Identifying Physician Practice Style for Mental Health Conditions

Kelli Marquardt*

March 2021

Click *here* for the latest version.

Abstract

This paper proposes new methods for identifying and estimating physician practice style in the context of mental health diagnoses. Empirical identification in these settings can be difficult as there is no formal biological/medical test to determine the presence of a mental health condition. Instead, diagnosis depends on physician-patient interviews to extract existence of behavioral symptoms and match to documented diagnostic guidelines. To address these identification concerns, I propose a two step estimation procedure. First, I describe unique text-analysis methods applied to digitized clinical doctor notes as a way to measure how closely the patient interview matches mental health diagnostic guidelines according to The Diagnostic and Statistical Manual of Mental Disorders (DSM-V). I then use this measure as a control in a reduced-form model to identify two components of physician practice style: *diagnostic intensity* (the mean propensity to diagnose) and *diagnostic compliance* (the weight that physicians place on official medical guidelines). As an application, I use electronic health record data to estimate physician practice style for a widely prevalent mental health condition, Attention Deficit Hyperactivity Disorder (ADHD). I find significant variation in physician practice style, with physician gender and experience being the strongest predictors of this variation. Finally, I discuss how mental health practice style estimates can be used to guide potential health care policies, and I provide a list of extensions and suggestions of how these methods can be used in future mental health care research.

*The University of Arizona; marquardtk@email.arizona.edu. 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.

1 Introduction

Approximately 1 in 5 adults and 1 in 6 children experience a mental health illness each year; however, roughly half of this population does receive professional mental health care or related treatments.[1] The first step to obtaining this appropriate mental health treatment is an appropriate diagnosis. Unfortunately, evidence suggests that mental health conditions are often inaccurately diagnosed. For example, Merten et al. (2017) review the literature on ADHD misdiagnosis, Bostwick and Rackley (2012) discuss the case of Depression, and Zimmerman et al. (2008) study inappropriate diagnoses for Bipolar Disorder. One reason why these mental health diagnostic errors exist is because diagnostic guidelines are often considered subjective and physicians are imperfect decision makers. This paper presents a way to quantitatively measure *how* physicians make a mental health diagnosis in the presence of uncertainty and discusses ways in which this information can be used to influence mental health care policy.

Identifying physician diagnosis quality is difficult in the mental health setting because diagnosis can not be determined via any biological test or medical imaging. Instead, a physician must conduct a behavioral interview and match present symptoms to those outlined in national guidelines. In the United States, these criteria are summarized in *The Diagnostic and Statistical Manual of Mental Disorders*, which is currently released in its 5th edition (DSM-V).[2]

The aim of this paper is to quantify physician diagnosis quality in the mental health setting. I first present a natural language text processing procedure to be applied to clinical doctor notes in order to determine how "appropriate" a patient is for a mental health diagnosis based on his/her behavioral symptom concerns. I then use a simple regression framework comparing the physician diagnosis decision to the mental health text match value estimated in the first stage. 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 making their diagnosis decision. Together, these two components define the physician practice style, providing a quantitative measure of mental health diagnosis quality.

To demonstrate how the proposed methods can be applied in practice, I focus on a specific mental

[1] See https://www.nami.org/mhstats

[2] The DSM-V largely overlaps with The International Classification of Diseases, which is the standard system used internationally.

health condition, Attention Deficit Hyperactivity Disorder (ADHD). I choose this application 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. An estimate of the annual economic impact of ADHD ranges from $143-$266 billion US dollars (Doshi et al., 2012). 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). Therefore, 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 center in Arizona, which importantly include access to de-identified clinical doctor notes. I use these data to estimate physician practice style for ADHD diagnoses. 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 exactly and diagnoses 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. In my specific ADHD application, I estimate a median diagnostic compliance of 0.44 and diagnostic intensity of -0.17. The negative intensity estimate suggests that the median physician *dislikes* diagnosing ADHD on average. 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 further explore how physician practice style varies with physician characteristics. 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.

