Mis(sed) Diagnosis: Physician Decision Making and ADHD

Kelli Marquardt*

January 2, 2021

Click *here* for the latest version.

Abstract

While the presence of disparities in healthcare is well documented, the mechanisms of such disparities are less understood, particularly in relation to mental health. This paper develops and estimates a structural model of diagnosis for the most prevalent child mental health condition, Attention Deficit Hyperactivity Disorder (ADHD). The model incorporates both patient and physician influences, highlighting four key mechanisms of mental health diagnosis: true underlying prevalence, patient stigma, diagnostic uncertainty, and physician costs from type I and type II diagnostic errors. I estimate sex-specific structural model parameters using novel electronic health record data on doctors' notes together with machine learning and natural language processing techniques. In raw comparisons, males are 2.3 times more likely to be diagnosed with ADHD than females. Counterfactual simulations using model estimates show that less than one-half of this disparity can be explained by true differences in underlying ADHD prevalence, very little explained by patient preferences, and about 50% attributed to differences in physician decision-making. I show that physicians view *missed diagnosis* to be costlier than *misdiagnosis*, especially for their male patients. Back of the envelope calculations estimate the national economic impact of ADHD diagnostic errors to be \$60-\$117 billion US dollars, suggesting a need for public and/or medical policy responses aimed at increasing diagnostic accuracy and reducing disparities in mental health care.

Keywords: *child mental health, physician decision making, ADHD misdiagnosis*
JEL Classification: I14, D81, J16, C34, C80

*The University of Arizona; marquardtk@email.arizona.edu. I am grateful for helpful feedback from my dissertation committee: Gautam Gowrisankaran, Keith Joiner, Juan Pantano, and Tiemen Woutersen. I also thank Keith Ericson, Christoph Kronenberg, Ashley Langer, Daniel Millimet, Jessamyn Schaller, Chase Eck, Jill Furzer, Rachel Mannahan, Evelyn Skoy, and conference participants at ASHEcon, Western Economic Association Annual Conference, and seminar participants at the University of Arizona. 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

Healthcare disparities, traditionally defined as differences in health treatment and outcomes across population groups, are of substantial concern in the United States.[1] While overall health disparities have been declining recently, mental health disparities show the opposite trend (AHRQ, 2019). Within mental health, disparities are particularly salient for Attention Deficit Hyperactivity Disorder (ADHD). Approximately 10% of children are diagnosed with ADHD, and males are diagnosed and treated 2 to 3 times more frequently than females.[2] The psychology literature suggests that this clinical diagnostic difference is larger than what can be explained by true underlying prevalence rates, with evidence showing over-diagnosis of males and under-diagnosis of females on average (Bruchmüller et al., 2012; Hinshaw, 2018). Both *missed* and *mis*-diagnoses are costly, including lower productivity and human capital accumulation for untreated ADHD and harmful side-effects from over-treatment.[3] Ensuring accurate diagnosis for ADHD is essential because its annual economic impact is large, ranging from $168 to $312 billion U.S. dollars (Doshi et al., 2012).[4]

This paper develops and estimates a model of ADHD diagnosis in order to explore the potential causes of differing diagnosis rates across male and female children. I propose and analyze four key mechanisms of ADHD diagnostic disparities: (1) differences in patient preference to seek mental health care, (2) varying rates of diagnostic uncertainty, (3) heterogeneous physician preferences for ADHD diagnosis, and (4) underlying differences in the true prevalence of ADHD between boys and girls. Importantly, the model also allows me to identify the extent of ADHD diagnostic errors (both *missed* and *mis*-diagnosis) according to national guidelines as well as the potential heterogeneous impact across patient sex.

My model has three distinct stages to reflect how the mental health diagnosis decision is made. In the first stage, patients (and their parents) who see a physician decide whether to discuss behavioral concerns. Patients bear a cost of discussing symptoms that includes a potentially sex-specific cost

[1] U.S. Congress mandates annual *National Healthcare Quality and Disparities Reports* in accordance with the Healthcare Research and Quality Act of 1999. State governments have also enacted legislation in response to healthcare disparities (see: `https://www.ncsl.org/research/health/health-disparities-laws.aspx`).

[2] 14.8% of males and 6.7% of females diagnosed with ADHD according to the 2018 National Center for Health Statistics report.

[3] Diagnosed ADHD is often managed with stimulant medications that fall under the CDC schedule IIN controlled substance category associated with "high potential for abuse which may lead to severe psychological or physical dependence." See: `https://www.deadiversion.usdoj.gov/schedules/`

[4] Inflated to 2019 U.S. dollars using consumer-price-index from the U.S. Bureau of Labor Statistics.

of mental health diagnosis, which I term *patient stigma*. Second, physicians conduct a behavioral assessment for this subset of patients and record/document the patient responses in a clinical doctor note. The physicians use this information to update their belief as to whether the patient matches national guidelines for ADHD diagnosis via a Bayesian learning process. In the final stage, physicians decide whether or not to diagnose the patient with ADHD. They do so if the patient specific posterior belief of ADHD match is above a sex-specific diagnostic threshold. This threshold is set by the physician ex-ante and is a function of the costs they bear from potential diagnostic errors. I allow for both *diagnostic uncertainty* and *diagnostic preferences* to vary by patient sex to emphasize how the physician decision-making process contributes to diagnostic disparities.

I empirically analyze the sex-specific ADHD diagnostic disparity using data derived from electronic health records from 2014 to 2017 provided by a large healthcare system in Arizona. The dataset includes over 136,000 pediatric visits for approximately 30,500 patients. In the raw data, 8% of males and 3% of females are diagnosed with ADHD, implying a male-to-female ADHD diagnostic disparity of 2.26:1. This disparity persists even after controlling for a variety of patient observables in reduced form analyses, supporting the need for a structural model and estimation approach.

I first construct mental health variables necessary for structural estimation using novel data that includes clinical doctor notes and advanced data analytic techniques including machine learning and natural language processing. Specifically, I determine whether patients discuss behavioral concerns with a physician using a machine learning prediction approach based on a training set of appointments in which this label is readily observed in the electronic health record. For the set of patients that seek mental health care, I also use the information provided in the clinical doctor note to construct an observable proxy for the ADHD match signal that physicians receive during the behavioral assessment. To do this, I use natural language processing techniques to measure how closely the encounter summary provided in the doctor note matches with national diagnostic guidelines for ADHD which are outlined in *The Diagnostic and Statistical Manual of Mental Disorders*, currently in its 5th edition (DSM-V).

I then use the constructed mental health variables and clinical diagnoses to estimate the underlying parameters of my structural model. My first stage presents a selection problem in which the ADHD match signal is only observed if the patient first chooses to schedule an appointment to discuss behavioral concerns with a diagnosing physician. While this diagnosing physician may be

chosen endogenously, I assume that the patients' choice of *original* primary care physician is orthogonal to behavioral symptom development. I show that these base primary care physicians have different eventual mental health discussion rates, providing me with an exclusion restriction that allows identification of patient cost of discussing symptoms (stigma). This also allows me obtain required selection-adjusted estimates of the population mean ADHD risk for males and females via extrapolations of ADHD match signals on quasi-exogenous mental health discussion rates. This exogenous extrapolation approach is similar to the methods proposed in Arnold et al. (2020).

Finally, the outcome for the third stage is the patients' clinical diagnosis, assigned by the physician and observed in the electronic health record. I estimate the components of diagnostic uncertainty and physician preferences by exploring differences in diagnosis rates across sex conditional on the constructed ADHD match signal. The weight that the physician places on this signal identifies varying levels of diagnostic uncertainty, with higher weights corresponding to stronger signal quality. I then show that conditional on diagnostic uncertainty, the mean diagnosis rates for each sex is a function of physician prior beliefs and physician disutility from diagnostic errors. I am able to separately identify these two values using estimates of mean ADHD risk by sex obtained in the initial selection stage.

The structural parameter estimates show that less than half of the observed ADHD diagnostic disparity between male and females patients can be attributed to differences in the underlying ADHD risk distribution by patient sex, with the rest explained by variation in physician decision-making across male and female patients. In particular, I find that physicians perceive female ADHD signals to be more informative of true health states and thus place more weight on female patient behavioral assessments when making a diagnosis decision. This is consistent with the vignette study by Bruchmüller et al. (2012) which finds that physicians rely on heuristics rather than official DSM-V criteria when diagnosing males with ADHD. I also find that physicians use significantly lower diagnostic thresholds for male patients, suggesting that physicians bear greater costs of inaccurately diagnosing male patients than females.

Using the diagnosis model and parameter estimates, I run counterfactual simulations to examine the extent of over and under diagnosis by patient sex. I find that physicians view *missed diagnosis* to be more costly than *misdiagnosis* on average, though the cost is much larger for males than females. While this finding may suggest that ADHD is over-diagnosed, counterfactual simulations that additionally account for patient stigma and physician uncertainty show that this condition

is slightly underdiagnosed in the sample population. Based on the DSM-V definition of ADHD, I estimate that 4.5% of the adolescent population is over-diagnosed (6.4% of males and 2.7% of females) and 5.3% of the adolescent population is under-diagnosed (6.1% of males and 4.5% females). Importantly, simulations show that the majority of the *missed diagnosis* rate is due to high patient costs of seeking care, suggesting a potential need to increase mental health education and reduce stigma in the general population.

These results add to the existing literature exploring the potential for ADHD diagnostic errors. For example, in the health economic literature a list of papers show 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 (e.g. Elder, 2010; Layton et al., 2018; Furzer et al., 2020). Understanding ADHD diagnosis is also explored in the medical and public health literature, including meta analyses on diagnostic differences (e.g. Sciutto and Eisenberg, 2007; AHRQ, 2011; Hinshaw, 2018), physician and patient surveys (e.g. Visser et al., 2015; Chan et al., 2005), and vignette studies exploring variation in ADHD diagnosis decisions by patient groups (e.g. Morley, 2010; Bruchmüller et al., 2012). My paper adds to this literature by presenting new estimates of over/under diagnosis along with a structural model to identify where these errors come from.

My paper also contributes to the vast literature on explaining variation and disparities in healthcare. This includes papers estimating physician practice style (e.g. Epstein and Nicholson, 2009; Currie et al., 2016; Gowrisankaran et al., 2017), structural models of physician decision making under uncertainty (e.g. Abaluck et al., 2016; Currie and MacLeod, 2017; Chan et al., 2019), and identification of physician prejudice (e.g. Balsa et al., 2005; Chandra and Staiger, 2010; Anwar and Fang, 2012). This existing literature typically focuses on physical health applications and thus relies on two assumptions that do not hold in mental health settings. The first is that patient preferences play a small role in explaining variation in healthcare (Cutler et al., 2019). While this assumption of insignificant demand-side influences might be supported in physical health applications, it is not the case with mental health in which patient stigma plays a potentially large role in determining a diagnosis. My paper develops a novel model of mental health diagnosis taking insights from this literature and adding a patient selection stage in order to explore how both demand-side and supply-side factors can lead to diagnostic disparities in mental health. Second, the extant literature assumes that health states or true diagnoses are observed on some level, which is not the case in

mental health applications as diagnosis is based on the presence of behavioral symptoms and cannot be confirmed via medical testing. My paper innovates to deal with this challenge by instead using clinical doctor note data and text analysis techniques to construct a proxy for ADHD match according to national diagnostic guidelines.

Finally, the methods I use in this paper also add to the more recent literature on using Text Analysis, Machine Learning and Natural Language Processing techniques in economic research (see Currie et al., 2020, and cites therein). In this paper, I combine machine learning methods outlined in Clemens and Rogers (2020) with text analysis methods proposed in Marquardt (2020) to construct key mental health variables which I then use in a structural model to estimate variation in both patient and physician decision-making. While I focus on ADHD in particular, the methods I propose can be used in a variety of settings where researchers have access to clinical doctor notes, especially those focused on mental health in which diagnosis depends on subjective interviews documented via text as opposed to biological testing/ medical imaging.

The remainder of this paper is structured as follows. Section 2 provides medical details on ADHD diagnosis to help motivate the theoretical model, which is then outlined in Section 3. In Section 4, I summarize the electronic health record data with reduced form comparisons and describe the the machine learning/natural language processing techniques used to extract important information from clinical doctor notes. Section 5 discusses the empirical strategy and identification. Section 6 presents the structural model estimates which are then used in ADHD diagnostic simulations in Section 7. Finally, Section 8 concludes.

2 Background and Medical Details

I study the physician decision to diagnose Attention Deficit Hyperactivity Disorder in children and young adolescents. ADHD is a chronic mental disorder associated with symptoms of inattention, hyperactivity, and impulsivity. These symptoms are associated with lower educational attainment (Currie and Stabile, 2006) in addition to long term effects on earnings and employment opportunities (Fletcher, 2014; Knapp et al., 2011). Importantly, treatment through stimulant medication and/or behavioral therapy has been shown to reduce the symptoms and associated costs related with this condition (Jensen et al., 2001), making accurate ADHD diagnosis and subsequent treatment essential for human capital development.

While the exact cause of ADHD is unknown, the medical literature agrees there is a strong heritability component. However, genetics alone do not indicate a diagnosis, and there is less consensus regarding other environmental and structural factors (Hinshaw, 2018).[5] There is no biological or medical test to determine the presence of ADHD in a given patient. Instead, an ADHD diagnosis is defined by a list of behavioral symptoms outlined in *The Diagnostic and Statistical Manual of Mental Disorders*, currently in its fifth edition (DSM-V).[6] There are three possible types or presentations of ADHD: inattentive, hyperactive-impulsive, and combined type. A child (ages 4-18) meets the clinical definition of ADHD if they meet 6 or more behavioral symptoms presented in Table 1. In addition, these symptoms should be present in two or more settings (e.g. home and school) and experienced before age 12.

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).

[5]Common risk factors mentioned in the medical literature include: low birth-weight, prenatal toxins, and exposure to lead. A list of more debated causes include: food additives/diet, in-utero cellphone radiation, and excess exposure to television/video games.

[6]The 5th edition of the DSM was released in May 2013; however, guidelines for ADHD in particular did not change significantly from the DSM-IV edition(Epstein and Loren, 2013).

It should be noted that the DSM-V does *not* make any differentiation in either symptom definitions or diagnostic guidelines according to the patient's sex. This is important for modeling and counterfactual diagnosis purposes as it explicitly restricts differences in ADHD prevalence by patient sex to come only from differences in symptom expression rates in the general population. Bruchmüller et al. (2012) discuss the medical and epidemiological literature on ADHD presentation and diagnosis, and conclude it is "unlikely that gender differences in the expression of ADHD can fully account for the fact that boys with ADHD receive treatment two to three times more often than girls with ADHD." This motivates the question: what other factors contribute to the large difference in ADHD diagnosis rates between boys and girls? To answer this question I first outline how an ADHD diagnosis is made.

In order to receive a clinical diagnosis, a physician must first conduct a behavioral assessment. In most cases, mental health concerns are first brought to the child's primary care physician during annual wellness checks in which physicians are encouraged to ask about school performance (American Academy of Pediatrics, 2011).[7] If warranted by the response, the primary care physician will then encourage the parent to schedule a follow-up appointment (either with themselves, another pediatrician, or a psychiatrist) so that a full behavioral assessment can be conducted. However, it is ultimately up to the parent to decide whether or not to schedule a follow-up behavioral assessment appointment for their child.

According to pediatric best-care practices outlined in American Academy of Pediatrics (2011), a behavioral assessment should include an interview with the patient, the parent, and a teacher or alternative care-giver. Physicians may use published ADHD rating-scales along with open-ended questions, but should consult the DSM and document the presence of relevant symptoms. Based on this assessment, the physician should diagnose ADHD if they believe the patient meets the minimum requirements for diagnosis outlined in the DSM-V.

