Mis(sed) Diagnosis: Physician Decision Making and ADHD

Kelli Marquardt*
Federal Reserve Bank of Chicago†
The University of Arizona

February 2023

Abstract

While the presence of disparities in mental healthcare is well documented, the mechanisms of such disparities are less understood. In this paper, I develop and estimate a structural model of diagnosis for a prevalent child mental health condition, Attention Deficit Hyperactivity Disorder (ADHD). The model incorporates both patient and physician influences to highlight various mechanisms of mental health diagnosis and sources of disparities. Using electronic health record data and novel natural language processing techniques applied to doctor note text, I estimate gender-specific model parameters and decompose the male:female ADHD diagnostic difference of 2.5:1 observed in the data. Counterfactual simulations show that only 33% of this diagnostic difference can be explained by underlying symptom prevalence, with the remainder driven by differences in diagnostic thresholds. I find that physicians view *missed diagnosis* to be costlier than *misdiagnosis*, especially for their male patients, and I discuss reasons why this may be economically warranted.

Keywords: ADHD, child mental health, health disparities, physician decision-making
JEL classification: I14, D81, C5

*I am grateful for helpful feedback from my dissertation committee: Gautam Gowrisankaran, Keith Joiner, Ashley Langer, Juan Pantano, and Tiemen Woutersen. I also thank Keith Ericson, Bo Honore, Christoph Kronenberg, Jessamyn Schaller, and conference and seminar participants at various universities and institutions. 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. Author email: kmarquardt@frbchi.org

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

1 Introduction

Healthcare disparities are traditionally defined as differences in health treatment and outcomes across population groups in ways that cannot be explained by underlying health status or preference differences.[1] While overall health disparities have declined, mental health disparities show the opposite trend (AHRQ, 2019). Determining whether differences in mental health diagnosis rates across groups is an unwarranted disparity, however, requires knowledge of true prevalence rates. This is especially difficult to isolate in the mental health context due to the subjective nature of diagnosis along with concerns of social stigma and selection into reporting and/or treatment.

In this paper, I develop a model of mental health diagnosis that depends on both underlying health status, patient preferences, and physician decision-making under uncertainty. I restrict my focus to a common child mental health diagnosis, Attention Deficit Hyperactivity Disorder (ADHD), which has a particularly salient diagnosis rate difference by gender. I then use electronic health record data to estimate gender-specific model parameters, which allows me to quantify the male/female ADHD diagnostic disparity and isolate mechanisms contributing to differences in diagnosis rates.[2]

Approximately 10% of children are diagnosed with ADHD, and males are diagnosed and treated 2 to 3 times more frequently than females. 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. Both *missed* and *mis*-diagnoses are costly, including lower productivity and human capital accumulation for untreated ADHD

[1] See the annual *National Healthcare Quality and Disparities Reports*, mandated by U.S. Congress in accordance with the Healthcare Research and Quality Act of 1999.

[2] Within the medical community, it remains an open question as to whether the difference in ADHD prevalence stems from biological (sex) or social/cultural (gender) factors. In reference to ADHD prevalence differences in general, Hinshaw (2018) writes: "All-biological or all-cultural perspectives are therefore reductionist and short-sighted." To be consistent within this paper, I refer to differences in male and female model parameters and outcomes as gender-specific rather than sex-specific differences.

and harmful side-effects from over-treatment.[3]

My model has three distinct stages to reflect how the mental health diagnosis decision is made. In the first stage, patients (or rather their caregivers) decide whether or not to schedule a behavioral assessment with a diagnosing physician. This is a function of underlying unobserved symptom severity in addition to mental healthcare utilization costs. 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 symptom match is above a gender-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.

Taken as a whole, the model highlights four key mechanisms of mental health diagnosis that can potentially vary by patient gender and therefore contribute deferentially to observed diagnostic differences. These key mechanisms are: (1) underlying differences in the true prevalence of ADHD symptoms between male and female children, (2) patient preferences/costs of seeking mental health care, (3) varying rates of diagnostic uncertainty, and (4) heterogeneous physician preferences/costs for ADHD diagnosis.