My paper adds to the robust literature on estimating physician practice style (e.g. Abaluck et al., 2016; Chan, 2016; Epstein and Nicholson, 2009; Van Parys, 2016). This literature typically estimates practice style by examining the variation in observed physician treatment decisions or medical costs *conditional* on a specific patient diagnosis. In contrast, I focus on the *initial* diagnosis decision. While this adds uncertainty into the model, it allows me to explore an important yet

understudied component of the physician decision-making process.[3] I am also one of the first in this specific literature to consider the mental health setting. One reason why physician diagnosis quality for mental health has not been studied in this literature is due to the subjectivity of the diagnosis decision and the lack of necessary medical data needed for identification. While this idea of mental health care quality has been explored in the medical literature, it is noted that identification is weak mainly because these quality measures rely on physician (subjective) opinions (Kronenberg et al., 2017). As a way to address such data and identification concerns, I propose adding a first stage text analysis procedure to the classic physician practice style models. I describe the specific details of this procedure in Section 2.1.

My paper also adds to the existing literature on ADHD misdiagnosis. A list of papers show that where a child's birth-date falls in relation to the school entry cut-off date is a strong predictor of ADHD diagnosis, implying that teachers are subjectively comparing the younger students in the class to older students and mistaking immaturity for ADHD (Elder, 2010; Evans et al., 2010; Schwandt and Wuppermann, 2016). Another recent study explores how diagnosis depends on patient insurance. Chorniy et al. (2018) find a significant increase in ADHD diagnosis following a switch from Fee-For-service to Medicaid Managed Care which can be partly explained by physician financial incentives and Medicaid quality control initiatives. There are also medical/public health literature exploring the topic of ADHD. 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 and find that the ADHD diagnosis decision may be influenced by patient characteristics such as gender, race, and income (Bruchmüller et al., 2012; Lowe et al., 2007; Morley, 2010). My paper contributes to the discussion on ADHD diagnosis by *quantifying* the role that the physician plays in the ADHD diagnostic decision. I find evidence of substantial variation in physician practice style, indicating differences in ADHD diagnosis decisions for similar patients and providing evidence that diagnostic errors do exist, though the direction and magnitude of misdiagnosis is left for future analyses.

The rest of the paper is outlined as follows. In the next section, I present the proposed methods

[3]Two exceptions are the recent work by Chan et al. (2019) and Gowrisankaran et al. (2017) who model physician diagnosis decisions. However, these papers focus on only physical health applications: pneumonia, angina, appendicitis, and TIA.

for physician practice style estimation. This includes a first stage natural language text processing procedure and a second stage standard regression model. I also note the data that is necessary for estimation. Section 3 applies the proposed methods to the case of Attention Deficit Hyperactivity Disorder, using electronic health record data provided by a large healthcare system in Arizona. I analyze predictors of physician practice style and health care policy implications in Section 4. Finally, in Section 5, I wrap up with a conclusion and mention directions for future work.

2 Methods

I aim to estimate two dimensions of physician practice style: *diagnosis intensity*, the average propensity to diagnose, and *diagnostic compliance*, the weight that the physician places 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, such identification 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 mental health symptom severity using natural language text analysis techniques applied to clinical doctor notes. 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 severity measure needed for identification. Second, I estimate physician practice style by comparing physician diagnosis decisions *conditional* on the constructed symptom severity measure from step 1.

Four key pieces of data are needed for estimation: a patient/encounter identifier, the diagnosing physician identifier, the diagnosis decision, and the clinical doctor note summarizing what was discussed during the patient encounter. All of this information can be found in the patient electronic health record (EHR).

2.1 Text Analysis: Natural Language Processing

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. In physical health applications, this value can be observed via imaging or lab results. However, in a mental health application, physicians make the diagnosis decision based on the presence of symptoms listed in the DSM-V.

Each mental health condition has it's own section in the DSM-V, detailing specific behavioral symptoms and other diagnostic criteria. For example, Table 1 presents the DSM-V symptoms for Attention Deficit Hyperactivity Disorder.

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 avoids, dislikes, or 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).