While American Academy of Pediatrics (2011) outlines best-practices for ADHD diagnosis, they also admit that these guidelines are often difficult for pediatricians and primary care physicians to follow in practice "because of the limited payment provided for what requires more time than

[7] In many cases, a child's teacher or guidance counselor may be the first to address behavioral concerns with a parent, however only licensed physicians are authorized to clinically diagnose and treat ADHD, therefore requiring a clinical appointment.

most of the other conditions they typically address." Due to time, payment, or a variety of other constraints, it is unlikely that physicians are able to strictly follow these best-practice guidelines. In fact, surveys suggest that only about 60% of physicians incorporate these guidelines into their practice (Rushton et al., 2004; Chan et al., 2005). This fact, along with the institutional features of non-mandatory mental health screening, motivates the need for a structural model of ADHD diagnosis that incorporates these different elements of diagnosis in order to identify the different mechanisms leading to diagnostic disparities.

3 Conceptual Framework

I start with the standard binary classification problem, closely following the set-up in Chan et al. (2019), who model radiologist decision-making for pneumonia patients. I then adjust the standard framework to include a patient selection component which is important in mental health applications as screening is not mandatory.

Let $S_i \in \{0, 1\}$ indicate whether child i has ADHD according to the DSM-V definition of the condition, and $D_i \in \{0, 1\}$ indicate whether child i receives a clinical ADHD diagnosis. The physician's goal is to align the binary health state S_i with the diagnosis decision D_i. Because there is no medical/biological test to confirm S_i, this state is unobserved in clinical practice. Therefore, physicians must make diagnostic decisions based on noisy signals of S_i which inevitably leads to some rate of diagnostic errors, including both False Negatives (FN) and False Positives (FP). Panel A of Figure 1 presents the classification matrix in this standard form. As shown in equation 1, diagnosis rate can be re-written as a prevalence adjusted function of false positive rates ($FPR = \frac{FP}{TN+FP}$) and false negative rates($FNR = \frac{FN}{TP+FN}$), where $\overline{S}$ denotes the true prevalence (i.e. the fraction of patients with $S_i = 1$). Following Chan et al. (2019), I assume that these diagnostic errors can be attributed to diagnostic uncertainty and/or physician preferences across errors.[8] Thus, in the standard setting, diagnostic disparities across groups can come from difference in prevalence rates, differences in diagnostic uncertainty, and/or differences in physician preferences.

[8]Diagnostic uncertainty can be defined in a variety of ways. Chan et al. (2019) assume diagnostic uncertainty comes from *physician skill*. Other papers with similar modeling structures use the terms *signal strength*, *signal quality*, or *test accuracy*. In this paper, I use the terms 'diagnostic uncertainty' and 'signal quality' interchangeably when referencing this potential source of diagnostic errors.

$$\%Diagnosed = (1 - FNR) \times \overline{S} + FPR \times (1 - \overline{S}) \tag{1}$$

Figure 1: Classification Matrices

(a) Standard Form

True ADHD	Clinical Diagnosis: $D_i = 0$	$D_i = 1$
$S_i = 0$	TN	FP
$S_i = 1$	FN	TP

(b) Complete Form

True ADHD	Clinical Diagnosis: $D_i = 0$	$D_i = 1$
$S_i = 0$	$TN_{Q_i=1}$ / $TN_{Q_i=0}$	$FP_{Q_i=1}$ / —
$S_i = 1$	$FN_{Q_i=1}$ / $FN_{Q_i=0}$	$TP_{Q_i=1}$ / —

However, this standard form misses a potentially important mechanism of diagnostic disparities, namely *patient preferences*. A child cannot be diagnosed with ADHD unless they first schedule an appointment to discuss mental health symptoms. Letting $Q_i \in \{0, 1\}$ indicate this patient choice, the updated classification matrix that corresponds to this setting is presented in panel B of Figure 1. A *misdiagnosis* (or False Positive-FP) occurs when a child does not have ADHD but receives a clinical diagnosis. Because a child can only be diagnosed if they first seek mental health care, this corresponds to $FP_{Q_i=1}$ in Figure 1. A *missed diagnosis* (or False Negative-FN) occurs when a child has ADHD but does not receive a clinical diagnosis. This can happen if a child does not seek care ($FN_{Q_i=0}$) or if they do seek care but the physician does not diagnose ($FN_{Q_i=1}$). As before, the diagnosis rate can be written as a prevalence adjusted sum of error rates. However, in this updated framework, the different rates can be separated into a physician error component and a patient discussion component. The false negative rate attributed to the physician is defined as $FNR_{phys} = \frac{FN_1}{TP_1+FN_1}$ and represents the percent of patients with ADHD that the doctor sees but does not diagnose. The false negative rate attributed to the patient is defined as $FNR_{pat} = \frac{FN_0}{TP_1+FN_1+FN_0}$ and represents the percent of patients that have ADHD but do not seek care from a physician. This physician-patient decomposition applies to the false positive rates as well.

$$\%Diagnosed = \big((1 - FNR_{phys}) \times (1 - FNR_{pat}) \times \overline{S}\big) + \big((FPR_{phys}) \times (FPR_{pat}) \times (1 - \overline{S})\big) \tag{2}$$

Equation 2 shows that in addition to the diagnostic error factors from the standard framework, mental health diagnosis rates also depend on patient rates of seeking care. Therefore, mental health diagnostic disparities across groups can be attributed to differences in prevalence rates, differences

in symptom discussion error rates, and differences in conditional diagnosis error rates. As Q is a patient decision, the first error comes from patient 'cost' of discussing mental health concerns with a physician, which I term *stigma*. The latter errors are a result of the physician decision-making process which depends on *diagnostic uncertainty* and *diagnostic preferences*. In the next section, I present a full theoretical model that captures these three mechanisms, while simultaneously allowing for differences in underlying prevalence of ADHD.

3.1 Theoretical Model of ADHD Diagnosis

The model is composed of three stages: a patient selection stage, a physician learning stage, and a clinical diagnosis stage. In the first stage, patients choose to discuss mental health symptoms with a physician if their behavioral concerns outweigh costs associated with symptom discussion. If behavioral concerns are addressed, the patient enters the second stage of the model in which the physician conducts a behavioral assessment, learns about the relevant symptoms, and develops a posterior probability of ADHD match. In the final stage, the physician will choose a diagnosis decision based on ADHD posterior risk and the costs he bears from diagnostic errors. The model allows for prevalence rates, patient costs, physician costs, and physician learning rates to vary by patient sex as a way to capture the varying components of diagnosis implicitly defined in equation 2.

ADHD Prevalence

Each child has some unobserved latent ADHD risk, v_i, which measures the extent of ADHD related symptoms. This comes from a continuous distribution $F_\theta(v)$, where θ indicates whether patient sex is male or female: $\theta \in \{m, f\}$. For computational simplicity, I assume $F_\theta(v)$ is a Normal CDF, though this assumption is not essential for identification, further discussed in Section 5.

$$v_i \sim N(\mu_\theta, \sigma_\theta^2) \tag{3}$$

This continuous mental health risk is in line with the medical literature that suggests ADHD symptoms present on a continuum (AHRQ, 2011). Despite this fact, ADHD diagnosis is binary by construct. Following the diagnostic guidelines in defining ADHD, a child has ADHD if and only if they meet all the requirements for diagnosis outlined in the DSM-V. Therefore, $S_i = 1(v_i > \overline{v})$ where $\overline{v}$ is the DSM-V defined minimum requirement for diagnosis, which by definition does not

vary by patient sex.[9] Thus, differences in true ADHD prevalence by patient sex depend only on differences in ADHD risk distribution parameters, with prevalence increasing in population mean risk, μ_θ.

Stage 1: Patient Choice to Discuss Behavioral Symptoms

In the first selection stage of the model, the patient/parent must decide whether or not to discuss symptoms by scheduling a behavioral appointment.[10] Parents will schedule a symptom discussion appointment if the child's behavioral symptoms outweigh any costs of discussing mental health with a physician, c_i, which includes stigma associated with mental health labels, c_θ, and an idiosyncratic cost $\varepsilon_i \sim N(0,1)$. I allow patient stigma to vary by patient sex (e.g. parents may feel more comfortable with a male child being labeled as 'ADHD' than a female child), but assume all other individual specific costs (e.g. time constraints, distance to clinic, etc.) are exogenous and independent of patient sex.[11] I assume the patient observes their costs $c_i = c_\theta + \varepsilon_i$ and their symptoms, v_i, but does not have enough medical information to know $\overline{v}$, thus motivating them to seek professional opinion. Denoting Q_i as an indicator for mental health symptom discussion, I define $Q_i = \mathbb{1}(v_i > c_i)$. Equation (4) defines the sex-specific symptom discussion probability rate, which follows from (3) and the assumption that $\varepsilon_i \sim N(0,1) \perp\!\!\!\perp v_i$.

$$\Pr\left(Q_i = 1 \mid \theta\right) = \Phi\left(\frac{\mu_\theta - c_\theta}{\sqrt{1+\sigma_\theta^2}}\right) \tag{4}$$

Stage 2: Physician Learning via Behavioral Assessment

I assume that the physician knows the sex-specific ADHD risk distribution, but does not know patient specific ADHD risk, v_i, nor the patient specific discussion costs, c_i. Thus, the physician prior can be defined by (3) and is a function of ADHD risk distribution parameters μ_θ and σ_θ.[12] If

[9] In the 2013 DSM-V release, guidelines were updated to reflect varying levels of symptoms severity. While these are associated with different CPT codes in how a physician is reimbursed, ICD-9 and ICD-10 codes were not adjusted and still reflect binary indicators, validating the assumption to use a single-valued cut-off. In the main estimation section of this paper, I do not assume a $\overline{v}$ value. However, this is necessary in counterfactual simulations which I discuss further in Section 7.2.

[10] Because I focus on children as patients, I assume the parent and child make joint decisions and thus simply refer to "patient" throughout the model.

[11] Out of Pocket monetary costs of treating ADHD are generally low. Medicaid completely covers traditional therapy and ADHD medications. Patients covered by commercial plans will pay more, but can opt for generic medications to reduce costs.

[12] This prior implicitly assumes that the physician does not know the patient symptom discussion cost distribution parameters, or at least does not use these in the ADHD risk learning process.

a patient chooses to discuss symptoms, the physician will learn about the patient specific ADHD risk, v_i, via a behavioral assessment. Through this process, the physician receives a noisy signal, x_i, of the true ADHD risk v_i, defined by equation 5. The signal is unbiased and correlated with the true state through $\rho_\theta \in (0,1)$. I allow correlation to vary by patient sex as a way to capture variation in diagnostic uncertainty coming from signal quality.[13]

$$\begin{pmatrix} v_i \\ x_i \end{pmatrix} \Bigg| \theta \sim N\left(\begin{pmatrix} \mu_\theta \\ \mu_\theta \end{pmatrix}, \begin{pmatrix} \sigma_\theta^2 & \rho_\theta \sigma_\theta^2 \\ \rho_\theta \sigma_\theta^2 & \sigma_\theta^2 \end{pmatrix} \right) \tag{5}$$

The physicians then use this information to update their belief of ADHD risk via a Bayesian updating process. After observing $x_i = x$ the physicians update their prior, resulting in the posterior ADHD risk distribution defined in (6). Notice that the updated risk posterior mean is a weighted average of patient observed signal, x, and the physician prior risk mean, μ_θ, where the weight placed on the signal depends on the signal quality ρ_θ.

$$v_i \mid x \sim N\left(\left(\rho_\theta x + (1-\rho_\theta)\,\mu_\theta \right), \sigma_\theta^2 \sqrt{1-\rho_\theta^2} \right) \tag{6}$$

Stage 3: Physician Diagnosis Decision

Finally, the physician makes a binary diagnosis decision, $D_i \in \{0,1\}$. I follow the literature in assuming the goal of the physician is to match the diagnosis decision to the true health state, and thus minimize diagnostic errors. This can be modeled as a a risk-threshold decision rule where physicians diagnose ADHD to patients whose posterior risk of ADHD is above a diagnostic threshold, τ_θ.

$$D_i \mid x, \theta = \mathbb{1}(v_i \mid x \geq \tau_\theta) \tag{7}$$

In Appendix C.1, I present a physician utility framework and derive this risk-threshold decision rule to show how τ_θ can be interpreted as physician preferences over diagnostic errors. Intuitively, if physicians view *misdiagnosis* as costly, they are worried about diagnosing children on the margin of ADHD according to risk and will thus apply a higher diagnostic threshold. On the other hand,

[13]This health signaling structure is very similar to that defined in Chan et al. (2019), but assumes that signal strength varies across patient types as opposed to physician types.

if physicians view *missed diagnoses* as costly, they would prefer to diagnose children on the margin of ADHD and will thus apply a lower diagnostic threshold. I allow these thresholds to differ by patient sex to capture potential differences in physician perceived cost of diagnostic errors.[14]

Using the physician posterior in equation 6, the probability a patient is diagnosed, conditional on symptoms discussion and received signal, is:

$$\Pr\left(D_i = 1 \mid Q_i = 1, x_i, \theta\right) = \Phi\left(\frac{1}{\sigma_\theta\sqrt{1-\rho_\theta^2}}\left(\rho_\theta x_i + \left(1-\rho_\theta\right)\mu_\theta - \tau_\theta\right)\right) \tag{8}$$

3.2 Mechanisms of Diagnosis and Diagnostic Disparities

Combining equations 4 and 8 yield the sex-specific diagnosis rate equation:

$$\begin{aligned}\Pr\left(D_i = 1 \mid \theta\right) &= \Pr\left(D_i = 1 \mid Q_i = 1, x_i, \theta\right) \times \Pr\left(Q_i = 1 \mid \theta\right) \\ &= \underbrace{\Phi\left(\frac{1}{\sigma_\theta\sqrt{1-\rho_\theta^2}}\left(\rho_\theta x_i + \left(1-\rho_\theta\right)\mu_\theta - \tau_\theta\right)\right)}_{\text{Physician Diagnosis Rate}} \times \underbrace{\Phi\left(\frac{\mu_\theta - c_\theta}{\sqrt{1+\sigma_\theta^2}}\right)}_{\text{Patient Discussion Rate}}\end{aligned} \tag{9}$$

As discussed in the beginning of this section and outlined in equation 2, diagnosis rates are a function of prevalence, patient discussion errors due to stigma, and physician diagnostic errors due to preferences and/or diagnostic uncertainty. My structural model captures each of these elements via μ_θ, c_θ, τ_θ, and ρ_θ respectively.

The comparative statics of population-group diagnosis rates are quite intuitive. Groups with higher prevalence, captured by mean risk, μ_θ, are associated with higher diagnosis rates.[15] This increase can be attributed to both the patient discussion channel ($\frac{\partial Pr(Q_i)}{\partial \mu_\theta} > 0$) and the physician conditional diagnosis channel ($\frac{\partial Pr(D_i|Q_i)}{\partial \mu_\theta} > 0$), where the latter is due to higher physician prior

[14] In analogous models coming from the physician bias literature, this threshold is often referred to as taste-based discrimination as it captures the difference in diagnosis rates for identical patients in terms of risk. However, it may be that the cost of diagnosis errors differ by patient sex, in which case the heterogeneous thresholds are justified. I leave this distinction to the medical literature and instead refer to differences in τ_θ as differences in *perceived* cost of errors, remaining agnostic about its medical accuracy.