I estimate the model parameters and empirically analyze the male/female ADHD diagnostic gap using data derived from electronic health records from 2014 to 2017 provided by a large healthcare system in Arizona. The dataset includes over 35,000 pediatric visits for approximately 11,000 patients. In the raw data, 7% of males and 3% of females are diagnosed with ADHD, implying a male-to-female ADHD diagnostic difference of roughly 2.5:1. This gap persists even after controlling for a variety of patient

[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." Further, (Doshi et al., 2012) estimate the annual economic impact of ADHD diagnosis at 168-312 billion U.S. dollars (inflated to 2019 $ with CPI).

observables, supporting the need for a structural model and estimation approach.

I first apply novel natural language processing and machine learning techniques to clinical doctor note text as a way to construct mental health related variables necessary for model estimation. Specifically, I determine whether patients receive a behavioral assessment 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 the national diagnostic guidelines for ADHD.

I then use the constructed mental health variables and clinical diagnoses to estimate the underlying parameters of the 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 a behavioral assessment 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 risk-adjusted referral rates, providing an exclusion restriction that allows identification of patient costs from scheduling a behavioral assessment (mental health utilization costs). This also allows me to obtain selection-adjusted estimates of the population mean ADHD risk for males and females via extrapolations of the observed ADHD-match signals on quasi-exogenous behavioral assessment propensity. This exogenous extrapolation approach is similar to the methods proposed by Arnold et al. (2022), who measure racial discrimination in judge bail decisions.

Finally, I recover the remaining model parameters with a method of moments approach that leverages variation in 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 analyzing differences in diagno-

sis rates by patient gender 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 and patient selection, the mean diagnosis rates for each gender 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 gender-specific ADHD risk obtained in the initial selection stage.

Counterfactual diagnostic simulations using model parameter estimates show that only one-third of the observed ADHD diagnostic difference between male and female patients can be attributed to differences in the underlying ADHD risk distribution, with the rest explained by variation in physician decision-making across patient gender. 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 symptoms when making a diagnosis decision. I also find that physicians view *missed diagnosis* to be more costly than *misdiagnosis* for male patients, denoted by lower male diagnostic thresholds. This difference in diagnostic thresholds by gender is inconsistent with clinical guidelines, yet explains about two-thirds of the gap in male/female ADHD diagnosis rates. I argue that this use of gender-specific thresholds may be economically warranted as males are more likely to express the externally costlier symptoms of ADHD and females are more likely to already receive treatment from internalizing comorbid mental health conditions like anxiety or depression (Hinshaw et al., 2022).

These results add to the existing literature exploring the potential for ADHD diagnostic errors. For example, in the health economic literature, multiple 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; Persson et al., 2021). 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; 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., Bruchmüller et al., 2012). My paper adds to this existing literature by estimating a structural model needed to decompose the underlying sources that contribute to the male/female diagnostic difference and quantify how much of this diagnostic gap aligns with medical guidelines. Results from the model simulation exercises can also help guide where policies might best focus efforts to reduce sources of medically-unwarranted diagnostic differences.

More broadly, my paper 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., 2022), structural models of physician decision-making under uncertainty (e.g., Abaluck et al., 2016; Currie and MacLeod, 2017; Chan et al., 2022), and identification of physician prejudice (e.g., Balsa et al., 2005; Chandra and Staiger, 2010; Anwar and Fang, 2012).

Importantly, this extant 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 health care (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 stigma plays a potentially large role in determining mental healthcare utilization. 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 disparities in mental health diagnosis. Second, this 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 traditional medical testing. My paper innovates to address this challenge by using clinical doctor note data and text analysis techniques to construct a proxy for ADHD symptom match based on clinical 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 in economic research (see Gentzkow et al., 2019; Currie et al., 2020, and citations therein). In this paper, I combine machine learning methods outlined in Clemens and Rogers (2020) with text analysis methods proposed in Marquardt (2022) 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 interview notes that inform agent decision-making, especially those in which true outcomes cannot be observed directly.