Even detailed electronic health records do not report readily observable patient behavioral symptoms. Instead, this information is found in the clinical doctor note (under the assumption that physicians appropriately document what is discussed during a physician-patient behavioral assessment). 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, it produces an unbiased yet noisy proxy for patient symptom match based on the official DSM-V text as reference.

NLP Algorithm

In the Natural Language Processing (NLP) literature where training data is limited, the typical method for calculating document similarity is a Bag of Words Model (BOW) with cosine similarity measures.[4] 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 write the DSM-V words exactly. Therefore, I propose a version of the traditional BOW framework, making necessary adjustments to keep semantic context.

I now present a natural language processing algorithm that produces a metric for the overlap between doctor note, i, and DSM-V symptom text, s. 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). While the algorithm is presented more generally, I demonstrate how it should be used in practice in Section 3.2. In the Appendix, I include an overly simplified example to demonstrate each algorithm step and document external resources used.

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.

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

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 number of words becomes quite large and synonym search is computationally expensive.

- Determine the part of speech (POS) for each word in each document.
- For computational purposes, I reduce the 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" → "good").

To further allow for variation in natural language, I also determine the set of 10 "closest" 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 as example, the 10 closest words for the term "inacurate" 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 uni-grams and bi-grams. For example, the phrase "patient is not sad" becomes the vector [patient, is, not, sad, patient is, is not, not sad]. I include bi-grams to allow for negation which further keeps semantic context. It 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. Table A1 in the Appendix 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 by the researcher.

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 severity. Specifically, higher values of x_i^* denote a closer overlap between patient note and DSM-V diagnostic criteria compared to lower values of x_i^*.

2.2 Estimating Physician Practice Style

After obtaining the proxy for patient mental health 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 chose to diagnose patient i with the mental health condition of interest.

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

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. 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*. While the ideal point value for β_j and α_j will depend on the application and aggregation choices made in the stage 1 NLP procedure, one would expect that relatively high values of compliance (β) and low values of intensity (α) should lead to more accurate mental health diagnoses on average.

Within each physician, the estimate $(\hat{\alpha_j}, \hat{\beta_j})$ is identified by comparing diagnosis decisions for patients with varying x_i^*. These will be unbiased estimates of the true practice style under the following condition: $E[\varepsilon_{ij}|x_i^*] = 0$ for all patients i of physician j. Assuming that a physician will document their notes in the same way across each individual patient, this condition is satisfied.

These measures can additionally be compared *across* physicians (as I show in Section 4), though the assumption here is slightly stronger, requiring $E[\varepsilon_{ij}|x_i^*] = E[\varepsilon_{kj'}|x_k^*] = 0$ for each patient i,k and each physician j, j'. This assumption is satisfied if (i) patients are randomly assigned or (ii) patients do not select physicians based on mental health factors that are observable to the physician

but unobservable to the physician. I discuss possible tests and implications of this assumption in Section 4.

3 Empirical Application: ADHD

In this section, I demonstrate how the proposed methods can be applied in practice by estimating physician practice style for a specific mental health condition, Attention Deficit Hyperactivity Disorder. ADHD, recognized by symptoms of inattention, hyperactivity, and impulsivity, is the most common and fastest growing child mental health condition. Today, it is estimated that 10% of children have ADHD, though research has suggested that this condition is often improperly identified. While this paper does not take a direct stance on the extent of ADHD over-diagnosis, the large diagnosis rates makes ADHD a great application to show (1) how to estimate physician practice style for a mental health condition, and (2) how these estimates can be used to potentially guide mental health care policy.

3.1 Data

I obtain electronic health record data (EHR) from a large healthcare system located in Arizona. The data range from January 2014 through September 2017. Because the DSM-V diagnosis criteria for ADHD differs for children and adults, in this application I consider only on children ages 5-17. I have access to patient identifiers, physician names, associated diagnoses (if any), and de-identified clinical 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.[5]

Patient Set

The patient set contains all patients (ages 4-17) that have an encounter with a pediatrician or child psychologist during the sample period. This results in 12,311 unique patients which includes both patients with and without an ADHD diagnosis. Patient summary statistics are presented in Table 2.