[15] Prevalence rates are technically defined as $P(S = 1|\theta) = P(v_i > \overline{v}|\theta)$ where $\overline{v}$ is the DSM-V specified cut-off rule. Provided $\overline{v}$ is not too large, it follows from $v_i \sim N(\mu_\theta, \sigma_\theta^2)$ that there is a one-to-one monotonic correspondence between prevalence and mean risk.

beliefs. On the other hand, high values of patient stigma imply lower diagnosis rates because less patients choose to seek mental health care ($\frac{\partial Pr(Q_i)}{\partial c_\theta} < 0$). In terms of physician preferences, high diagnostic thresholds, corresponding to large cost of misdiagnosis, are associated with lower diagnosis rates ($\frac{\partial Pr(D_i|Q_i)}{\partial \tau_\theta} < 0$). Finally, groups with lower diagnostic uncertainty (i.e. higher ρ_θ) will have higher population diagnosis rates ($\frac{\partial P(D_i=1|Q_i=1)}{\partial \rho_\theta} > 0$ in the selected sample).[16]

I intuitively extend the discussion of population group comparative statics to explaining mechanisms of diagnostic disparities between males and females: $\Delta = \frac{P(D|\theta=m)}{P(D|\theta=f)}$. Diagnosis rates increase with population prevalence and signal quality and decrease with patient stigma and diagnostic thresholds. Therefore, the ADHD diagnostic disparity seen between males and females may be attributed to higher male prevalence ($\mu_m > \mu_f$), higher signal strength for male patients ($\rho_m > \rho_f$), lower stigma for male children ($c_m < c_f$), or lower diagnostic thresholds applied to male patients ($\tau_m < \tau_f$). From a healthcare policy standpoint, it is essential to identify which of these mechanisms explain the diagnostic disparity and by how much. The direction and relative contribution of each mechanisms is an empirical question which I explore in the remainder of this paper.

3.3 Empirical Approach Outline

In this section I briefly outline my estimation approach to help motivate the empirical portion of this paper that follows.

To identify the mechanisms of diagnostic disparities, I separately estimate the structural parameters for both male and female patients: $(\mu_\theta, \sigma_\theta, c_\theta, \rho_\theta, \tau_\theta)$ for $\theta \in \{m, f\}$. I use electronic health record data and estimate equation 9 separately for male and female patients. Estimation requires the following steps: health variable construction, selection methods to obtain population ADHD risk, and maximum likelihood estimation to recover remaining parameters.

The variables required to estimate sex-specific diagnosis rates (9) are clinical diagnosis decision, D_i, symptom discussion indicator, Q_i, ADHD risk signal, x_i, and patient sex, θ_i. However, the only variables directly observed in the electronic health record is D_i (via associated ICD-10 codes)

[16] $\frac{\partial P(D_i=1|Q_i=1)}{\partial \rho} = \phi(\frac{\rho(x-\mu)+\mu-\tau}{\sigma(1-\rho^2)^(1/2)})(\frac{x-\mu+\rho(\mu-\tau)}{\sigma(1-\rho^2)^(3/2)})$. By contradiction, assume this partial derivative is negative. As $\sigma > 0$ and $\rho \in (0,1)$, this implies that $\rho(x-\mu)+(\mu-\tau)$ and $x-\mu+\rho(\mu-\tau)$ have opposite signs. For the selected sample with $Q_i = 1$, symptoms are on average higher than underlying risk implying $x > \mu$. Additionally, assuming physicians would diagnose less than 50% of population, $\tau > \mu$. Therefore, partial derivative is negative if and only if $\rho > \frac{\tau-\mu}{x-\mu}$ and $\rho > \frac{x-\mu}{\tau-\mu}$ which violates the requirement that $\rho \in (0,1)$. Thus, it must be that $\frac{\partial P(D_i=1|Q_i=1)}{\partial \rho_\theta} > 0$ for selected sample.

and patient sex, θ_i. Even though symptom discussion, Q_i, and patient signals, x_i, are not directly imputed to electronic health record systems, this information is available via the clinical doctor note, assuming the physicians document (1) if the patient brings up behavioral concerns, and (2) what is discussed during behavioral assessment. In Section 4 I obtain proxies for both Q_i and x_i using text analysis techniques applied to doctor notes.

I then use these observed and constructed variables to estimate the structural model parameters. I break this down into two steps where the first recovers the sex-specific population mean ADHD risk parameter, μ_θ. This is required for identification of the remaining parameters which I obtain via maximum likelihood estimation of symptoms discussion and conditional diagnosis probabilities following equation (9), separately for male and female patient groups.

While the maximum likelihood estimation is straightforward, recovering population mean ADHD risk requires a selection approach. In an ideal setting in which risk signals are observed for all patients, one could estimate μ_θ with sample averages of x_i. However, x_i is not observed for all patients, and is only observed if patients first discuss symptoms with their physician. Further, symptom discussion is increasing in risk v_i which by construction is positively correlated with signal x_i. Therefore, using sample averages of x_i would over-estimate the mean risk μ_θ. In Section 5.1 I show how this parameter can be estimated using quasi-exogenous variation in symptom discussion rates following an approach applied in Arnold et al. (2020). Section 5.2 then follows with the maximum likelihood estimation procedure and discussion of remaining parameter identification.

4 Data and Variable Construction

The data come from de-identified electronic health records provided by a large healthcare center in Arizona. I obtain encounter level data for all pediatric patients (age<18) who had a health appointment with a diagnosing physician at some point during the sample period of January 2014 to September 2017.[17] I first exclude children younger than 5 years old where rates of ADHD diagnosis and treatment are very low and require peer-to-peer review and prior authorization (N=11,183).[18] I then drop erroneous encounters, encounters with insufficient documentation, or patients with

[17] A diagnosing physician is identified as one who diagnosed ADHD at least once during the sample period. There are 220 diagnosing physicians in my dataset.

[18] see `http://lawatlas.org/query?dataset=adhd-prior-authorization-policies`

missing demographic information (N=1784). The remaining data encompass 37,021 unique patient encounters, for 11,397 unique patients. Patient characteristics include: birthdate, sex, race, original primary care physician, and insurance status. Encounter characteristics include: appointment date, physician seen, associated diagnoses (if any), and most importantly, the clinical doctor note summarizing the encounter.

As ADHD is a chronic condition, the unit of observation in the structural model is at the patient level. I label a patient as clinically diagnosed with ADHD ($D_i = 1$) if the patient has an encounter during the sample period in which one of the first three associated diagnosis codes reflect an ADHD diagnosis.[19] Summary statistics are available in Appendix Tables A1 and A2.

4.1 Reduced Form Comparisons

Of the roughly 11,000 patients seen from 2014 to 2017, 6.24% have a clinical ADHD diagnosis.[20] Males are diagnosed with ADHD significantly more than females. The raw diagnostic disparity is 2.33:1, with 8.68% of males receiving a clinical diagnosis and only 3.72% of females.[21] In Table 2, I show that the male-female diagnostic disparity cannot be explained by differences in observable patient characteristics. This table presents the estimated coefficient on patient sex from a OLS regression of ADHD clinical diagnosis on patient controls. Column 1 shows baseline differences without any added controls. Column 2 shows results after adding demographic observables: patient age, insurance status, and race/ethnicity. Column 3 additionally adds healthcare utilization observables which include: # of doctors seen, # of appointments, appointment year fixed effects, and indicators for other mental health diagnosis, wellness visits, and visits with a psychiatrist.

The results of this reduced form analysis show that the male-female ADHD diagnostic disparity persists even after controlling for a variety of observable patient characteristics, supporting the need for a structural estimation approach.

[19] The ICD-9 codes include 314.00 and 314.01, and the ICD-10 codes include F90.0, F90.1, F90.2. I group together the different types of ADHD into a single diagnosis category as a way to increase power in the presence of small sample sizes.

[20] The in-sample ADHD diagnosis rate is slightly lower than the national average during this time period. This is likely due to the fact that a large portion of the population is of Hispanic ethnicity, and research suggests a significantly lower diagnosis rate for this group (see Morgan et al., 2013). I discuss the implications of this bias in Section 8.

[21] The male-female diagnostic disparity is defined as the ratio of male ADHD diagnosis rate over female ADHD diagnosis rate.

Table 2: Reduced Form ADHD Diagnostic Comparisons

	(1)	(2)	(3)
Male	0.048***	0.048***	0.039***
	(0.004)	(0.004)	(0.004)
Added Patient Observables:			
Demographic Variables	N	Y	Y
Healthcare Utilization Variables	N	N	Y
Adj. R-squared	0.010	0.014	0.072
Observations	11,265	11,265	11,265

Note: This table presents the estimated coefficient on patient sex from a OLS regression of ADHD clinical diagnosis on patient controls. Demographic Variables: patient age, insurance status, and race/ethnicity. Healthcare Utilization Variables: # of doctors seen, # of appointments, appointment year fixed effects, and indicators for other mental health diagnosis, wellness visit, visit with psychiatrist. All controls based on average (or max) across patient appointments, with only those prior to ADHD diagnosis appointment for patients with a clinical diagnosis. Robust standard errors in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3 presents summary statistics for the key variables needed to estimate the structural model.

Table 3: Mental Health Observational Comparisons

	Total	Male	Female	Difference
Full Sample				
ADHD Dx.	0.0624	0.0868	0.0372	0.0495***
	(0.242)	(0.282)	(0.189)	
Discuss Sx. (Q_i)	0.169	0.194	0.143	0.0505***
	(0.375)	(0.395)	(0.350)	
N	11397	5786	5611	
Discuss Sx. Subsample ($Q_i = 1$)				
ADHD Dx.	0.357	0.433	0.250	0.183***
	(0.479)	(0.496)	(0.433)	
ADHD Match Signal (x_i)	0.314	0.320	0.305	0.0149*
	(0.170)	(0.162)	(0.179)	
N	1923	1120	803	

Note: ADHD Dx. (D_i) based on ICD codes in EHR. Discuss Sx. rates (Q_i) and ADHD Match Signal measures (x_i) are constructed using machine learning and natural language processing techniques outlined in Sections 4.2 and 4.3 respectively. Differences calculated as female means subtracted from male means, and significance based on two-sample T-test difference in means. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The top panel of Table 3 presents ADHD diagnosis rates for the full sample and highlights the diagnostic disparity between males and females. As discussed in Section 3.3, there are two key mental health variables that are unobserved to the econometrician yet play a central role in the physician diagnosis decision. These are (1) Q_i which is an indicator for whether a patient schedules an appointment to discuss behavioral concerns with a physician, and (2) x_i which is the patient

specific ADHD match signal observed conditional on behavioral assessment. In the next two sections I discuss how both of these variables are defined and constructed using clinical doctor note data combined with machine learning and natural language processing techniques respectively.

4.2 Mental Health Discussion- Q_i

The electronic health record does not specifically indicate whether a behavioral assessment was conducted during the visit.[22] Therefore, I manually construct this variable from the data by applying machine learning techniques to clinical doctor notes as a way to predict whether a behavioral assessment was conducted during an appointment using the content of the doctor note. I now give a general outline of the procedure.

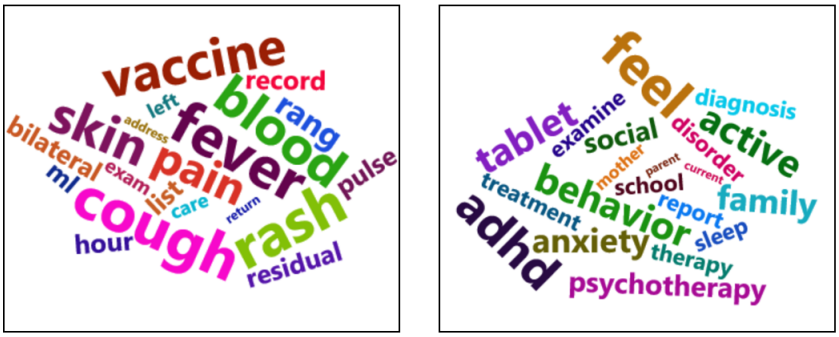
I first take a subset of appointments in which the mental health discussion variable is known with almost certainty. For model training purposes, this must include appointments with a positive discussion label and appointments with a negative discussion label. I assume that a behavioral assessment was conducted if the encounter is associated with an ADHD diagnosis, a differential mental health diagnosis (e.g. bipolar disorder), or a comorbid condition (e.g. generalized anxiety disorder) as noted by the DSM-V. The negative labeled appointments are those with an associated diagnosis that is never co-diagnosed with a mental health condition. These include conditions such as Strep throat, skin rashes, and sinus infections. Table B9 presents the full list of icd9 codes included under each hand label. The remaining appointments are 'unlabeled' due to either no associated diagnoses or appointments with ambiguous icd9 codes that could be related to either mental or physical health concerns (e.g. abdominal pain can be associated with anxiety or a virus). The purpose of this machine learning approach is to determine whether behavioral symptoms were discussed and ADHD diagnosis considered by the physician during these unknown appointments.

I first determine a set of model features using information from the clinical doctor notes of the labeled dataset. I consider 41 features, including note length, relative frequency of the top 20 'positive' label words, and relative frequency of the top 20 'negative' label words. Figure 2 provides a visual of these features with a word cloud representation broken up by mental health discussion label, where $Q_{ij} = 1$ indicates a behavioral assessment was included in notes for appointment j,

[22] There is a variable labeled "visit type", however categories are broad and reporting of this variable appears to be inconsistent.

and $Q_{ij} = 0$ indicates no behavioral symptom discussion. As expected, the positive mental health discussion label includes words related to behavioral symptoms such as: *school*, *social*, *behavior*, *family*, and *feel*. The negative mental health discussion label includes words more related to physical rather than mental health concerns. These include words such as: *pain*, *fever*, *cough*, and *rash*.

Figure 2: Mental Health Symptom Discussion Word Clouds



Non-Behavioral Words **Behavioral Words**

Note: Word clouds based on relative frequency of word-stems in labeled appointments used for machine learning model training, shown separately for Non-Behavioral ($Q_{ij} = 0$) and Behavioral ($Q_{ij} = 1$) labeled appointments respectively. This figure presents full words, whereas actual stems used for prediction are listed in Appendix B.1.

Finally, I use the labeled data and selected features to train a random forest machine learning algorithm, which I then apply to the unlabeled dataset in order to predict whether behavioral symptoms were discussed during the appointment based on the information in the clinical doctor note. I then take the maximum of this prediction across patient encounters to obtain the symptom discussion indicator Q_i used in model estimation, based on a 0.5 prediction cut-off.

There is an important distinction to be made here between full symptom discussion (i.e. behavioral assessment) and symptoms mentioned casually during wellness checks. As ADHD diagnosis can only be made following the former, I require $Q_i = 1$ if and only if the patient receives a behavioral assessment during the sample period. This follows naturally by the construction of the labeled set used in training the machine learning model. The machine learning algorithm will only assign a positive prediction label if the words in the doctor note closely align with the words in the set of appointments with a positive label. By construction, these appointments were ones associated with a mental health diagnosis code and thus the notes most likely reflect a full behavioral assessment. Therefore, appointments in which only a few symptoms were mentioned in passing will not be assigned a positive symptom discussion label by the machine learning prediction. This includes patients who only briefly (or not at all) talk about child behavior when asked during annual wellness

checks. Thus the only way in which patient i receives a positive label $Q_i = 1$ is if either (i) patient i receives a clinical mental health diagnosis during the sample period and thus falls in the training set with positive label, or (ii) at least one of the doctor notes associated with an appointment for patient i contains enough mental health symptom words to be labeled as a behavioral assessment by the machine learning prediction. Additional details behind this machine learning estimation approach are provided in Appendix B.1.

These predicted symptom discussion (behavioral assessment) rates are presented in the top panel of Table 3. The model predicts that approximately 17% of children discuss mental health concerns with a physician, with males presenting with concerns more than females. This average estimate is in line with the *American Academy of Pediatrics* Clinical Guidelines for ADHD which states: "Primary care pediatricians and family physicians recognize behavior problems that may affect academic achievement in 18 percent of the school-aged children seen in their offices and clinics" (Herrerias et al., 2001). Panel B of Table 3 presents diagnosis rates for the set of patients with $Q_i = 1$. The conditional diagnosis rates are still significantly different between males and females, suggesting that symptom discussion rates alone cannot explain the diagnostic disparity.