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 a reduced form analysis and observational comparisons. I also describe the the machine learning and natural language processing techniques used to extract important variables from clinical doctor notes. In Section 5, I outline the empirical strategy and parameter identification. Section 6 presents the model estimates and results from model simulations used to isolate and quantify mechanisms of disparities. I interpret these results and discuss both medical and economic implications 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). 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).[4] 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).[5]

Table 1: DSM-V Symptoms for ADHD

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

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

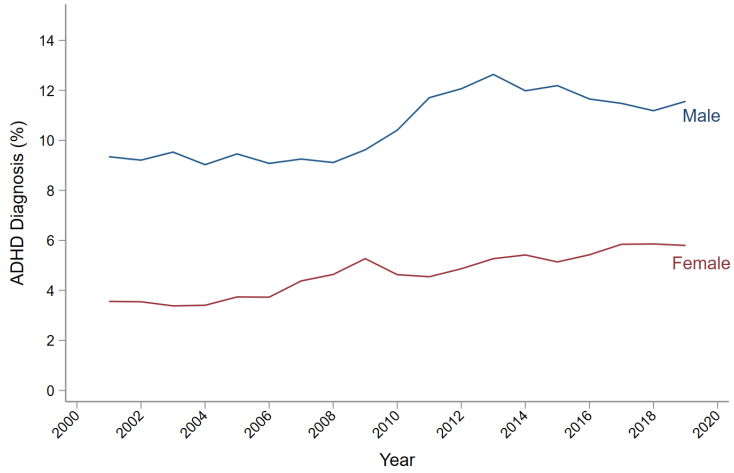
There are three possible types or presentations of ADHD: inattentive, hyperactive-impulsive, and combined type. Male children with ADHD are more likely to have the

[4]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.

[5]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).

Hyperactive or Combined type, and female children with ADHD are more likely to have the Inattentive type (Hinshaw et al., 2022). However, the clinical requirements for diagnosis are the same regardless of type; A child meets the clinical definition of ADHD if they experience 6 or more behavioral symptoms of a given sub-type 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.

Figure 1: National Trends in ADHD Diagnosis



Note: This figure plots the ADHD diagnosis rates for male and female children aged 5-17, based on data from the National Health Interview Survey (NHIS), 2000-2021. Yearly rates are weighted by the NHIS person sample weights, and figure plots the 3-year moving average.

Figure 1 displays the national trend in ADHD diagnosis rates for male and female children. These average diagnosis rates have increased over time, but the male/female diagnostic difference has remained relatively constant around 2.3:1. It is important to reiterate that while male and female children differ in which sub-type of ADHD they are most likely to experience, the DSM-V does *not* have different clinical requirements for these sub-types or conditional on patient gender. For both conceptual modeling and estimation purposes, this fact explicitly restricts differences in overall ADHD prevalence to come only from differences in number and severity of symptoms between male and female children. 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 patient must schedule and receive a behavioral assessment from a diagnosing physician. Scheduling this assessment is not required for all children, but may be encouraged based on feedback from teachers, guidance counselors, or primary care physicians during annual wellness checks.

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-V 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 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 finding, along with the institutional features of non-mandatory mental health screening, motivates the need for a structural model of ADHD diagnosis that incorporates these various elements of diagnosis in order to separately identify the key mechanisms leading to diagnostic differences.

3 Conceptual Framework

In traditional models of decision-making under uncertainty, deciding agents receive a noisy signal of the true state of the world, use the signal to update their prior beliefs, and make a decision to maximize utility. These types of models have been empirically estimated in healthcare settings (e.g., Anwar and Fang, 2012; Chan et al., 2022) in addition to other applications such as the judicial system (e.g., Arnold et al., 2022). What is missing from these models, however, is individual selection, which I show is an important mechanisms to understanding disparities in outcomes across patient groups, specifically in relation to mental health. In what follows, I present a model of ADHD diagnosis that pairs a physician decision-making under uncertainty model with a first-stage selection component that endogenizes the patient decision to seek mental health care (selection). I allow, but do not enforce, key model parameters to vary based on patient gender. I then discuss comparative statics to highlight the four potential mechanisms underlying ADHD diagnostic differences between boys and girls: true symptom prevalence, patient utilization costs, diagnostic uncertainty, and physician preferences.

3.1 Diagnosis Model with Endogenous Selection

The model is composed of three stages: patient selection, physician learning, and clinical diagnosis. In the first stage, patients choose to schedule a behavioral assessment if their ADHD symptoms outweigh any costs associated with mental healthcare utilization. Conditional on selecting into care, 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 likelihood. In the final stage, the physician will choose a diagnosis decision based on ADHD posterior risk and the costs they bear from making a diagnostic error. Underlying the model is a gender-specific ADHD risk distribution that captures differences in true prevalence rates. The model allows patient mental health utilization, physician preference thresh-

olds, and physician learning rates to vary by patient gender as a way to capture the varying components of mental health diagnostic disparities.