[5]Due to sample size limitations, I group together each sub-type of ADHD. See American Psychiatric Association (2013) for information on each type.

Table 2: Patient Summary Statistics

	Mean	S.D
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	

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).[6] A majority of the patients have an appoitments 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").

Physician Set

The 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. While only a unique physician identifier is needed for estimation of physician practice style, in an extended analysis I explore which physician attributes are correlated with practice style (see Section 4). Therefore I take some additional steps to collect physician specific information.

From the electronic health record data directly, I record 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

[6]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. Because this is not related to the physician diagnosis decision, it is unlikely to introduce sample bias.

physician gender, where they attended medical school, and what year they graduated. I determine medical school rank based on US News 2018 rankings.[7] I code a physician as attending a top Medical School if their associated school is ranked in the top 100.

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 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.

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			

3.2 Patient Symptom Match

The goal of the natural language processing algorithm presented in Section 2.1 is to measure how closely patient symptoms match the DSM-V criteria for a given mental health diagnosis. While the algorithm is presented more generally, I make a variety of application-specific choices.

First, 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 documents to be the entire DSM-V text for ADHD (i.e. the entire text from Table 1).

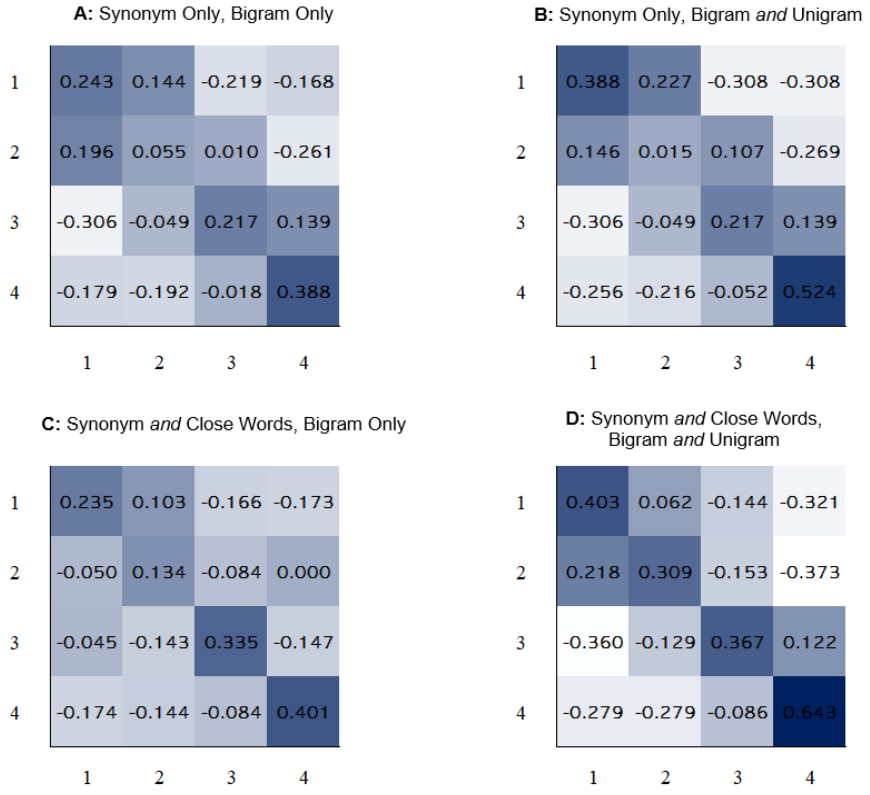
I also test the validity of the algorithm by comparing the outcome measure to a hand-coded sub sample of patient records. Specifically, I collect a random subset of doctor note records for 100

[7]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".

patients and read each note, documenting the number of ADHD symptoms that appear. I then compare the NLP algorithm rank to the hand-coded rank of patient. I do this for the full NLP algorithm in addition to a variety of algorithm-adjusted predictions. The adjustments I consider are to algorithm Step 2 (synonym only vs. synonym *and* 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-preform those using bigrams alone. Additionally comparing B to D suggests the need for both synonym *and* close word match for algorithm Step 2. Therefore, I use the plot D algorithm (full procedure in Section 2.1), when estimating ADHD symptom match for the entire sample.