4.3 ADHD match signal- x_i

Recall that v_i is the (unobserved) true health state and represents a measure of ADHD risk based on behavioral symptoms, and x_i is an unbiased yet noisy signal of v_i that physicians observe during patient behavioral assessment. Because ADHD diagnosis is defined by a list of behavioral symptoms (see Table 1), I interpret v_i as a composite measure summarizing number and severity of symptoms *experienced* by patient i. Following this logic, x_i is then a composite measure summarizing number and severity of symptoms *discussed* with a physician during behavioral assessment.

Even detailed electronic health records do not report readily observable patient behavioral symptoms. Instead, this information is collected during interview and documented in the clinical doctor note. With access to these clinical doctor notes, I construct a proxy for x_i using natural language processing techniques originally proposed in Marquardt (2020). Essentially, I calculate the overlap between words in the DSM-V symptom criteria list (see Table 1) and words in the collective doctor notes for a given patient, making necessary adjustment to account for semantic content. This text-constructed value is a proxy for the signal observed by the physician assuming they follow clinical guidelines in documenting all "relevant behaviors of inattention, hyperactivity, and impulsivity from

the DSM" (American Academy of Pediatrics, 2011).

As x_i is defined on the patient level, I first combine patient notes across encounters into a single document. I combine only notes with that were labeled as 'positive symptom discussion' by the machine learning prediction described in the previous section. For patients without an ADHD diagnosis code, I include notes from all such behavioral appointments. For patients with an eventual ADHD diagnosis code, I include the note associated with the first appearance of ADHD diagnosis and behavioral notes from earlier encounters. I also include notes that occur within 60 days after the initial diagnosis to account for the fact that behavioral assessments may expand over multiple visits and physicians are not always consistent on when diagnosis codes are assigned during this process.[23]

With the behavioral assessment notes combined into one document per patient, I then calculate ADHD signal match, x_i. I follow the natural language processing algorithm proposed in Marquardt (2020), in which patient documents and DSM-V symptom requirements are compared using an Adjusted Bag-of-Words Model. I now outline the general procedure and provide additional details in Appendix B.2.

I first pre-process the clinical texts following standard procedure (e.g. spell check, abbreviation replacement, and size reductions). I next group words according to contextual meaning which requires part-of-speech tagging and synonym replacement. Each document is then broken into uni-gram and bi-gram tokens, where the latter is included to preserve meaning from negation. Using these tokenized documents, I build the adjusted Bag-of-Words (BOW) matrix where rows (i) represent documents, columns (k) represent bi-grams of word groups, and binary matrix elements indicate the presence of bi-gram k in document i. In this application, I consider N+3 documents. The first N correspond to the patient doctor notes for the N patients that receive behavioral assessments. The latter 3 documents correspond to (1) the list of Inattentive symptoms (Type I in Table 1), (2) the list of Hyperactive/Impulsive symptoms (Type II in Table 1), and (3) the combined list of Type I and Type II symptoms from Table 1. In the notation of Marquardt (2020), $s = \{1, 2, 3\}$ corresponds the 3 types of ADHD, Inattentive, Hyperactive/Impulsive, and Combined Type. Fi-

[23]Of the children that are diagnosed with ADHD in my sample, 33% have a symptom discussion appointment within 30 days of the initial diagnosis and 42% have a symptom discussion appointment with 60 days of the initial diagnosis. This suggests that physicians may be breaking up behavioral assessments into multiple visits and assigning ADHD diagnosis codes slightly before the assessment is fully complete.

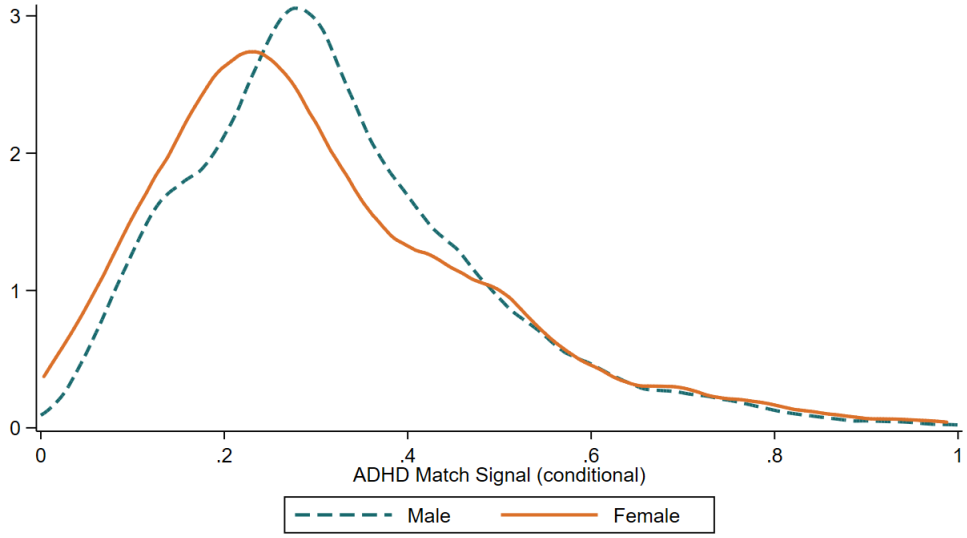
nally, patient-type specific match values, x_{is} are calculated by taking the cosine similarity measure between the BOW row vector for patient i, and the BOW row vector for ADHD Type s. Because I do not distinguish between the different diagnosis types when defining a clinical diagnosis in the data, I construct the patient overall ADHD match signal as the maximum of the patient match value across types. In other words, I calculate $x_i = max\{x_{i1}, x_{i2}, x_{i3}\}$. The Appendix in Marquardt (2020) provides additional intuition for this natural language processing procedure with a simple 3 patient, 1 symptom example.

The average values of x_i for the subset of patients in which it is observed ($Q_i = 1$), are presented in the bottom panel of Table 3. Across both males and females, the average signal match is is 0.314 with a standard deviation of 0.170. For reference, a value of $x_i = 1$ indicates that the note for patient i references *all* symptoms in either the Inattentive list, the Hyperactive/Impulsive List, or the Combined List, and a value of $x_i = 0$ indicates no reference to any symptoms.[24] The signal for males is slightly larger than for females, however the difference is only significant at the 10% level. Figure 3 presents a visual for the ADHD match signal distribution by patient sex. This provides only suggestive evidence of true prevalence differences as the plot represents the match for the (endogenous) set of patients that first discuss mental health concerns with a physician. Therefore, in the general population, ADHD risk signal distributions would be shifted to the left, though the magnitude of the shift and change in dispersion depend on costs symptom discussion and mental health stigma, which I assume is heterogeneous by patient sex.

Interestingly, even though the difference in signals is small, the difference in conditional ADHD diagnosis rates (see Table 3) is large and significant with 44% of males who discuss behavioral concerns receiving an ADHD diagnosis and only 25% of females. This suggests that elements of the physician decision making process (both diagnostic uncertainty and preferences) play a potentially large role in determining diagnostic differences between males and females.

[24]Recall that only a sub-set of symptoms are necessary for appropriate diagnosis, which implies there is some threshold $\overline{x}$ of which $x_i > \overline{x}$ implies ADHD. I remain agnostic about the this threshold value in estimation of the general model, and discuss potential values of this cut-off in Section 7.2 along with its implications on diagnostic errors.

Figure 3: Observed ADHD Match Signal by Patient Sex



Note: Figure shows sex-specific distribution of constructed ADHD match signals x_i based on NLP techniques described in Section 4.3. This implicitly covers the set of patients who discuss mental health symptoms, $Q_i = 1$, thus shows only a truncated distribution of the true population ADHD risk.

5 Empirical Strategy and Identification

With data on ADHD diagnosis D_i, symptom discussion Q_i, patient sex θ_i, and conditional ADHD risk signal x_i, I estimate sex-specific parameters of the structural model: $(\mu_\theta, \sigma_\theta, c_\theta, \rho_\theta, \tau_\theta)$ for $\theta \in \{m, f\}$. As discussed in Section 3.3, the structural parameter estimation procedure requires two steps where the first recovers sex-specific population mean ADHD risk parameter, μ_θ, and the remaining parameters are obtain via maximum likelihood estimation of symptoms discussion and conditional diagnosis probabilities following equation (9), estimated separately for male and female patient groups.

5.1 First Stage: ADHD Population Risk

The reason for a first stage estimation of population mean ADHD risk μ_θ is shown mathematically in (9) but also intuitively following the comparative statics discussion in Section 3.2. Specifically, symptom discussion rates are increasing in mean risk μ_θ and decreasing in patient stigma, c_θ. Additionally, conditional diagnosis rates are increasing in mean risk μ_θ and decreasing in diagnostic threshold τ_θ. This makes it difficult to separately identify the three components even with information on Q_i, x_i, and D_i. In an ideal setting in which risk signals are observed for all patients, one

could estimate μ_θ using sample average risk, x_i. However, x_i is only observed for the subset of patient that discuss behavioral concerns with their physician. Because symptom discussing increases with ADHD risk, the average value of *observed* signals will over-estimate the population risk mean, as shown by equation (10).

$$E[x_i|Q_i = 1] = E[x_i|v_i > c_i] = \mu_\theta + \underbrace{\rho_\theta \sigma_\theta^2 \frac{\phi(\frac{c_i - \mu_\theta}{\sigma_\theta})}{1 - \Phi(\frac{c_i - \mu_\theta}{\sigma_\theta})}}_{\text{upward bias}} \tag{10}$$

I use quasi-exogenous variation in symptom discussion rates to recover unbiased estimates of mean population risk for males and females. To build intuition for this approach, consider a set of patients with extremely low symptom discussion costs, c_i. For low enough levels of c_i, the probably of symptom discussion is approximately 1, so the patient will schedule a behavioral assessment and thus ADHD risk signals, x_i, will be observed. Further, the bias term in (10) for these patients with low c_i goes to 0, and thus sample mean of x_i for patients with low symptoms discussion costs (or conditionally high probability of symptom discussion) provides an unbiased estimation of population mean risk, μ_θ.

As c_i is unobserved in application, I instead estimate individual propensity to discuss symptoms using quasi-exogenous "cost-shifters". An individual factor, Z_i, is a valid cost-shifter under the following two conditions:

(a) Z_i is correlated with symptom discussion rates through patient costs, c_i.

(b) Z_i is independent of patient ADHD risk, v_i.

I use primary care physician identifiers as the source of quasi-exogenous symptom discussion in this application. The electronic health record includes both the *diagnosing physician* as well as the patients' *original primary care physician (PCP)* where the former denotes who the patient meets with during a given appointment, and the later is the PCP originally seen when patient first entered the health system. Because *diagnosing physicians* may be chosen endogenously, I instead focus on the *original primary care physician* and define Z_i as a vector of size p, where $Z_{ip} = 1$ if child i is a patient of PCP p.[25]

[25] I use the *original primary care physician* as opposed to the *diagnosing physician* as the later is likely chosen endogenously. Patients with behavioral concerns may specifically schedule appointments with physicians who specialize in mental health. This would suggest a positive relationship between the diagnosing physician and v_i which violates

To see how the original PCP identifier is correlated with patient symptom discussion costs, it is relevant to recall Section 2 where I discuss the institutional details of behavioral assessment scheduling. Parents may schedule these appointments independently based on own concerns or concerns of teachers. However in many cases, behavioral assessments follow from casual discussions with a primary care physician. PCPs are trained to ask about child's school performance and behavioral concerns during annual wellness visits (American Academy of Pediatrics, 2011), and if warranted by the response may encourage the parent to schedule a follow-up appointment (either with themselves, another pediatrician, or a psychiatrist) so that a full behavioral assessment can be conducted. This discussion and subsequent recommendation from the child's original primary care physician can reduce the cost of scheduling a full behavioral assessment through increased mental health awareness, help with internal scheduling, comfortability with health system personnel, etc., thus satisfying the relevance condition (a).

Importantly, PCPs have discretion over what to address during routine check-ups and whether or not to suggest the patient seek follow-up mental health care. Some may be more thorough during these wellness checks in regard to questions about child behavior, and thus differ in the rates at which they suggest their patients seek follow-up care and schedule behavioral assessments. Appendix Figure A1 shows the variation in follow-up behavioral assessment rates across primary care physicians. To verify that the PCP identifier meaningfully influences the patient probability of full symptom discussion, I conduct a first stage analysis, regressing patient symptom discussion indicator, Q_i, on patient controls and original PCP fixed effects. I test for and find strong joint significance of PCP fixed-effects, results presented in Appendix Table A3.

Condition (b) is satisfied if original PCPs are chosen or assigned independently of true ADHD risk, v_i. As v_i is unobserved, I cannot test for this directly, though a list of observations and institutional details provide support for its validity. First, primary care physicians are typically selected by patients before age 5, which is the age at which behavioral symptoms may develop. This timing structure shows that parents do not chose primary care physicians selectively after observing v_i. Second, there are 600 *original* primary care physicians covering the patients in my sample, but only 24 of these ever diagnose ADHD.[26] So while PCPs may differ in the number of

requirement (b).

[26]There are 220 diagnosing physicians in my sample. 24 of these are the original primary care physician of the

patients they encourage to seek follow-up mental health care, they do not actually diagnose ADHD themselves, suggesting that patients set up behavioral assessments with alternative physicians, again implying no relation between the original PCP and patient v_i. Finally, while patients may not select PCP based on v_i directly, condition (b) would still be violated if PCP selection is based on other factors, W_i, that are correlated with ADHD risk, such as age, race/ethnicity, and income. I test for this by analyzing an ordinary least squares regression of PCP eventual discussion rate on various patient demographics. I define PCP eventual discussion rates as the leave-one-out average symptom discussion rates among all other patients of the given PCP. Appendix Table A5 presents the coefficients from this regression, which are not significantly different from zero, providing support for balance across original primary care physicians.[27]

Under conditions (a) and (b), I can recover population ADHD risk estimates for male and female patients by taking the vertical intercept at one from the fitted relationship between observed ADHD signals and exogenous probability of symptoms discussion. Empirically, I first conduct a probit regression of symptom discussion, Q_i, on patient sex θ_i, other patient covariates W_i, and the PCP fixed effects vector Z_i. I use the results to obtain exogenous symptom discussion probabilities, $\widehat{P_\theta(Q_i|Z_i)}$, by predicting symptom discussion for each sex, holding W_i at sample means. In the absence of sufficient number patients with $\widehat{P_\theta(Q_i|Z_i)} \approx 1$, values of μ_θ can be estimated via extrapolations of observed ADHD match signals on exogenous mental health discussion probability. Specifically, I fit a model of observed ADHD signals, x_i, on $\widehat{P_\theta(Q_i|Z_i)}$ for both male and female patients, and obtain bias-adjusted values of μ_m and μ_f by evaluating the fitted model at $\widehat{P_\theta(Q_i|Z_i)} = 1$ for $\theta \in \{m, f\}$ respectively. This exogenous extrapolation approach is similar to the methods proposed in Arnold et al. (2020) and in line with the literature on identification in selection models (see Chamberlain, 1986; Heckman, 1990). Results of this procedure are presented in Section 6.

patient they diagnose. The remaining 196 physicians are either pediatricians or psychologists that conduct behavioral assessments for patients referred to them by other PCPs in the system.

[27] There may still be concern that patients choose PCPs based on unobserved factors that are correlated with ADHD risk, leading to biased estimates of μ_θ. However, so long as these unobserved factors are independent of patient sex, the relative difference between male and female ADHD risk is unaffected.