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 gender 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.

$$v_i \sim N(\mu_\theta, \sigma_\theta^2) \tag{1}$$

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 clinical diagnosis is binary by definition. 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, letting $S_i \in \{0, 1\}$ denote the true ADHD status, we have $S_i = 1(v_i > \overline{v})$ where $\overline{v}$ is the DSM-V defined minimum requirement for diagnosis, which by definition does not vary by patient gender.[6] Thus, differences in true ADHD prevalence between boys and girls depend only on differences in ADHD risk distribution parameters, with prevalence increasing in population mean risk, μ_θ.

Stage 1: Patient Choice to Schedule Behavioral Assessment

In the first selection stage of the model, the patient/parent must decide whether or not to schedule a behavioral assessment.[7] Parents will schedule a behavioral assessment if the child's behavioral symptoms outweigh any mental healthcare utilization costs,

[6] 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, but rather test if doctors use different thresholds based on patient gender, a practice that implies deviation from the official DSM guidelines.

[7] 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.

c_i, which includes a mean component, c_θ, and an idiosyncratic cost, $\varepsilon_i \mid v_i \sim N(0,1)$. Because health insurance typically covers behavioral assessments with little to no out of pocket expenditures, c_i includes non-monetary constraints (or conversely nudges) impacting the decision to schedule a behavioral assessment. This can include parent time constraints, distance to the nearest health center, recommendations from school teachers, or information obtained from primary care physicians during annual wellness visits. It may also include any stigma surrounding potential mental health diagnosis. In other words, c_i captures everything that impacts the decision to seek mental health care net of child symptom level, v_i. I allow for differences in the gender-specific mean utilization cost, c_θ, but do not enforce a difference empirically.

I assume the patient observes their costs, c_i, and their symptoms, v_i, but does not have enough medical information to know $\overline{v}$, thus motivating them to seek a professional opinion. Denoting Q_i as an indicator for behavioral assessment, I define $Q_i = \mathbb{1}(v_i > c_i)$. Equation (2) defines the gender-specific behavioral assessment rate, which follows from (1) and the assumption that $c_i = c_\theta + \varepsilon_i \perp\!\!\!\perp v_i$.

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

Stage 2: Physician Learning via Behavioral Assessment

I assume that the physician knows the gender-specific ADHD risk distribution, but does not know patient specific ADHD risk, v_i, nor the patient specific mental health utilization costs, c_i. Thus, the physician prior can be defined by (1) and is a function of ADHD risk distribution parameters μ_θ and σ_θ.[8]

If a patient chooses to schedule a behavioral assessment, the physician will learn about the patient specific ADHD risk, v_i. Through this process, the physician receives a noisy signal, x_i, of the true ADHD risk v_i, defined by equation (3). The signal is

[8]This assumption allows me to interpret the diagnostic threshold parameter τ_θ as physician preferences over diagnostic errors. In Appendix C.2, I discuss the benefits of this assumption and implications if it fails.

unbiased and correlated with the true state through $\rho_\theta \in (0,1)$. I allow correlation to vary by patient gender as a way to capture variation in diagnostic uncertainty coming from signal quality.[9]

$$\begin{pmatrix} v_i \\ x_i \end{pmatrix} \Bigg| \theta \sim N\left(\begin{pmatrix} \mu_\theta \\ \mu_\theta \end{pmatrix}, \begin{pmatrix} \sigma_\theta^2 & \rho_\theta \sigma_\theta^2 \\ \rho_\theta \sigma_\theta^2 & \sigma_\theta^2 \end{pmatrix} \right) \tag{3}$$

The physician then uses this information to update their belief of ADHD risk via a Bayesian updating process. After observing x_i the physician updates their prior, resulting in the posterior ADHD risk distribution defined in (4). Notice that the updated risk posterior mean is a weighted average of patient observed signal, x_i, and the physician prior risk mean, μ_θ, where the weight placed on the signal depends on the signal quality ρ_θ.

$$v_i \mid x_i \sim N\left(\left(\rho_\theta x_i + (1-\rho_\theta)\,\mu_\theta \right), \sigma_\theta^2 (1-\rho_\theta^2) \right) \tag{4}$$

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.

$$D_i \mid x_i, \theta = \mathbb{1}(v_i \mid x_i \geq \tau_\theta) \tag{5}$$

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, if physicians view *missed*

[9]This health signaling structure is very similar to that defined in Chan et al. (2022), but assumes that signal strength varies across patient types as opposed to physician types.

