development, evaluation, and application. *Journal of medical Internet research*, 20(11):e10497.

Lowe, J., Pomerantz, A. M., and Pettibone, J. C. (2007). The influence of payment method on psychologists' diagnostic decisions: Expanding the range of presenting problems. *Ethics & Behavior*, 17(1):83–93.

Marquardt, K. R. (2021). Mis(sed) Diagnosis: Physician Decision Making and ADHD. Available at: www.KelliMarquardt.com.

Merten, E. C., Cwik, J. C., Margraf, J., and Schneider, S. (2017). Overdiagnosis of mental disorders in children and adolescents (in developed countries). *Child and adolescent psychiatry and mental health*, 11(1):1–11.

Morgan, P. L., Staff, J., Hillemeier, M. M., Farkas, G., and Maczuga, S. (2013). Racial and ethnic disparities in ADHD diagnosis from kindergarten to eighth grade. *Pediatrics*, 132(1):85–93.

Morley, C. P. (2010). The effects of patient characteristics on ADHD diagnosis and treatment: A factorial study of family physicians. *BMC Family Practice*, 11(1):1–10.

Open Minds (2020). The U.S. Mental Health Market: $225.1 Billion In Spending In 2019. Open Minds Market Intelligence Report: May 6, 2020.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Persson, P., Qiu, X., and Rossin-Slater, M. (2021). Family spillover effects of marginal diagnoses: The case of ADHD. National Bureau of Economic Research, w28334.

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

Schnell, M. and Currie, J. (2018). Addressing the opioid epidemic: Is there a role for physician education? *American journal of health economics*, 4(3):383–410.

Schwandt, H. and Wuppermann, A. (2016). The youngest get the pill: ADHD misdiagnosis in Germany, its regional correlates and international comparison. *Labour Economics*, 43:72–86.

Sciutto, M. J. and Eisenberg, M. (2007). Evaluating the evidence for and against the overdiagnosis of ADHD. *Journal of attention disorders*, 11(2):106–113.

Van Parys, J. (2016). Variation in physician practice styles within and across emergency departments. *PloS one*, 11(8):e0159882.

Visser, S. N., Zablotsky, B., Holbrook, J. R., Danielson, M. L., and Bitsko, R. H. (2015). Diagnostic experiences of children with attention-deficit/hyperactivity disorder. *National health statistics reports*, (81):1–7.

Data: The data was 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 Informatics Services.

Appendices

This appendix presents a very simple example to demonstrate the natural language text processing (NLP) procedure described in Section 3. I also note code and resources used in each step.

Consider the following 3 "fake" documents. The first is an example DSM-V text and the remaining are doctor notes for 2 different patients. In what follows, I title each step and then note how the documents are transformed. I leave out details that have already been presented in the main text.

Raw Text

DSM text: "They will express gloominess."

Note 1: "Mom doesn't express anxty"

Note 2: "pt says they are sad"

Step 1: Text Cleaning & Pre-Processing

An example Medical Dictionary can be downloaded here: `https://github.com/glutanimate/wordlist-medicalterms-en`. Additional medical abbreviations and their expansions are available here: `https://www.tabers.com/tabersonline/view/Tabers-Dictionary/767492/all/Medical_Abbreviations`.

DSM text: "they will express gloom"

Note 1: "mom does not express anxiety"

Note 2: "patient says they are sad"

Step 2: Obtain Word Groupings and Reduce Size

To determine word part of speech, I run the entire patient document through Python's POS tagger found in the NLTK library. Documentation can be found here: `https://www.nltk.org/book/ch05.html`. Most common synonyms are determined using "WordNet" which is available for download here: `https://wordnet.princeton.edu/download`. For additional word groupings, I use pre-trained GloVe word embeddings which can be downloaded here:

https://nlp.stanford.edu/projects/glove/.

DSM text: "express sad"

Note 1: "mom not express anxiety"

Note 2: "patient express sad"

Step 3: Tokenize

DSM text: [express, sad, express sad]

Note 1: [mom, not, express, anxiety, mom not, not express, express anxiety]

Note 2: [patient, express, sad, patient express, express sad]

Step 4: Build the Adjusted BOW Model Matrix

	anxiety	express	express anxiety	express sad	mom	mom not	not	not express	patient	patient express	sad
document 1	0	1	0	1	0	0	0	0	0	0	1
document 2	1	1	1	0	1	1	1	1	0	0	0
document 3	0	1	0	1	0	0	0	0	1	1	1

Table A1: BOW Model Matrix Example

Step 5: Measuring Content Overlap: x^*_{is} and x^*_i

The content overlap value is the cosine similarity (weighted dot product) of each patient vector with the DSM-V vector. The length of each vector is 3, 7, and 5 respectively. The dot product between the DSM-V and each patient vector is 1 and 3 respectively. Therefore, the following match value for the two patients are:

Patient 1 (document 2): $\frac{1}{\sqrt{3*7}} = .218$

Patient 2 (document 3) : $\frac{3}{\sqrt{3*5}} = .775$

In this example there are no other DSM-V symptoms to consider, thus aggregation is not needed. So, the final measures for patient 1 and patient 2 are $x^*_1 = 0.218$ and $x^*_2 = 0.775$. Looking back at the raw text, notice that patient 2 is contextually similar to the DSM-V criteria, yet none of the words are exact matches. My proposed text analysis procedure accounts for this content overlap which is why the resulting metric for patient 2 is appropriately higher than for patient 1 (0.775 >> 0.218).