5.2 Second Stage: Conditional Maximum Likelihood

I estimate the remaining model parameters via maximum likelihood estimation of diagnosis probability (previous equation 9), separately for male and female patient groups.

$$\begin{aligned}\Pr\left(D_i = 1 \mid \theta\right) &= \Pr\left(D_i = 1 \mid Q_i = 1, x_i, \theta\right) \times \Pr\left(Q_i = 1 \mid \theta\right) \\ &= \underbrace{\Phi\left(\frac{1}{\sigma_\theta\sqrt{1-\rho_\theta^2}}\left(\rho_\theta x_i + \left(1-\rho_\theta\right)\mu_\theta - \tau_\theta\right)\right)}_{\text{Physician Diagnosis Rate}} \times \underbrace{\Phi\left(\frac{\mu_\theta - c_\theta}{\sqrt{1+\sigma_\theta^2}}\right)}_{\text{Patient Discussion Rate}}\end{aligned}$$

With μ_θ estimated in first stage, it is clear how remaining parameters are identified *up to ADHD risk dispersion* σ_θ. Patient stigma, c_θ, is identified through variation in symptom discussion rates *conditional* on mean ADHD risk parameter μ_θ. Both diagnostic uncertainty (ρ_θ) and diagnostic thresholds (τ_θ) are identified in the conditional physician diagnosis probability equation. The correlation between physician diagnosis, D_i, and patient ADHD match signal, x_i, identifies the signal strength ρ_θ. The diagnostic threshold, τ_θ, is identified by mean diagnosis rates *conditional* on ADHD signals, x_i, and mean risk, μ_θ.

Up to this point, the parameter identification has not relied on any functional form assumptions, and thus would follow through if instead ADHD risk and signals were modeled using alternative distributions (e.g. the Beta distribution). However, estimation of the final parameter, ADHD risk dispersion (σ_θ^2), requires an additional moment that depends on this parametric form. Specifically, I identify σ_θ using the moment defined by equation 11 which follows from the truncated normality of selected risk signals. Thus σ_θ is identified by the difference between observed risk signals and population mean risk, adjusting for selection due to stigma and signal strength.

$$\overline{x_{obs}}|\theta = E[x_i|v_i > c_i] = \mu_\theta + \rho_\theta\sigma_\theta\frac{\phi\left(\Phi^{-1}(1-\widehat{Q|\theta})\right)}{\widehat{Q|\theta}} \tag{11}$$

6 Structural Model Estimates

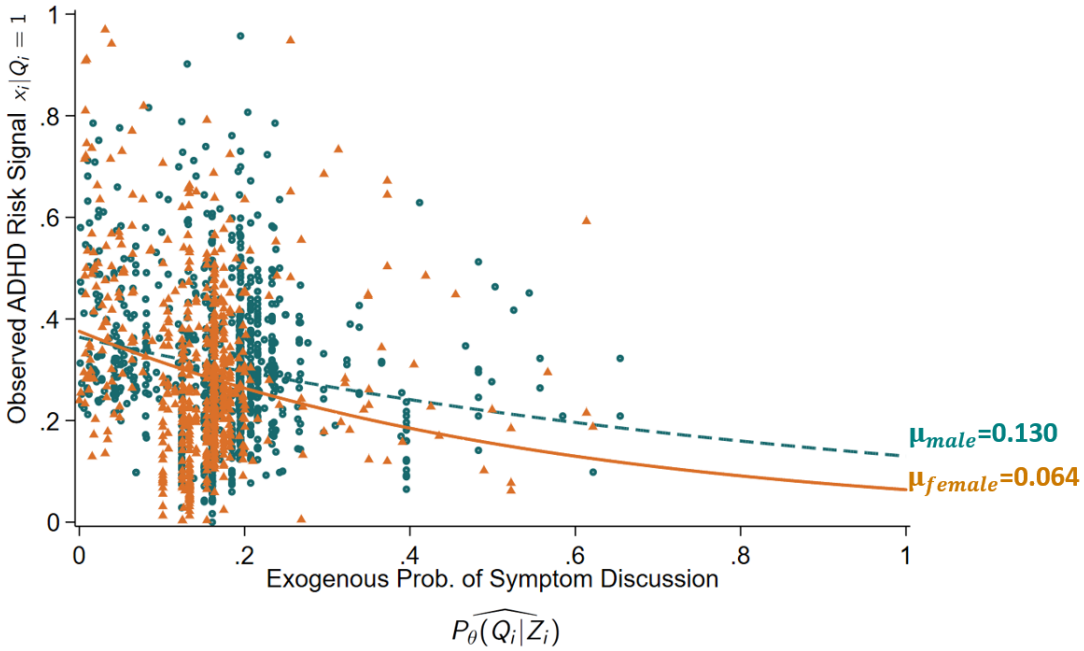
I now present and discuss the results from the 2-stage estimation procedure. Figure 4 illustrates the relationship between observed ADHD signals and exogenous probability of symptom discussion.

The vertical axis plots patient ADHD match signal, x_i, for the set of patients in which it is observed ($Q_i = 1$). The horizontal axis plots estimates of $\widehat{P_\theta(Q_i|Z_i)}$ obtained from the probit estimation:

$$P(Q_i = 1) = \Phi\left(W_i\beta + \delta Male_i + Z_i\gamma\right) \tag{12}$$

Q_i denotes the symptom discussion indicator construction in Section 4.2, W_i includes a set of demeaned patient controls, and Z_i denotes original PCP identifiers. Details and first stage coefficients are presented in Appendix Table A4. With W_i demeaned, $\widehat{P_\theta(Q_i|Z_i)} = \Phi(\hat{\delta}Male_i + \hat{\gamma}_{PCP_i})$ and is interpreted as the regression-adjusted exogenous probability of symptom discussion due to quasi-exogenous variation in original PCP follow-up visit propensity. Figure 4 shows significant variation in $\widehat{P_\theta(Q_i|Z_i)}$, with maximum values near 0.65.

Figure 4: Symptom Discussion Rates and Observed ADHD Risk



Note: This figure plots sex-specific observed ADHD risk signals on predicted symptom discussion probabilities from equation (12), with demeaned patient controls set to 0, for the set of patients with $Q_i = 1$. The figure also plots sex-specific exponential curves of best fit and the associated male and female intercept at 1.

Consistent with the theory, observed ADHD risk signals, x_i, are decreasing in exogenous $\widehat{P_\theta(Q_i|Z_i)}$. A low value of $\widehat{P_\theta(Q_i|Z_i)}$ implies that child i is a patient of a PCP with generally low follow-up behavioral appointment rates, perhaps due to lack of thoroughness at annual wellness visits. Thus, these patients are ex-ante unlikely to schedule a behavioral assessment appointment. Despite this, the patient appears in the data as discussing symptoms anyway, which means that they must have a high ADHD risk draw, v_i, consistent with high observed signal, x_i. On the other hand, a large value of $\widehat{P_\theta(Q_i|Z_i)}$ implies the child is a patient of a PCP whose set of patients often schedule behavioral assessments, suggesting that the PCP is thorough in her wellness assessments and may even provide follow-up scheduling assistance for her patients. Therefore, these patients are more likely to schedule behavioral assessments regardless of true risk, and thus have lower *observed* risk signals on average.

The two dashed lines in Figure 4 represent the sex-specific lines of best fit through the data. These are obtained via non-linear least squares estimation, specifying an exponential functional form to ensure estimates above 0. Appendix Table A6 presents the estimated model fit coefficients for both males and females. The vertical intercept at one of the sex-specific curves provides an estimate of population mean ADHD risk, μ_θ. These values are reported in the figure and again in Table 4 which presents the full set of parameters estimates. I estimate population mean ADHD risk for males to be $\mu_m = 0.130$ and mean ADHD risk for females to be $\mu_f = 0.064$, with bootstrap standard errors of 0.039 and 0.053 respectively.

Table 4: Model Parameter Estimates

	Male	Female	Difference
Pop. Mean Risk μ_θ	0.130	0.064	0.066***
	(0.037)	(0.047)	
Pop. Risk Dispersion σ_θ	0.384	0.375	0.009
	(0.067)	(0.067)	
Patient Stigma c_θ	0.459	0.463	-0.004*
	(0.041)	(0.053)	
Signal Quality ρ_θ	0.351	0.408	-0.057***
	(0.054)	(0.062)	
Diagnostic Threshold τ_θ	0.258	0.398	-0.140***
	(0.019)	(0.030)	

Note: Standard errors in parenthesis based on 1000 bootstrapped patient samples. Differences calculated as female parameter estimate subtracted from male parameter estimate with significance based on paired T-test difference in means using bootstrap sample estimates. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4 presents the full set of results for male and female patients. This includes population mean risk estimates from the first stage as well as the remaining model parameters obtained via sex-specific maximum likelihood estimation of equation 9.

The differences in model parameters in Table 4 are informative about which mechanisms lead to ADHD diagnostic disparities and in what direction. As discussed in Section 3.2, diagnostic disparities between male and female patients can be attributed to differences in prevalence, stigma, diagnostic uncertainty, and diagnostic thresholds. The results in Table 4 suggest that each of these channels play an important role in explaining diagnostic disparities. First, the population mean risk for males is significantly higher than that of females (difference of 0.066), which increases diagnostic disparities through both the patient symptom discussion channel and through higher physician posterior risk values. This finding is directionally consistent with the medical literature that shows ADHD prevalence in males is likely higher than in females AHRQ (2011). Second, males and females have similar mental health stigma suggesting patient preferences do not drive differences in ADHD diagnosis rates. I find that physicians put more weight on female ADHD risk signals ($\rho_f > \rho_m$), which by construction measures the overlap between patient symptoms and DSM-V symptoms. This finding is consistent with the results in Bruchmüller et al. (2012), who show that physicians are more likely to follow DSM-V criteria when diagnosing female patients and rely on heuristics for male patients. Finally, I find that physicians use lower diagnostic threshold for male patients ($\tau_m < \tau_f$). This means that physicians are more likely to diagnose a male patient over a female patient with identical posterior ADHD risk, suggesting that perceived cost of missed-diagnosis is higher for male patients.

7 ADHD Diagnosis Simulations

In the previous section I presented estimates of model parameters and discussed differences across patient sex. I now use the model parameters in Table 4 and run ADHD diagnosis simulations using the structural model in Section 3.1. This allows me to (1) identify how much of the diagnostic disparity can be attributed to the different mechanisms of diagnosis, and (2) provide estimates of both missed and mis-diagnosis for male and female patients based on the DSM-V definition of ADHD.

In Table 5, I show how well the simulated model matches key moments of the observed data, both overall and for male and female subset of patients. The simulated model does extremely well at

predicting average diagnosis rates (D) and symptom discussion rates (Q). It slightly over-estimates mean ADHD match signals ($x|Q$) and conditional diagnosis rates ($D|Q$), but differences are small.

Table 5: Observed verses Simulated Rates

	Observed			Simulated		
	Total	Male	Female	Total	Male	Female
ADHD Dx. (D)	0.063	0.088	0.037	0.062	0.086	0.038
Discuss Sx. (Q)	0.170	0.195	0.144	0.169	0.196	0.143
ADHD match ($x|Q$)	0.305	0.320	0.305	0.314	0.319	0.308
Cond. Dx. ($D|Q$)	0.357	0.433	0.250	0.365	0.439	0.265

Note: This table presents average values across patients of ADHD diagnosis, Symptom discussion, ADHD risk signals, and conditional diagnosis. Means are calculated for full set, and subset of patients according to patient sex. Those in the Observed columns are based on the EHR data and those in the Simulated columns based on diagnostic simulations using model parameters in Table 4 and model outlined in Section 3.1.

7.1 Mechanisms of Diagnostic Disparities

To show how the various mechanisms contribute to the ADHD diagnostic disparity measure, I analyze simulated diagnosis rates under counterfactual scenarios that place restrictions on the source of sex-specific variation. The results of this analysis are presented numerically in Table 6 and visually in Figure 5.

The first row of Table 6 shows no diagnostic disparity (1.00:1), in which parameters are restricted to be identical across patient sex. The second panel shows the results when only ADHD risk distribution parameters μ_θ and σ_θ are allowed to vary. The remaining parameters are held constant at either the male or female estimates. When only ADHD underlying risk varies by patient sex, the diagnostic disparity increases from 1.00:1 to 1.57:1 or 1.63:1 depending on which estimates the remaining parameters are held at. This represents 45.2% or 50.0% of the observed disparity, suggesting that only half of the diagnostic disparity can be attributed to differences in underlying prevalence between male and female patients.

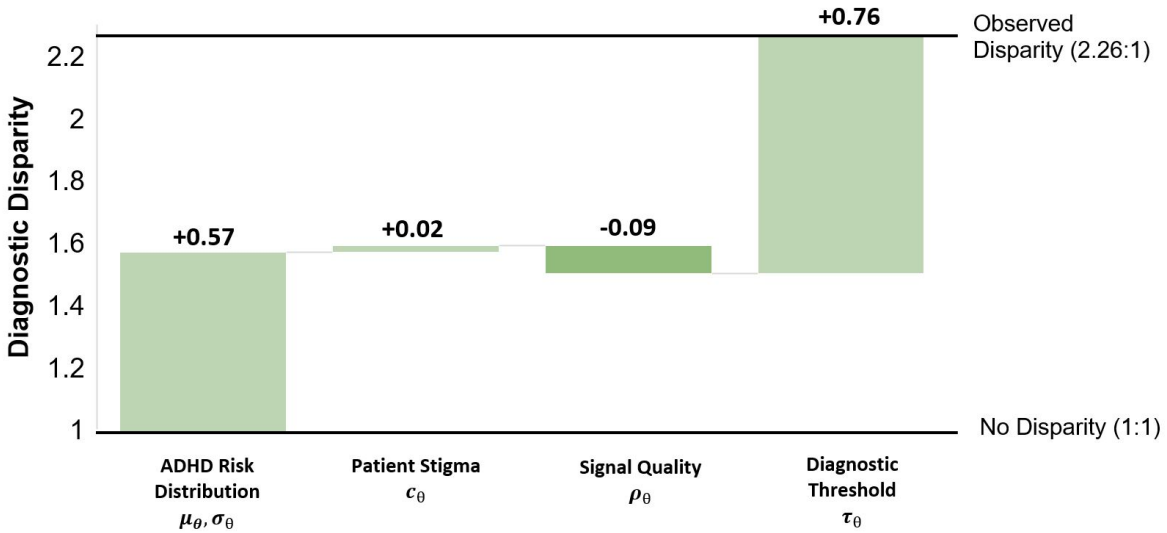
When patient stigma is also allowed to vary by patient sex (Patient Contribution panel in Table 6), diagnostic disparities increase only slightly, suggesting that the very little of the diagnostic disparity can be attributed to variation in patient preferences across sex. Finally, to analyze the physician decision making contribution, I relax the restrictions on signal quality and physician thresholds sequentially. The differences in signal quality actually reduces the diagnostic disparity, but this is more than made up for in differences in diagnostic thresholds that explain between 56.3% to 60.3% of the diagnostic gap between male and female patients.

Table 6: Disparity Mechanism Contribution

	Diagnostic Disparity	Disparity Effect	Relative Contribution
No Disparity	**1.00**	-	-
Prevalence Contribution			
ADHD Risk Distribution: μ_θ and σ_θ			
at Male estimates	1.57	+0.57	45.2%
at Female estimates	1.63	+0.63	50.0%
Patient Contribution			
Patient Stigma: c_θ			
at Male estimates	1.59	+0.02	1.6%
at Female estimates	1.65	+0.02	1.6%
Physician Contribution			
Signal Quality: ρ_θ			
at Male estimates	1.50	-0.09	-7.1%
at Female estimates	1.55	-0.10	-7.9%
Diagnostic Thresholds: τ_θ			
at Male estimates	2.26	+0.76	60.3%
at Female estimates	2.26	+0.71	56.3%
Observed Disparity	**2.26**	+1.26	100%

Note: This table presents results from diagnostic simulations with sequential restrictions the model parameters. Rows show which parameters are varied, starting with no variation, and adding variation until all parameters are at estimated value. Diagnostic disparity is calculated as simulated male diagnosis rate divided by simulated female diagnosis rate. Disparity effect calculates the net difference from disparities in previous simulation. Relative Contribution calculated as disparity effect divided by total effect of 1.26.