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 gender to capture potential differences in physician perceived cost of diagnostic errors.[10]

Using the physician posterior in equation (4), the probability a patient is diagnosed, conditional on behavioral assessment and received signal, is:

$$\Pr\left(D_i = 1 \mid Q_i = 1, x_i, \theta\right) = \Phi\left(\frac{1}{\sigma_\theta\sqrt{1-\rho_\theta^2}}\left(\rho_\theta x_i + (1-\rho_\theta)\,\mu_\theta - \tau_\theta\right)\right) \tag{6}$$

3.2 Mechanisms of Diagnosis and Diagnostic Disparities

Combining equations (2) and (6) yields the following gender-specific diagnosis rate:

$$\begin{aligned}\Pr\left(D_i = 1 \mid \theta\right) &= \Pr\left(D_i = 1 \mid Q_i = 1, x_i, \theta\right) \times \Pr\left(Q_i = 1 \mid \theta\right) \\ &= \underbrace{\Phi\left(\frac{1}{\sigma_\theta\sqrt{1-\rho_\theta^2}}\left(\rho_\theta x_i + (1-\rho_\theta)\,\mu_\theta - \tau_\theta\right)\right)}_{\text{Physician Diagnosis Rate}} \times \underbrace{\Phi\left(\frac{\mu_\theta - c_\theta}{\sqrt{1+\sigma_\theta^2}}\right)}_{\text{Patient Assessment Rate}}\end{aligned} \tag{7}$$

Diagnosis rates are a function of underlying prevalence, mental healthcare utilization costs, diagnostic uncertainty, and physician preferences/diagnostic thresholds. My structural model captures each of these elements via μ_θ, c_θ, ρ_θ, and τ_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.[11] This increase can be attributed to both the patient selection 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

[10] 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 true cost of diagnostic errors differ by patient gender, in which case the heterogeneous thresholds are justified and should not be considered "discrimination." In this paper, I refer to differences in τ_θ as differences in *perceived* cost of errors, and I discuss whether this is economically and even medically warranted in Section 7.

[11] 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.

the latter is due to higher physician prior beliefs. On the other hand, high values of patient utilization costs imply lower diagnosis rates because fewer 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).[12]

These population-group comparative statics map directly into mechanisms explaining diagnostic differences 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 utilization costs and diagnostic thresholds. Therefore, the ADHD diagnostic difference seen between males and females may be attributed to some combination of the following:

- higher male prevalence ($\mu_m > \mu_f$)
- higher signal strength for male patients ($\rho_m > \rho_f$)
- lower utilization costs for male children ($c_m < c_f$)
- lower diagnostic thresholds applied to male patients ($\tau_m < \tau_f$)

From a health care policy standpoint, it is essential to identify whether true prevalence is the driving factor of differing diagnosis rates, or if these other mechanisms contribute to diagnostic disparities. The direction and relative contribution of each mechanism is an empirical question which I explore in the remainder of this paper.

3.3 Empirical Approach Outline

To identify the mechanisms leading to different male/female diagnosis rates, I separately estimate the model parameters for both male and female patients: $(\mu_\theta, \sigma_\theta, c_\theta, \rho_\theta, \tau_\theta)$

[12] $\frac{\partial P(D_i=1|Q_i=1)}{\partial \rho} = \phi(\frac{\rho(x-\mu)+\mu-\tau}{\sigma(1-\rho^2)(1/2)})(\frac{x-\mu+\rho(\mu-\tau)}{\sigma(1-\rho^2)(3/2)})$. 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, this 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.

for $\theta \in \{m, f\}$. I use electronic health record data and estimate equation (7) separately for male and female sub-samples.

The variables required to estimate gender-specific diagnosis rates (7) are clinical diagnosis decision, D_i, behavioral assessment indicator, Q_i, ADHD match signal, x_i, and patient gender, θ_i. However, the only variables directly observed in the electronic health record are D_i (via associated ICD-10 codes) and patient gender, θ_i. Even though behavioral assessment, Q_i, and ADHD match signals, x_i, are not directly imputed into electronic health record systems, I show how both variables can be recovered from clinical doctor note text.