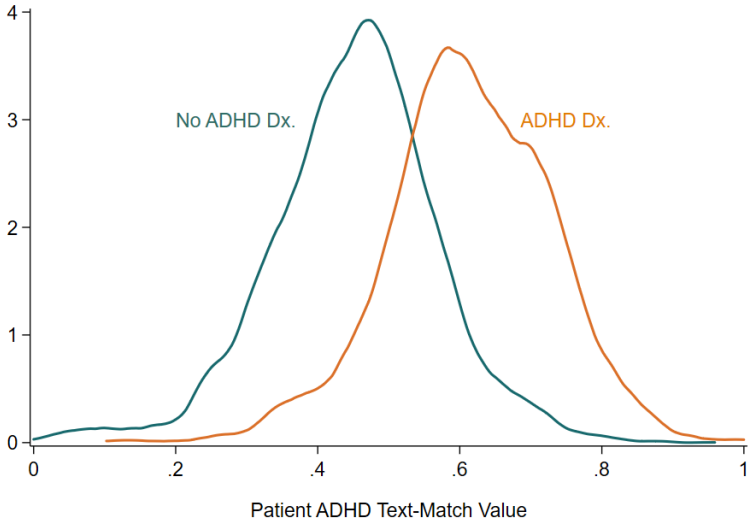
Figure 1: NLP Algorithm Correlation Plots



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.

It takes approximately 48 hours to run the first stage text analysis procedure outlined in Section 2.1 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. Figure 2 displays the distribution of the estimated x_i^* value separated by children with and without an assigned ADHD diagnosis. As expected, patients with an ADHD diagnosis have a higher x_i^* value compared to those patients without the diagnosis.

Figure 2: Density of x_i^*, by Diagnosis



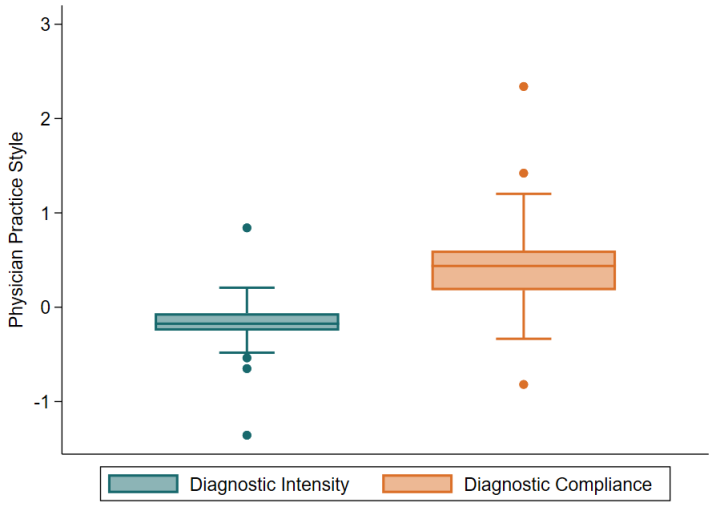
3.3 Physician Practice Style

Next, I estimate physician practice style for each of the 129 physicians in the data according to Equation (1). To visualize physician practice style point estimates and variation, I present the box-plot distribution in Figure 3.

The physician practice style estimates suggest that the median physician prefers *not* to diagnose ADHD ($\hat{\alpha}_{med} = -0.17$) and places 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 DSM-V diagnostic guidelines, they are also likely to rely on prior beliefs and/or other patient signals when making the official diagnosis decision.

The point value interpretation for physician practice style estimates depends on the researchers choice of symptom aggregation in the NLP first stage along with the distributional assumption on ε in estimating Equation 1. In this specific ADHD application, x_i^* is the patient note match across *all* DSM-V symptoms for ADHD, including those for each sub-type, and Equation 1 is a linear probability model of diagnosis. Thus, at point value, the estimates suggest that one standard deviation increase in patient ADHD symptom match will increase the probability of ADHD diagnosis by 11 percentage points given the patient is seen by the median physician.[8] However, Figure 3 shows that variation in practice style exists. Thus, the actual diagnosis response to an increase in symptom severity level will depend on which physician is making the diagnosis decision.