Figure 5: Cumulative Disparity Mechanism Effect



Note: This figure shows the cumulative effect of each mechanism in explaining ADHD diagnostic disparity. Values come from Column 2 of Table 6, where parameter restrictions in simulations are set at male parameter values.

7.2 Estimates of Mis(sed) Diagnosis

As less than one-half of the ADHD diagnostic disparity can be attributed to true underlying ADHD prevalence differences, it remains that the remaining difference in ADHD diagnosis rates between male and female patients is unwarranted, at least according to the DSM-V guidelines. In this section, I use the model estimates along with the DSM-V defined definition of ADHD, to determine the extent of over and under diagnosis for both male and female patients. I define ADHD diagnostic errors as any variation in the diagnosis decision from the DSM-V defined definition of ADHD. A patient is *misdiagnosed* if the symptoms in her doctor notes do not meet the criteria in the DSM-V, yet she is diagnosed with ADHD by a physician. On the other hand, a patient's condition is *missed* if her described symptoms meet the DSM-V diagnostic criteria for ADHD, yet she does not receive a clinical diagnosis.

I find that physicians are making diagnostic errors in the sense that they do not follow the DSM-V guidelines. However, I preface this section by noting that from a medical standpoint, these errors may be justified. It may be the case that the DSM-V definition of ADHD is outdated or too terse, in which case physician discretion and variation from these guidelines is warranted. In fact, this is a common consensus among psychologists. A recent *Psychology Today* article written by a psychiatrist and pediatric neurologist states: "...behavioral issues that patients face are not so easily cataloged as medical books (including the DSM) might tempt a person to believe. The DSM is just a tool designed to categorize human behavior in a clinically useful way — but it is inherently artificial, and must be taken with a grain of salt (and preferably used by a well-trained clinician with plenty of practical experience and good judgment)" (Cheyette and Cheyette, 2020). Therefore, while I estimate rates of ADHD diagnostic errors, I remain agnostic about the resulting policy implications. It may be that physicians require more training in recognizing ADHD, or it may imply a need to adjust DSM-V definition of this condition. I leave the interpretation and implications of the following estimates to the medical profession.

Defining 'True' ADHD

Recall that while ADHD risk, v_i, presents itself on a continuum, a DSM-V defined definition of ADHD is binary by construct. I follow the DSM-V definition of ADHD and assume that a child has ADHD, $S_i = 1$, if and only if they meet all the requirements for diagnosis outline in the DSM-V. In other words, $S_i = \mathbb{1}(v_i > \overline{v})$ where $\overline{v}$ is implicitly defined as the DSM-V minimum requirement

of diagnosis, which by definition does not vary by patient sex.

Thus far I have remained agnostic about the value of $\overline{v}$ as it is not necessary to estimate diagnostic disparities across patient sex. However, to examine inaccuracies in diagnosis according to guidelines, it is important to use this value. For purposes of classifying ADHD diagnostic inaccuracies, I refer back to the DSM-V guidelines for ADHD that requires a patient meets 6 (or more) of the 9 specified ADHD symptoms (see Table 1). As ADHD signal, x_i, and therefore ADHD risk, v_i, measures the fraction of DSM-V symptoms experienced by patient i (see construction of x_i in Section 4.3), the DSM-V defined minimum threshold is $6/9 = .66$, corresponding to a $\overline{v}$ value of 0.66.

Estimates of DSM-V defined errors

Using $\overline{v} = 0.66$ along with population risk distribution parameters, μ_θ and σ_θ, I can simulate DSM-V defined ADHD prevalence rates by patient sex. Combining this with the full diagnosis model allows me to simulate the extent of over/under diagnosis across patient sex, as well as potential sources of error. I present the results of this simulation exercise in Table 7.

Table 7: Mis(sed) Diagnosis Simulations

	% Misdiagnosed	% Missed Diagnosis
Panel A: Total		
DSM-V defined ADHD: 6.97%		
Clinical Dx: 6.19%		
Overall	4.55	5.34
Patient Effect	-	4.14
Physician Effect	4.55	1.20
Panel B: Male		
DSM-V defined ADHD: 8.34%		
Clinical Dx: 8.56%		
Overall	6.39	6.14
Patient Effect	-	4.91
Physician Effect	6.39	1.24
Panel c: Female		
DSM-V defined ADHD: 5.60%		
Clinical Dx: 3.79%		
Overall	2.72	4.53
Patient Effect	-	3.73
Physician Effect	2.72	1.16

Note: This table shows simulated diagnosis rates D_i based on simulated diagnosis decisions and S_i based on simulated risk v_i larger than $\overline{v} = 0.66$. DSM-V defined ADHD is then proportion of children with $S_i = 1$ and Clinical Dx is proportion with $D_i = 1$. Misdiagnosis is defined by $S_i = 0, D_i = 1$ and Missed Diagnosis by $S_i = 1, D_i = 0$. Within column 3, the Patient Effect denotes set of patients with $S_i = 1, D_i = 0, Q_i = 0$ and Physician Effect set of patients with $S_i = 1, D_i = 0, Q_i = 1$. Panel A shows results for full sample, and Panel B and C broken up by patient sex subsamples.

There are three key takeaways from this table. First, comparing simulated DSM-V defined ADHD and clinical diagnosis decisions allows me to estimate net over and under diagnosis. Based on the simulation results in Panel A, approximately 7% of children meet the diagnostic criteria for ADHD, 4.55% of children are misdiagnosed, and 5.34% of children have ADHD but do not receive clinical diagnosis, resulting in a small net under-diagnosis estimate of 0.79%. Both the prevalence rates and error rates are heterogeneous across patient sex, as shown in Panels B and C. 8.34% of males and 5.6% of females meet the DSM-V defined requirement for ADHD, however on net, 0.22% of males are over-diagnosed compared to a net female under-diagnosis of 1.81%.

Referencing back to the classification matrix in Figure 1 Panel B, whereas there is only way to obtain a false positive ($FP_{Q=1}$), there are two potential sources of false negatives, one from a patient error ($FN_{Q=0}$) and one from a physician error ($FN_{Q=1}$). The second insight from Table 7 is the break down of these sources of errors. Panel A shows that 4.14% of children who have ADHD do not seek care from a physician, and 1.2% of children who have ADHD do seek care, but do not receive an appropriate diagnosis. This suggests that the majority of missed diagnosis is due to patients not appropriately seeking care as opposed to physicians missing a warranted diagnosis. This is true in the male and female sub-samples as well.

Finally, the simulations also provide insight on the impacts of heterogeneous physician decision making. In total, physicians are much more likely to misdiagnose than to miss a diagnosis (4.55% to 1.20% in Panel A). While this is true for both male and female patients, the relative difference in error rates in heterogeneous. The rate of missed diagnosis from a physician error is similar for both male and female patients (1.24% and 1.16% respectively); however, the misdiagnosis rate is much higher for male patients than female (6.39% and 2.72% respectively). These findings suggest that the perceived cost of *missed diagnosis* is larger than cost of *misdiagnosis* on average, where the relative cost is larger for male patients.

7.3 Economic Impact of ADHD Diagnostic Errors

Both types of these ADHD diagnostic errors are costly. A *misdiagnosis* can lead to excess medical and educational spending, along with indirect costs associated with treatment side effects and psychological stigmatization. A *missed diagnosis* can lead to decreased educational attainment (Currie and Stabile, 2006) and lower earnings/employment (Fletcher, 2014).

Thus far, the literature that monetizes the economic impact of ADHD has only provided evidence

of incremental costs associated with having an ADHD *clinical diagnosis*. Adjusting to 2019 U.S. dollars, Doshi et al. (2012) estimate annual economic of ADHD diagnosis to be between $168 to $312 billion dollars. However, this is based on diagnosis rates alone and therefore does not consider how much of the estimated costs come from misdiagnosis, and additionally excludes costs associated with missed diagnoses. In this section, I derive an estimate of the economic impact associated with both over and under diagnosis of ADHD using per-person cost components from the literature combined with rates of ADHD diagnostic errors shown in Table 7.

I use ADHD cost estimates from Table 2 (pg 996) in Doshi et al. (2012), which provides monetizes ranges of ADHD economic impact based on a comprehensive literature review. Importantly, the authors break up the over-all economic impact into different categories: health care, productivity/income loss, justice system, and education. As these are based on those with an ADHD clinical diagnosis, I make some assumptions about how these categories carry over unto those with ADHD diagnostic inaccuracies. Specifically, I assume that those who are misdiagnosed incur the health care costs (e.g. through treatment and follow-up visits) and the educational costs, which includes special education services used for children with diagnosed ADHD. I assume that those whose ADHD is missed (missed diagnosis), do not have to incur the direct medical spending for treatment in childhood, but as a result experience the productivity and income loss as adults.

Appendix Table A8 provides the relevant table from Doshi et al. (2012), and highlights the costs I consider for this analysis. I make necessary adjustments to the 'Per-Person Incremental Costs' column in order to re-monetize accounting for inflation and rates of diagnostic errors. I first inflate costs to 2019 U.S. dollars based on CPI and medical care CPI from the U.S. Bureau of Labor Statistics.[28] As the literature has not fully explored differential costs by patient sex, I must assume per-person costs within each category are the same for males and females. Therefore, differences in costs across sex comes from differences in diagnostic error rates. I determine the number of population incurring cost by multiplying the diagnostic error rates in Table 7 with 2019 Child Population by sex estimates from the U.S. Census. Finally, I calculate national incremental costs of ADHD diagnostic errors by multiplying the per-person cost by the number of children in each ADHD diagnostic error category. Table 8 presents the results from this back-of-the-envelope calculation.

[28]Health care costs are adjusted using medical care component of CPI and all others using CPI-U.

Table 8: Cost of ADHD Mis(sed) Diagnosis

	Population Incurring Cost	Per-Person Cost of Error	National Cost of Errors (billions)
Misdiagnosis	3,355,893	\$3402-\$8989	\$11.4-\$30.2
Males	2,384,024		\$8.11-\$21.43
Females	971,869		\$3.31-\$8.74
Missed Diagnosis	3,909,343	\$12,593-\$22,156	\$49.2-\$86.6
Males	2,290,752		\$28.8-\$50.8
Females	1,618,591		\$20.4-\$35.9

Note: Table reflects estimates of costs associated with ADHD diagnostic errors, separated into misdiagnosis and missed diagnoses, by patient sex. All costs reported in 2019 U.S. dollars. Population counts based on 2019 Census population estimates and rates of errors in Table 7. Per-Person costs from Doshi et al. (2012), and are the same for males and females within each category.

The annual economic impact of ADHD diagnostic errors is \$60.6-\$116.8 billion U.S. dollars, with \$11.4-\$30.2 due to excess medical and education spending for those misdiagnosed, and \$49.2-\$86.6 due to productivity and income loss following a missed diagnosis. My findings suggest that the national estimate provided by Doshi et al. (2012) underestimates the cost of ADHD by at least \$37.8-\$56.4 billion dollars.[29]

The per-person cost of missed diagnosis is about 3 times larger than the cost of misdiagnosis. This suggest that physicians may in fact be optimal in their average diagnostic threshold, which recall reflects a higher relative cost of missed diagnosis. However, as these per-person cost estimates were not broken down by patient sex, this table does not yet provide support for why physicians use significantly lower thresholds for males, suggesting higher relative per-person costs for male patients. Future research analyzing per-person excess expenditures by patient sex is warranted.

Interestingly, while females are under-diagnosed more than males on net, almost two-thirds of the economic impact is incurred by males. This comes from the important break-down of net diagnosis rates in Table 7 which shows males are both misdiagnosed and missed-diagnosed more than females. The former is attributed to higher diagnostic uncertainty for male patients (i.e. lower signal quality estimate ρ_θ), and the later comes from lower diagnostic thresholds. This demonstrates the extreme importance of examining both *misdiagnosis* and *missed diagnosis* as opposed to net rates of errors, and further exploring the differential impact across patient sex.

[29] This underestimate is determined by subtracting cost of misdiagnosis and adding cost of missed diagnosis to Doshi et al. (2012) estimates.

Finally, I am conservative in the choice of cost categories for each error so that these estimates would underestimate the true cost of misclassified ADHD. For example, they do not include the potential spill-over effects of misdiagnosis (Persson et al., 2020) or the productivity loss of family members (Birnbaum et al., 2005). They also do not reflect health costs associated with over-use of stimulants, or personal costs through hindered peer relationships and self-esteem (Coghill, 2010). On the other hand, these estimates would overstate the true cost of diagnostic errors if physicians use the DSM-V with discretion and adjust the definition to fit each patient accordingly. As estimates of diagnostic errors are significant, understanding how the DSM-V defines errors and how additional indirect costs affect children and society is an important area of future research.

8 Conclusion

Attention Deficit Hyperactivity Disorder is the most diagnosed child mental health condition in the United States. Yet, recent research presents evidence of improper ADHD diagnosis decisions and documents heterogeneous national diagnosis rates by patient sex, with 14.8% of males diagnosed with ADHD and 6.7% of females. In this paper I combine structural modeling, selection estimation techniques, and text analysis procedures, to explore mechanisms of ADHD diagnosis and show how these contribute to the significant diagnostic disparity between male and female patients.

I develop a model of ADHD diagnosis, composed of three district stages, to demonstrate how both patient and physician factors contribute to the ADHD diagnosis rate. Importantly, each stage of the model depends on an unobservable patient ADHD risk value, coming from a sex-specific risk distribution to account for variation in true ADHD prevalence between male and female children. My model highlights four key mechanisms of ADHD diagnostic disparities: (1) differences in patient preference to seek mental health care, (2) varying rates of diagnostic uncertainty, (3) heterogeneous physician preferences for ADHD diagnosis, and (4) underlying differences in the true prevalence of ADHD between boys and girls.

I estimate the sex-specific structural parameters using electronic health records and clinical doctor notes. I address the lack of necessary observable mental health variables by using clinical doctor note data combined with natural language processing and text analysis techniques to create proxies for two mental health related variables- the patient decision to discuss behavioral symptoms, and an ADHD match signal measuring how closely the behavioral assessment aligns with DSM-V

criteria. In a first stage selection approach, I use quasi-exogenous variation coming from primary care physician quality to estimate population mean ADHD risk for males and females. I then estimate the remaining parameters via maximum likelihood estimation, separately by patient sex. I find that males have higher ADHD prevalence, higher diagnostic uncertainty, and lower diagnostic thresholds than their female counterparts.

I then use these estimated parameters and structural model to simulate ADHD diagnosis rates in order to (1) identify the mechanism contribution in explaining ADHD diagnostic disparities and (2) provide estimates of over and under diagnosis for males and females. The raw ADHD male-to-female diagnostic disparity is 2.26:1. I show that less than half of this disparity can be explained by differences in true underlying prevalence rates. The remaining difference is due to variation in physician decision making across male and female patients.