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 gender-specific population mean ADHD risk parameter, μ_θ. Because ADHD match signals are only observed for an endogenously selected sample, I recover this parameter using quasi-exogenous variation in scheduling costs following an approach outlined in Arnold et al. (2022). Once male and female population mean risk are estimated, the remaining parameters are identified and estimated from moments defined by behavioral assessment rates and the conditional diagnosis probit following equation (7). I further detail this estimation process in Section 5.

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.[13] I first exclude children younger than 5 years old, whose rates of ADHD diagnosis and treatment are very low and whose medical care requires peer-to-peer review and prior authorization. I then drop erro-

[13] A diagnosing physician is identified as one who diagnosed ADHD at least once during the sample period. There are 151 diagnosing physicians in the dataset.

neous encounters, encounters with insufficient documentation, and patients with missing demographic information. The remaining data encompass 35,793 unique patient encounters, for 10,950 unique patients. Patient characteristics include: age, gender, race/ethnicity, 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 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 the main associated ICD-9 or ICD-10 code corresponds to an ADHD diagnosis.[14] While the specific symptoms differ by sub-type of ADHD, the clinical requirements for diagnosis are the same (e.g. ≥ 6 symptoms). Therefore, I group together the different types of ADHD into a single diagnosis category, but appropriately adjust for the different symptom presentations when constructing the patient ADHD match signal, detailed in Section 4.2. Patient-level summary statistics are presented in Table 2.

Of the roughly 11,000 patients seen from 2014 to 2017, 5% have a clinical ADHD diagnosis. The in-sample ADHD diagnosis rate is slightly lower than the national average during this time period, but the male/female diagnostic difference is representative of national values.[15] Males are diagnosed with ADHD significantly more than females. The raw diagnostic difference is 2.56:1, with 7.13% of males receiving a clinical diagnosis but only 2.78% of females. On average, patients will be seen by two different physicians over an average of 3.3 appointments, and will have at least one appointment with someone other than their original primary care physician. In total, there are 303 primary care physicians and only 151 diagnosing physicians, who are mainly classified

[14]The ICD-9 codes include 314.00 and 314.01, and the ICD-10 codes include F90.0, F90.1, F90.2.

[15]This lower-than-average diagnosis rate is likely due to the fact that a large portion of the sample population is of Hispanic ethnicity (49.5%), and research shows a significantly lower diagnosis rate for this population (see Morgan et al., 2013). I discuss the generalizability and implications of this sample bias in Section 7.

Table 2: Summary Statistics

	Mean	Std. Dev.	Min.	Max.
ADHD Dx.	0.050	0.218	0	1
Male Dx.	0.071	0.257	0	1
Female Dx.	0.028	0.164	0	1
Male	0.507	0.500	0	1
Age	10.318	3.562	5	18
White	0.559	0.496	0	1
Hispanic	0.495	0.500	0	1
Medicaid	0.538	0.499	0	1
# of Physicians	1.937	1.507	1	15
# of Appt.	3.269	4.116	2	92
# of Appt. (not PCP)	1.386	1.266	1	21
# Yrs. in Sample	1.695	0.895	1	4
N Patients	10,950			
N Diagnosing Physicians	151			
N Primary Care Physicians	303			

Note: This table presents summary statistics for the full set of patients in estimation sample. # of physicians indicates the number of unique physicians the patient sees over sample period. # of Appt. (not PCP) denotes the total number of completed appointments where the designated provider was not the child's original primary care physician.

as being pediatricians or family medicine doctors.[16]

Table 3 presents reduced-form ADHD diagnostic regressions that control for any gender-specific differences in demographics and other healthcare utilization variables.[17] In all instances, male patients are significantly more likely to be diagnosed with ADHD than female patients. This analysis highlights the inability to explain the male-female diagnostic difference using only directly observable information in electronic health record (or claims-based) datasets.

However, 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

[16]The original primary care physician is defined as the specified PCP during the patient's first visit. This is binding for only a few patients that switch PCP's during the sample period. Only 7% of the diagnosing physicians are psychiatrists, psychologists, and/or behavioral specialists. This small percentage may be specific to the case of childhood ADHD where clinical diagnosis does not require psychiatric specialization.