Figure 3: Distribution of Practice Style Estimates



4 Correlates of Practice Style

As a demonstration of how physician mental health practice style estimation can be used to inform potential health care policies, I explore if physician practice style is correlated with physician attributes. 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

[8]The value 11 comes from multiplying the physician compliance estimate of 0.44 with the s.d. estimate of x_i^* which is 0.25.

include Currie et al. (2016) for heart attacks, Van Parys (2016) for minor injuries, and Chan (2016) for internal medicine resource use. The 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:

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

$$\hat{\beta_j} = \delta_0 + \delta_1 Male_j + \delta_2 TopMedSchool_j + \delta_3 Specialty_j + \sum_k [\delta_k experience_{jk}] + \eta_j^b \tag{3}$$

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.

Recall from the model in Section 2.2 that physician practice style estimates can be compared *across* physicians under the assumption that $E[\varepsilon_{ij}|x_i^*] = E[\varepsilon_{kj'}|x_k^*] = 0$ for each patient i,k and each physician j, j' in Equation 1. This assumption is satisfied if (i) patients are randomly assigned or (ii) patients do not select physicians based on mental health factors that are observable to the physician but unobservable to the physician. While random assignment may be suitable for some applications, it is not the case for the empirical exploration in this paper where pediatrician is the patient choice. The assumption may still hold under (ii) so long as physicians fully and accurately document patient symptoms in their note (or deviations from full documentation is idiosyncratic). In this case, any patient selection is observed to the econometrician and captured in x_i^*.

However, this assumption fails if physician deviation from full documentation is correlated with physician characteristics. For example, if all physicians from non-ranked medical schools do not document patient symptoms but still use the symptoms in their diagnosis decision, then these doctors may incorrectly appear more "intense" and less "compliant" than doctors from top-ranked medical schools. Mathematically, $\hat{\gamma}_2$ would be biased downwards and $\hat{\delta}_2$ biased upwards. That being said, even biased estimates from Equations (2) and (3) can still provide fruitful insights with respect to health care policy. I discuss these implications further in Section 4.2.

Table 4: 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

Standard errors in parentheses, clustered at physician level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: 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

Standard errors in parentheses, clustered at physician level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.1 Results

Tables 4 and 5 present the estimates from Equation (2) and Equation (3) respectively. Columns 1-2 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 each each table.

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 4 but positive in Table 5. 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.

4.2 Health Care Policy Discussion

As noted earlier, estimates from Tables 4 and 5 will be biased if physician documentation patterns are correlated with physician characteristics. Therefore, the first policy prescription is to ensure that all physicians comply with the standard behavioral assessment and pediatric best-care practices (American Academy of Pediatrics, 2011). If differences in practice style persist, additional interventions may be needed.

The substantial variation in physician practice style for the ADHD application indicates that physicians will make different diagnosis decisions even for similar patients, implying that ADHD diagnostic errors exist. Because ADHD misdiagnoses will decrease with diagnostic compliance and (likely) increase with diagnostic intensity, the results from Tables 4 and 5 can speak to some potential

policies aimed at reducing such misdiagnosis concerns.

The large 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 project, 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 (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, perhaps within the broad pediatrics specialty.

5 Extensions and Conclusion

This paper presents a new methodology that can be used to estimate physician practice style for mental health diagnoses, relying on two key pieces of data for estimation: the physician diagnosis decision and the clinical doctor note text. The doctor note is essential in the mental health setting as there is no medical/biological observable variable that indicates the presence of a mental health condition. I show how information within the doctor note text can be used to create an observable proxy for unobserved behavioral symptoms.

I detail a two step estimation procedure. The first step uses natural language processing techniques applied to clinical doctor note text in order to obtain a measure for mental health symptom match. The result is a composite measure that summarizes how closely the symptoms discussed during behavioral assessment match with national diagnostic guidelines in *The Diagnostic and Statistical Manual of Mental Disorders*. Second, I show how this value can be used as a control in the physician diagnosis decision model in order to identify two components of physician practice style. These components are estimated via simple linear (or probit) regressions with physician-specific coefficients, in which the slope identifies *diagnostic compliance*, the weight the physician places on

DSM-V documented criteria, and the intercept identifies *diagnostic intensity*, the physician average propensity to diagnose.