Using the DSM-V definition of ADHD, I find that males are slightly over-diagnosed and females under-diagnosed on net. This can be broken down into heterogeneous rates of misdiagnoses (6.4% males and 2.7% females) and missed-diagnoses (6.1% males and 4.5% females). I conduct back-of-the-envelope calculations and estimate an annual economic impact of ADHD diagnostic error in the range of $60.6 to $116.8 billion U.S. dollars. The cost of *missed* diagnosis is more than 3 times larger than the estimated cost of *mis*-diagnosis. This finding suggests that physicians may be acting optimal by internalizing these costs (on average) and setting diagnostic thresholds lower than that specified by the DSM-V guidelines. However, I also find that physicians use lower diagnostic thresholds for male patients than female patients with identical ADHD risk, implying that physicians perceive the relative cost of type II vs type I diagnostic error to be higher for male patients. The clinical support for these heterogeneous costs should be explored further, and perhaps even warrant a re-evaluation of how ADHD is defined in the DSM-V, noting its associated effects on male and female clinical diagnoses and subsequent treatment.

I also decompose ADHD missed diagnosis into rates due to physicians decisions and rates due to patient decisions. On the demand side, I find that approximately 80% of under-diagnosis for both male and females can be attributed to patient stigma- i.e. not seeking mental healthcare when warranted. As missed diagnoses are extremely costly, this suggests a potential policy response through targeted mental health education to reduce associated stigmas.

It is important to note the limitations of the results in this paper. First, the construction of mental health variables using clinical texts relies on the assumption that physicians accurately and

thoroughly document each patient encounter (or at least are consistent in documentation practices across patient sex). While it is difficult to test this assumption directly, possible directions for future work involve additional text analysis procedures that determine documentation similarity both across physician types and within physician across patients. The second limitation of this study comes from the quasi-exogenous variation in symptom discussion rates due to primary care physician quality which I use to estimate population mean ADHD risk for males and females. In an ideal (econometric) setting, patients would be assigned to PCPs randomly. However, in application, families may select their primary care physician. If this choice is correlated with unobserved ADHD risk, then my estimations of population mean risk will be biased, though the direction depends on the sign of this correlation which is ex-ante ambiguous. Although not feasible in this paper due to data constraints, an alternative source of exogenous symptom discussion costs would be primary care physician time pressures. If a patient has a wellness visit on a "busy" day, the PCP may be less able to provide a thorough evaluation and thus less likely to suggest follow-up mental health care. This idea is motivated by the recent work by Freedman et al. (2018), and should be explored further, specifically in relation to mental health care.

Finally, I emphasize that the suggested policy responses and diagnostic error estimates are potentially sample-location dependent. The in-sample ADHD diagnosis rate of 6.3% is lower than the national average during this time period, suggesting a potential under-estimate of misdiagnosis and over-estimate of missed diagnoses when compared to national rates. This is likely due to the fact that a large portion of the population in Arizona is of Hispanic ethnicity, and research suggests a significantly lower diagnosis rate for this group coming from cultural biases (Morgan et al., 2013). This is consistent with my large estimates of patient stigma, which may be lower in more nationally representative samples.

This paper addresses an understudied yet important area of research: mental health diagnostic errors and disparities. While the results are limited in external validity, the proposed model and methods are general enough to be applied to a variety of other applications. Mental health conditions are costly to both the individual and society, and thus understanding mechanisms across additional geographies, other disparity groups (e.g. by race/ethnicity), and alternative mental health conditions is an important area of future research.

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):3730–64.

AHRQ (2011). Attention Deficit Hyperactivity Disorder: Effectiveness of Treatment in At-Risk Preschoolers; Long-Term Effectiveness in All Ages; and Variability in Prevalence, Diagnosis, and Treatment. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm.

AHRQ (2019). National Healthcare Quality and Disparities Report. Available at: https://www.ahrq.gov/sites/default/files/wysiwyg/research/findings/nhqrdr/2018qdr-final-es.pdf.

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.

Anwar, S. and Fang, H. (2012). Testing for the role of prejudice in emergency departments using bounceback rates. *The BE Journal of Economic Analysis & Policy*, 13(3).

Arnold, D., Dobbie, W. S., and Hull, P. (2020). Measuring racial discrimination in bail decisions. Technical report, National Bureau of Economic Research.

Balsa, A. I., McGuire, T. G., and Meredith, L. S. (2005). Testing for statistical discrimination in health care. *Health Services Research*, 40(1):227–252.

Birnbaum, H. G., Kessler, R. C., Lowe, S. W., Secnik, K., Greenberg, P. E., Leong, S. A., and Swensen, A. R. (2005). Costs of attention deficit–hyperactivity disorder (adhd) in the us: excess costs of persons with adhd and their family members in 2000. *Current medical research and opinion*, 21(2):195–205.

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.

Chamberlain, G. (1986). Asymptotic efficiency in semi-parametric models with censoring. *journal of Econometrics*, 32(2):189–218.

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

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.

Chandra, A. and Staiger, D. O. (2010). Identifying provider prejudice in healthcare. Technical report, National Bureau of Economic Research.

Cheyette, B. and Cheyette, S. (2020). The relationship between autism spectrum disorder and adhd. *Psychology Today*.

Clemens, J. and Rogers, P. (2020). Demand shocks, procurement policies, and the nature of medical innovation: Evidence from wartime prosthetic device patents. Technical report, National Bureau of Economic Research.

Coghill, D. (2010). The impact of medications on quality of life in attention-deficit hyperactivity disorder. *CNS drugs*, 24(10):843–866.

Cronin, C. J., Forsstrom, M. P., and Papageorge, N. W. (2020). What good are treatment effects without treatment? mental health and the reluctance to use talk therapy. Technical report, National Bureau of Economic Research.

Cuddy, E. and Currie, J. (2020). Rules vs. discretion: Treatment of mental illness in us adolescents. Technical report, National Bureau of Economic Research.

Currie, J., Kleven, H., and Zwiers, E. (2020). Technology and big data are changing economics: Mining text to track methods. Technical report, National Bureau of Economic Research.

Currie, J. and MacLeod, W. B. (2017). Diagnosing expertise: Human capital, decision making, and performance among physicians. *Journal of labor economics*, 35(1):1–43.

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.

Currie, J. and Stabile, M. (2006). Child mental health and human capital accumulation: the case of adhd. *Journal of health economics*, 25(6):1094–1118.

Cutler, D., Skinner, J. S., Stern, A. D., and Wennberg, D. (2019). Physician beliefs and patient preferences: a new look at regional variation in health care spending. *American Economic Journal: Economic Policy*, 11(1):192–221.

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.

Epstein, J. N. and Loren, R. E. (2013). Changes in the definition of adhd in dsm-5: subtle but important. *Neuropsychiatry*, 3(5):455.

Fletcher, J. M. (2014). The effects of childhood adhd on adult labor market outcomes. *Health economics*, 23(2):159–181.

Freedman, S., Golberstein, E., Huang, T.-Y., Satin, D., and Barrie Smith, L. (2018). Docs with their eyes on the clock? the effect of time pressures on primary care productivity. Working Paper, Indiana University.

Furzer, J., Dhuey, E., Laporte, A., et al. (2020). Adhd misidentification in school: Causes and mitigators. Technical report.

Gowrisankaran, G., Joiner, K. A., and Léger, P.-T. (2017). Physician practice style and healthcare costs: evidence from emergency departments. Technical report, National Bureau of Economic Research.

Heckman, J. (1990). Varieties of selection bias. *The American Economic Review*, 80(2):313–318.

Herrerias, C. T., Perrin, J. M., and Stein, M. T. (2001). The child with adhd: Using the aap clinical practice guideline. *American Family Physician*, 63(9):1803.

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

Jensen, P. S., Hinshaw, S. P., Swanson, J. M., Greenhill, L. L., Conners, C. K., Arnold, L. E., Abikoff, H. B., Elliott, G., Hechtman, L., Hoza, B., et al. (2001). Findings from the nimh multimodal treatment study of adhd (mta): implications and applications for primary care providers. *Journal of Developmental & Behavioral Pediatrics*, 22(1):60–73.

Knapp, M., King, D., Healey, A., and Thomas, C. (2011). Economic outcomes in adulthood and their associations with antisocial conduct, attention deficit and anxiety problems in childhood. *Journal of mental health policy and economics*, 14(3):137–147.

Layton, T. J., Barnett, M. L., Hicks, T. R., and Jena, A. B. (2018). Attention deficit–hyperactivity disorder and month of school enrollment. *New England Journal of Medicine*, 379(22):2122–2130.

Marquardt, K. (2020). Identifying physician practice style for mental health conditions. available at: `www.kellimarquardt.com`.

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.

Persson, P., Rossin-Slater, M., and Qiu, X. (2020). Family spillover effects of misdiagnosis: The case of adhd. draft available soon at: `https://web.stanford.edu/~perssonp/research.html`.

Rushton, J. L., Fant, K. E., and Clark, S. J. (2004). Use of practice guidelines in the primary care of children with attention-deficit/hyperactivity disorder. *Pediatrics*, 114(1):e23–e28.

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.

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.

Appendices

A Additional Tables and Figures

Table A1: Additional Patient Demographics

	Mean	Std. Dev.	Minimum	Maximum
Medicaid	0.540	0.498	0	1
Private Ins.	0.419	0.493	0	1
White-Non Hispanic	0.345	0.475	0	1
Black-Non Hispanic	0.070	0.255	0	1
Hispanic	0.489	0.500	0	1
Psych Physician	0.069	0.254	0	1
Age	10.312	3.535	5	18
# of Appt.	3.248	4.043	1	85
# of Physicians	1.927	1.491	1	15
# Yrs. in Sample	1.693	0.891	1	4
N	11,397			

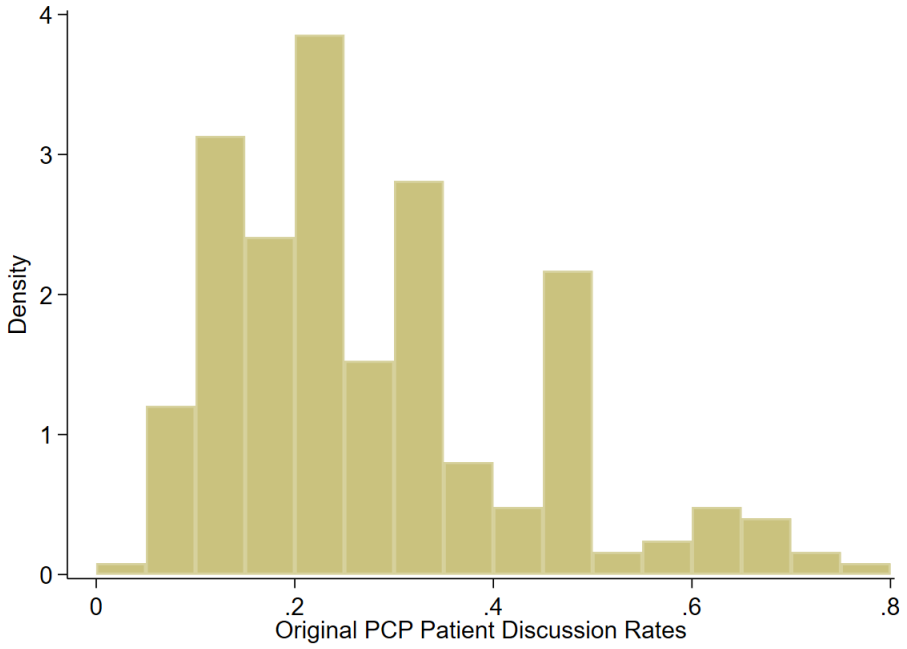
Note: Table presents summary statistics for full set of patients included in sample. Psych Physician indicates whether patient has an appointment with psychiatrist. Age is based on age at last appointment in sample. # of physicians indicates the number of unique physicians the patient sees over sample period. Alternative insurance category includes 'self-pay' and 'other'.

Table A2: Male/Female Difference in Observables

	Male	Female	Difference
Full Sample			
Age	10.165	10.463	-0.298***
Medicaid	0.535	0.545	-0.010
Private Ins.	0.421	0.417	0.005
White	0.346	0.345	0.001
Hispanic	0.483	0.496	-0.013
N	5,786	5,611	
Symptom Discussion Sample			
Age	10.227	11.757	-1.529***
Medicaid	0.519	0.497	0.022
Private Ins.	0.437	0.477	-0.040*
White	0.412	0.447	-0.035
Hispanic	0.466	0.440	0.026
N	1,120	803	

Note: Table presents sex-specific means and difference in means for full sample and symptom discussion subsample ($Q_i = 1$). Significance based on two-sample T test with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure A1: Variation Eventual Discussion Rates- *Original PCP*



Note: This figure plots histogram of eventual patient symptom discussion rates across *original primary care physicians* in the sample. Patient Discussion rates calculated as the fraction of patients of each original PCP that eventually appear in the electronic health record with $Q_i = 1$.

Table A3: Test of First Stage PCP Relevance

Wald Test for PCP Fixed-Effect Significance			
	Total	Male	Female
	(1)	(2)	(3)
Wald Chi-Squared Test Statistic	1773***	1352***	1378***
Degrees of Freedom	205	146	128
Patients	8934	4363	4258
Mean Discussion Rates	.173	.198	.150
Patient Controls			
Male, Age, Psych Referral, Medicaid, Private Ins., Hispanic, White, Appt. Type, # of Physicians, #of Appts. Year FE			

Note: This table shows results from Wald Chi-squared joint test of significance on original PCP fixed effects in a probit regression of patient symptom discussion on set of patient controls and PCP fixed effects. Results shown for three separate regressions based on total sample, male sample, and female sample respectively. The coefficients and construction of patient controls presented in Table A4. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A4: First Stage Symptom Discussion Coefficients

	Total (1)	Male (2)	Female (3)
Male	0.120***		
	(0.036)		
Age	0.021***	-0.005	0.053***
	(0.005)	(0.007)	(0.008)
Psych Referral	2.039***	2.051***	2.389***
	(0.111)	(0.153)	(0.217)
Medicaid	-0.065	-0.136	0.099
	(0.099)	(0.135)	(0.159)
Private Ins.	-0.159	-0.212	-0.011
	(0.101)	(0.139)	(0.162)
Hispanic	0.119*	0.103	0.139
	(0.055)	(0.076)	(0.085)
White	0.365***	0.309***	0.442***
	(0.060)	(0.083)	(0.092)
Behavioral Appt.	2.352***	2.478***	2.198***
	(0.185)	(0.256)	(0.286)
Wellness Appt.	0.058	-0.034	0.110
	(0.074)	(0.107)	(0.108)
# of Phys.	-0.010	-0.029	0.020
	(0.024)	(0.035)	(0.036)
# of Appt.	-0.247***	-0.257***	-0.266***
	(0.033)	(0.048)	(0.049)
$\mathbb{1}(2014)$	0.292***	0.300***	0.313***
	(0.034)	(0.049)	(0.049)
$\mathbb{1}(2015)$	0.258***	0.276***	0.263***
	(0.034)	(0.050)	(0.049)
$\mathbb{1}(2016)$	0.136***	0.138*	0.133*
	(0.038)	(0.055)	(0.055)
PCP Fixed Effects	Y	Y	Y
Observations	8934	4363	4258

Note: This table shows patient control coefficients from probit regression of patient symptom discussion on demeaned patient controls and PCP fixed effects. Results shown for three separate regressions based on total sample, male sample, and female sample respectively. All controls are based on the average (or max) across patient appointments prior to and *including* symptom discussion appointment. All controls demeaned using sample average. Behavioral Appt, indicator based on previous other mental health diagnoses, and Wellness Appt. indicator based on broad appointment type categories. Psych Referral indicates whether patient was seen by a psychiatrist during first symptom discussion visit. Year fixed effect included to control for changes in mental health trends over time. Robust standard errors in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A5: Test of PCP Selection

	Full (1)	Male (2)	Female (3)
Male	0.002		
	(0.003)		
Age	-0.000	-0.000	-0.000
	(0.000)	(0.001)	(0.000)
Medicaid	-0.005	-0.006	-0.005
	(0.007)	(0.007)	(0.007)
Hispanic	0.001	0.005	-0.003
	(0.004)	(0.004)	(0.005)
White	0.005	0.011	-0.002
	(0.004)	(0.006)	(0.005)
N	8929	4463	4466
Joint F-Test (p-value)	.849	.320	.813

Note: This table presents results from patient level regression of leave-one-out PCP eventual symptom discussion rates on demeaned patient demographics. Leave-one-out PCP discussion rates determined using data from all other patients of the patient's PCP and calculated as the percent of other patients from each PCP that eventually discuss mental health symptoms in the data (i.e. percent with $Q_i = 1$). Robust standard errors in parenthesis, clustered at the PCP level. The table also reports the p-value associated with a joint test of patient demographic significance. Regression results provided for full sample and subsambles by patient sex.