[17]Appendix Table A1 shows that male patients are on average 4 months younger than female patients, though similar across all other observed demographics.

Table 3: Reduced Form ADHD Diagnostic Regressions

	(1)	(2)	(3)
Male	0.046***	0.046***	0.040***
	(0.004)	(0.004)	(0.004)
Added Patient Observables:			
Demographic Variables	N	Y	Y
Healthcare Utilization Variables	N	N	Y
Adj. R-squared	0.0292	0.0318	0.1131
N	10,950	10,950	10,950

Note: This table presents the estimated coefficient on patient gender from a OLS regression of ADHD clinical diagnosis on patient controls. Demographic Variables: age, insurance status, race/ethnicity. Health Care 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. Robust standard errors in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

decision. These are (1) Q_i, which is an indicator for whether a patient receives a behavioral assessment, 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.

4.1 Behavioral Assessment: Q_i

The electronic health record does not specifically indicate whether a behavioral assessment was conducted during the visit. 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 give a general outline of the procedure here and provide additional details in Appendix B.1.

I first take a subset of appointments in which the behavioral assessment indicator variable is known with almost certainty. This subset is constructed by assuming 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. The remaining appointments are considered 'unlabeled' due to either no associated diagnoses or appointments with ambiguous ICD-9 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 unlabeled set of appointments.

Using information from the clinical doctor notes in the labeled dataset, I next determine a set of machine learning model features. I consider 41 features: note length, the relative frequency of the top 20 'positive' label words, and the relative frequency of the top 20 'negative' label words. Figure 2 provides a visual of these features with a word cloud representation for both negative and positive behavioral assessment labels. As expected, the positive behavioral assessment label includes words related to behavioral symptoms such as: *school*, *social*, *behavior*, *family*, and *feel*. The negative behavioral assessment label includes words more related to physical rather than mental health concerns. These include words such as: *pain*, *fever*, *cough*, and *rash*.

Figure 2: Behavioral Assessment Indicator 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 set of appointments in order to predict whether behavioral symptoms were discussed during the appointment based

on the information in the clinical doctor note. I take the maximum of this prediction across patient encounters to obtain the patient-level behavioral assessment indicator Q_i used in model estimation.

The machine learning algorithm predicts that approximately 18% of children receive a behavioral assessment. 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). Table 4, presented in in the following section, compares behavioral assessment rate predictions by patient gender. Males are significantly more likely than females to schedule and receive a behavioral assessment, at 20.8% and 15.4% respectively.

4.2 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. The ADHD match signal, 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 an interview and documented in the clinical doctor note. With access to these clinical doctor notes, I construct a proxy for x_i using a natural language processing algorithm originally proposed in Marquardt (2022). Essentially, I calculate the overlap between symptoms in the DSM-V symptom criteria list (see Table 1) and symptoms in the collective doctor notes for a given patient, making necessary adjustments 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).[18]

As x_i is defined on the patient level, I first combine patient notes across encounters into a single document, keeping only those identified as behavioral assessments and occurring before or during initial ADHD diagnosis. I then calculate ADHD match signal, x_i, following the natural language processing algorithm proposed in Marquardt (2022), in which patient documents and DSM-V symptom requirements are compared using an Adjusted Bag-of-Words Model. I give a general outline of the procedure here and provide additional details in Appendix B.2.

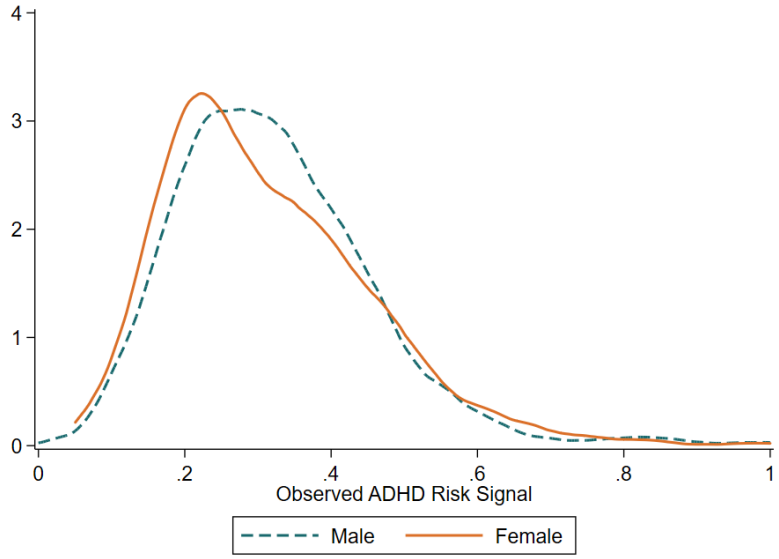
I first pre-process the clinical texts following standard medical text cleaning procedures (e.g., spell check, abbreviation replacement, and size reductions). I next group words according to contextual meaning which requires part-of-speech tagging and synonym replacement. Each document is then broken 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 matrix elements (i,k) are the "tf-idf" values indicating the relative frequency and importance of bi-gram k in document i.[19]