I then illustrate how the procedure can be applied in practice by estimating physician practice style for ADHD diagnosis using electronic health record data from a large healthcare system in Arizona. I find significant variation in practice style across physicians in the sample, noting that both median diagnostic intensity and median compliance are lower than what would be expected in the ideal setting. I further explored 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. I then discussed how these practice style estimates and correlation analysis can be used to influence mental health care policy.

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 a variety of different mental health conditions. Are physicians better at diagnosing certain 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 in Gowrisankaran et al. (2017) 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 males and females is quite large (14.8% vs 6.7%).

Additionally, these methods could be used to explore differences 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 spacial and/or time variation in diagnosis rates can be attributed to physician practice style for mental health conditions? These questions (and more) can all potentially be answered with the framework and methods presented in this paper.

While the proposed methods are novel and can be used in a variety of applications, this paper is not without its limitations. First, the methods rely on the assumption that physicians appropriately document what is discussed during patient interviews. Physician estimates may be biased if some physicians are better at documentation than others or if they change their documentation patterns

across patients. A potential test for this would be to compare document similarity both across and within physician notes. Additionally, the text analysis procedure uses only unigrams and bigrams, which can result in a crude measure of symptom overlap. Although computationally expensive, future work should explore more advanced text comparison procedures in order to reduce the level of noise and/or measurement error that exists in the current proposed procedure.

Understanding how physicians make diagnosis decisions in mental health applications is extremely important from a health care policy perspective. This paper provided a set of quantitative tools that researchers can use to address such mental health care questions. The goal of this paper is to encourage others to use the methods and add to the important yet understudied economics of mental health.

References

Abaluck, J., Agha, L., Kabrhel, C., Raja, A., and Venkatesh, A. (2016). The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care. *American Economic Review*, 106(12):3730–3764.

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.

Bostwick, J. M. and Rackley, S. (2012). Recognizing mimics of depression: The 8 d's. *Current Psychiatry*, 11(6):30–36.

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):128–138.

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

Chan, D., 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., Hodgkins, P., Kahle, J., Sikirica, V., Cangelosi, M., Setyawan, J., Erder, M. H., and Neumann, P. J. (2012). Economic impact of childhood and adult attention-deficit/hyperactivity disorder in the united states. *J. Am. Acad. Child Adolesc. 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.

Evans, W. N., Morrill, M. S., and Parente, S. T. (2010). Measuring inappropriate medical diagnosis and treatment in survey data: The case of ADHD among school-age children. *Journal of Health Economics*, 29(5):657–673.

Gowrisankaran, G., Joiner, K., and Léger, P.-T. (2017). Physician Practice Style and Healthcare Costs: Evidence from Emergency Departments. National Bureau of Economic Research w24155.

Jurafsky, D. and Martin, J. H. (2018). *Speech and Language Processing:An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd Edition Draft-Ebook. Available at: `https://web.stanford.edu/~jurafsky/slp3/`.

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. (2007). The influence of payment method on psychologists' diagnostic decisions: Expanding the range of presenting problems. *ETHICS & BEHAVIOR,*, 17(1):83–93.

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).

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.

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.

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.

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).

Zimmerman, M., Ruggero, C., Chelminski, I., and Young, D. (2008). Is bipolar disorder overdiagnosed? *Journal of Clinical Psychiatry*, 69(6):935–40.

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

In this Appendix, I use a very simple example to demonstrate the natural language text processing (NLP) procedure described in Section 2.1. My main resource for NLP in general is Jurafsky and Martin (2018). I will note other external resources used as I run through the procedure below.

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 text (refer back to Section 2.1).

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

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) : $\frac{1}{\sqrt{3*5}} = .775$

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 conceptually 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 much larger than patient 1 ($0.775 >> 0.218$).