Table A6: Exponential Fit of ADHD Risk Signal on Exogenous Symptom Discussion Probability

	Male (1)	Female (2)
$\widehat{\alpha_0}$	0.364	0.376
	(0.013)	(0.018)
$\widehat{\alpha_1}$	-1.030	-1.778
	(0.200)	(0.324)
N	878	668
Adj. R-sq.	0.811	0.759
Fitted μ_θ	0.130	0.064

Note: This table shows coefficients from non-linear least squares regression with exponential functional form: $Y = \alpha_0 exp(\alpha_1 X)$ where Y is the observed ADHD risk signal for patients who discuss symptoms and X is the predicted probability of symptom discussion coming from quasi-exogenous variation in patient primary care physician quality. This fits the data in text Figure 4 for male and female subset of patients separately. Fitted μ_θ denotes the intercept at 1 (i.e. $\mu_\theta = \widehat{\alpha_0} exp(\widehat{\alpha_1})$). Standard errors in parenthesis.

Table A7: Independent Disparity Effects

	Diagnosis Rates		Diagnostic
	Male	Female	Disparity
Baseline Differences	**0.086**	**0.038**	**2.26**
Panel A: Prevalence			
ADHD Risk Distribution: μ_θ and σ_θ			
at Male estimates	0.086	0.061	1.41
at Female estimates	0.055	0.038	1.44
Panel B: Patient Preferences			
Patient Stigma: c_θ			
at Male estimates	0.086	0.039	2.23
at Female estimates	0.848	0.038	2.24
Panel c: Physician Decision-Making			
Signal Quality: ρ_θ			
at Male estimates	0.086	0.035	2.46
at Female estimates	0.091	0.038	2.40
Diagnostic Thresholds: τ_θ			
at Male estimates	0.086	0.057	1.50
at Female estimates	0.059	0.038	1.55

Note: This table reflects diagnosis rates from a model simulation exercise that restricts variation in only one set of model parameters. The simulated sex-specific diagnosis rates are reported in columns 1 and 2 with the ratio in column 3. For reference, Panel A presents simulations that restrict ADHD risk distribution parameters to be equal for male and female patients and all other parameters allowed to vary and equal their estimated values in text Table 4. I include diagnosis rates when equalization is based on male estimate and female estimate. Panel B restricts variation in patient stigma, and Panel C restricts variation in physician parameters, signal quality and diagnostic thresholds respectively.

Table A8: ADHD Cost Estimate Table: (Doshi et al., 2012)

TABLE 2 National Incremental Costs of Attention-Deficit/Hyperactivity Disorder (ADHD) by Cost Category and Age Group

Cost Category	Age Group of Patients with ADHD	Number of Studies	Age Range across Studies	Population corresponding to Age Range[31,33]	ADHD Prevalence for Age Range	Other Multipliers[a]	Population Incurring Cost	Per-Person Incremental Cost, 2010 U.S. Dollars	National Incremental Cost, 2010 U.S. Dollars (Billions)
Health care									
Health care (patient)	children and adolescents	9	0–21	92,140,979	7.2%[3]	—	6,634,150	$621[37]–$2,720[23]	$4.12–$18.04
Health care (patient)	adults	6	18–64	194,296,087	4.4%[8]	—	8,549,028	$137[(NS)46]–$4,100[42]	$1.17–$35.05
Health care (family)	children and adolescents	2	0–18	74,181,467	7.2%[3]	2.92	15,595,912	$1,088[10]–$1,658[25]	$16.97–$25.86
Health care (family)	adults	1	19–44	108,305,787	4.4%[8]	2.92	13,915,128	$1,051[10]	$14.62
								subtotal	$37B–94B
Productivity and income losses									
Income losses (lower wages)	adults	1	19–25	30,433,583	4.4%[8]	—	1,339,078	$(3,744)[26]	$(5.01)
Income losses (unemployment and lower wages)	adults	1	18–64	194,296,087	4.4%[8]	—	8,549,028	$10,532[34]–$12,189[34]	$90.04–$104.20
Productivity losses (patient)	adults	6	18–64	194,296,087	4.4%[8]	67.6%	5,779,143	$209[45]–$6,699[41]	$1.21–$38.71
Productivity losses (family)	children and adolescents	2	0–18	74,181,467	7.2%[3]	2.0, 67.6%	7,221,121	$142[10]–$339[25]	$1.03–$2.45
Productivity losses (family)	adults	1	19–44	108,305,787	4.4%[8]	1.0, 67.6%	3,221,447	$174[10]	$0.56
								subtotal	$88B–$141B
Education	children	1	3–4	8,182,210	5.5%[3]	—	450,022	$12,447[35]	$5.60
	children and adolescents	2	5–18	58,480,960	7.2%[3]	—	4,210,629	$2,222[23]–$4,690[36]	$9.36–$19.75
								subtotal	$15B–$25B
Justice system	adolescents	1	13–17	21,238,249	9.3%[3]	—	1,975,157	$267[(NS)23]	$0.53
	adults	1	18–28	47,550,861	4.4%[8]	—	2,092,238	$1,204[24]–$2,742[24]	$2.52–$5.74
								subtotal	$3B–$6B
								total	$143B–$266B

Note: B = billions; NS = difference was not statistically significant in the original study.
[a]Figures used in "Other Multipliers" are described in the Method section.

Note: This table based on screenshot of Table 2 in Doshi et al. (2012) which reflects estimates ADHD diagnostic costs decomposed into categories and age. The highlighted estimates are the ones used in back of the envelope cost calculations in text Section 7.2. The "+" denotes which costs are used for misdiagnosis and the "-" denotes which costs are used for missed diagnoses.

B Variable Construction using Clinical Texts

B.1 Constructing Mental Health Discussion Variable: Q_i

In this appendix I present the Machine Learning Algorithm used to construct a proxy for the mental health discussion indicator, Q_i. This closely follows the *Text Analysis Appendix* in Clemens and Rogers (2020).

I first break the appointment level data into a labeled and un-labeled subsets, where i denotes patient and j denotes appointment. The labeled set is determined by icd9 codes where an appointments receive a positive label ($Q_{ij} = 1$) if the appointment is associated with an icd9 diagnosis related to mental health (Q1 Codes in table B9). An appointment receives a negative label ($Q_{ij} = 0$) if the appointment is associated with an idc9 diagnosis related to physical ailments (Q0 Codes in table B9). To ensure that there is no overlap with patients in both groups, I restrict the negative labeled set to only those patients that never receive a mental health diagnosis during the sample period. The un-labeled set contains all appointments in which there is no associated diagnoses or appointments with ambiguous icd9 codes that could be related to either mental or physical health concerns (e.g. abdominal pain can be associated with anxiety or a virus). This hand coded separation procedure results in 40,917 appointments and 14,092 patients in the labeled set (31,716 appointments with $Q_{ij} = 0$ and 9,200 with $Q_{ij} = 1$) and 105,054 appointments of 28,403 patients in the un-labled set.[30]

Q0 Codes	Q1 Codes
034, 055, 058, 078, 079, 080, 111, 113, 171, 192, 204, 250, 251, 273, 277, 278, 283, 287, 288, 289, 363-383, 389, 390, 462, 463, 466, 473, 474 478, 486, 488, 493, 494, 529, 537, 599, 600, 608, 612, 682, 683, 693, 697, 703, 707, 709, 710, 715, 719, 720, 725, 728, 729, 730, 733, 734, 744, 760, 781-791, 849, 907, 919, 920, 960	293-319, 331, V11, V15, V40, V41, V61, V62, V71, V79

Table B9: ICD-9 Labeled Dataset Codes

I next prepare the doctor notes for feature extraction. This includes traditional text preprocessing procedures: replace contractions, remove special characters and stop words, conversion

[30]These sample sizes are larger than the estimation sample as I choose not to make any sample restrictions in building the machine learning algorithm. Within the estimation sample, 6,711 appointments of 2658 patients are un-labeled.

to lowercase and stemming. For both computational and prediction purposes, I consider only 41 features: note length, relative frequency of top 20 predictive words in the positive labeled set, and relative frequency of top 20 predictive words in the negative labeled set. I determine these top predictive words by their "tf-idf" value in a constructed document term matrix.[31]

- Positive-label word stems: *school, mother, behavior, parent, report, current, social, disord, anxieti, famili, examin, activ, treatment, therapi, sleep, adhd, psychotherpi, tablet, feel, diagnosi*
- Negative-label word stems: *pain, fever, list, care, cough, blood, exam, address, rash, skin, return, vaccin, left, rang, bilater, ml, resid, hour, puls, record*

For cross-validation, I split the labeled data into a training and test set using 90-10 split. Using the training set, I define a random forest learner and tune hyperparameters using random grid search with hold-out re-sampling. I use false discovery rate (FDR) as the objective measure for hyperparameter tuning. The main hyperparameters and their tuned values are: number of trees to grow (ntree=398), number of variables at node split (mtry=3), and maximum number of observations in terminal nodes (nodesize=3).

Using the tuned hyperparameters, I then train the model on the training set, again specifying false discovery rate as the objective measure. The confusion matrix applied to the test set is presented below, with false discovery rate of 0.03487.

	Predicted-0	Predicted-1
True-0	3,153	28
True-1	129	775

Before analyzing the final model predictions, I look for issues with *context specificity*, or "limitations on a model's validity outside of its training set" (Clemens and Rogers, 2020). I take a random sample of 96 notes from the unlabeled dataset, read the unprocessed notes, and determine the appropriate hand label for mental health discussion using own discretion. Then using the training random forest algorithm, I obtain the model's predictions for these notes. I specify a probability threshold of 0.5. The confusion matrix is presented in the table below. 88 of the notes were cor-

[31] A document term matrix consists of documents i as rows, words j as columns, and matrix elements t_{ij} representing frequency of word j in document i. The tf-idf value is defined as $\frac{t_{ij}}{T_i} ln(\frac{D}{D_j})$ where T_i denotes the number of terms in document i, D denotes the total number of documents, and D_j denotes the number of documents with term j.

rectly determined via the random forest algorithm. 7 notes were incorrectly specified, with only 1 non-mental health related appointment receiving a positive label.

	Predicted-0	Predicted-1
True-0	70	1
True-1	6	18

I consider this performance and validity to be satisfactory, and thus apply the trained random forest algorithm to the full un-labeled set of appointments to obtain the complete set of predictions for mental health discussion. Approximately 9% of appointments receive a positive predicted label. Results at the patient level are shown in text Table 3.

B.2 Constructing Patient ADHD risk Signal: x_i

The content in this appendix is taken directly from Marquardt (2020) and included in this paper for ease of the reader. I present only the general algorithm here and direct readers to the full paper for further intuition and examples.

Procedure

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.[32] 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 run the following procedure for each doctor note, indexed by i, and each DSM-V symptom, indexed by s. The index, s, can reference a single symptom (a specific symptom in Table 1), a group of symptoms (list 1 or list 2 in Table 1), or the entire DSM-V text for a given mental health condition (all of text in Table 1).

Step 1: Text Cleaning & Pre-Processing

Traditional text is messy. This first step cleans the text and prepares it for mathematical analysis,

[32]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.

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 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.

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.

C Econometric Appendix

C.1 Physician Diagnostic Threshold

In this appendix, I present a physician utility framework that results in a risk-threshold diagnosis decision rule, where the threshold is a function of physician perceived cost of diagnostic errors.[33]

Let physician utility be defined by:

$$u_i|\theta = \begin{cases} -1 & \text{if } D_i = 0,\ S_i = 1 \\ -\beta_\theta & \text{if } D_i = 1,\ S_i = 0 \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

The utility of correct diagnoses are normalized to 0 so that physicians receive *disutility* from errors. With utility of missed diagnoses ($D_i = 0, S_i = 1$) standardized to -1, β_θ captures the potentially sex-specific disutility of misdiagnosis *relative* to missed diagnoses.

The physician chooses $D_i = 0$ or $D_i = 1$ in order to maximize his expected utility, where expectation is based on the posterior probability of $S_i = 1$. Let $p(x, \theta)$ denote this probability. $p(x, \theta)$ is expressed in equation 14, and follows from posterior ADHD risk in (6) and the DSM-V defined minimum diagnostic requirement, $\overline{v}$.

$$p(x, \theta) = Pr(v_i|x > \overline{v}) = \Phi\left(\frac{\rho_\theta x + (1 - \rho_\theta)\mu_\theta - \overline{v}}{\sigma_\theta\sqrt{1 - \rho_\theta^2}}\right) \tag{14}$$

The doctor will choose to diagnose a patient with ADHD if the expected utility of $D_i = 1$ is larger than the expected utility of $D_i = 0$. Based on the utility function (13), $E[u_i|D_i = 1, \theta] = -\beta_\theta(1 - p(x, \theta)) + 0(p(x, \theta))$ and $E[u_i|D_i = 0, \theta] = -1(p(x, \theta)) + 0(1 - p(x, \theta))$.

[33]This is similar to the utility in Chan et al. (2019), but with variation in cost across patient sex as opposed to variation across physicians.

Assuming misdiagnoses are costly (i.e. $\beta_\theta > 0$), then the doctor will choose $D_i = 1$ iff

$$\begin{aligned} & E[u_i|D_i = 1, \theta] \geq E[u_i|D_i = 0, \theta] \\ \implies & -\beta_\theta + \beta_\theta p(x, \theta) \geq -p(x, \theta) \\ \implies & p(x, \theta) \geq \frac{\beta_\theta}{1 + \beta_\theta} \end{aligned}$$

Plugging in equation (14) for $p(x, \theta)$, a physician will diagnose if $\Phi\left(\frac{\rho_\theta x + (1-\rho_\theta)\mu_\theta - \overline{v}}{\sigma_\theta\sqrt{1-\rho_\theta^2}}\right) \geq \frac{\beta_\theta}{1+\beta_\theta}$. Re-writing with posterior ADHD risk mean on the right-hand side results in the following sex-specific threshold value:

$$\tau_\theta = \overline{v} + \sigma_\theta\sqrt{1 - \rho_\theta^2}\Phi^{-1}\left(\frac{\beta_\theta}{1 + \beta_\theta}\right)$$

For $\beta_\theta \in (0, 1)$, $\Phi^{-1}\left(\frac{\beta_\theta}{1+\beta_\theta}\right) < 0$ which implies $\tau_\theta < \overline{v}$. In words, physicians will use thresholds lower than the DSM-V defined definition so that they diagnose patients on the margin of meeting ADHD criteria. Intuitively, this suggests that physicians view missed diagnoses as costlier than misdiagnosis, which is consistent with $\beta_\theta \in (0, 1)$ in (13).

On the other hand, $\beta_\theta > 1$ implies $\tau_\theta > \overline{v}$. In this case, physicians will use higher thresholds and will *not* diagnose patients on the margin of meeting ADHD criteria. This suggests that physicians view misdiagnosis as costlier than missed diagnosis, which is consistent with $\beta_\theta > 1$ in (13).