I then calculate the patient overall ADHD match signal by taking the cosine similarity measure between the patient document vector and the document vector constructed using official ADHD symptoms listed within *The Diagnostic and Statistical Manual of Mental Disorders*, (DSM-V). I do this for each sub-type of ADHD and define patient-level x_i as the similarity measure maximum across all three ADHD sub-types.

[18]In Appendix Tables B2 and B3, I show that male and female patients have similar doctor notes in terms of both note length and words predictive of high ADHD match. In Appendix C.2, I discuss the implications of this full documentation assumption. I argue that if the assumption fails equally for male and female patients, the diagnostic disparities and mechanism decomposition analysis remain unaffected.

[19]The "tf-idf" value is defined as $\frac{f_{ki}}{F_i} \times log(\frac{D}{D_k})$ where f_{ki} is frequency of bi-gram k in document i, F_i is length of document i, D is number of documents, and D_k is number of documents with bi-gram k.

Figure 3: Observed ADHD Match Signal by Patient gender



Note: This figure shows gender-specific distribution of constructed ADHD match signals x_i based on NLP techniques described in Section 4.2. This implicitly covers the set of patients with behavioral assessment, $Q_i = 1$, thus shows only a truncated distribution of the true population ADHD risk.

In total, the average signal match is is 0.318 with a standard deviation of 0.138. 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.[20] The signal for males is slightly larger than for females; however, the difference is only significant at the 10% level (see Table 4). Figure 3 presents a visual for the ADHD match signal distribution by patient gender. This provides only suggestive evidence of true prevalence differences as the plot represents the match for the (endogenous) set of patients that receive a behavioral assessment.

Table 4 presents summary statistics for the key variables needed to estimate the diagnosis model parameters. The top panel of Table 4 shows ADHD diagnosis rates and behavioral assessment rates for the full sample. While males do receive behavioral

[20]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 look only at differences between male and female patients.

Table 4: Mental Health Observational Comparisons

	Total	Male	Female	Difference
Full Sample				
ADHD Dx.	0.0499	0.0713	0.0278	0.0435***
	(0.218)	(0.257)	(0.164)	
Behav. Appt. (Q_i)	0.181	0.208	0.154	0.0538***
	(0.385)	(0.406)	(0.361)	
N	10950	5554	5396	
Behavioral Assessment Subsample ($Q_i = 1$)				
ADHD Dx.	0.275	0.343	0.180	0.1626***
	(0.447)	(0.475)	(0.385)	
ADHD Match Signal (x_i)	0.318	0.322	0.314	0.0078*
	(0.138)	(0.136)	(0.141)	
N	1987	1155	832	

Note: ADHD Dx. (D_i) based on ICD codes in EHR. Behavioral Assessment rates (Q_i) and ADHD Match Signal measures (x_i) are constructed using machine learning and natural language processing techniques outlined in Sections 4.1 and 4.2, respectively. Differences calculated as female means subtracted from male means, and significance based on two-sample T-test difference in means. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

assessments significantly more than females, this selection does not explain the entire diagnostic disparity as seen by the lower panel of Table 4. For those that receive a behavioral assessment, 34.3% of males will be diagnosed with ADHD and only 18% of females will be diagnosed. It is also unlikely that differences in symptom presentation fully explain the diagnostic gap as the difference in ADHD symptom match is only significant at the 10% level. This table provides suggestive evidence that the ADHD diagnostic difference is a function of selection, prevalence, *and* physician decision-making biases. I next outline the structural estimation approach which allows me to separate out the magnitude and direction of these underlying mechanisms.

5 Model Parameter Estimation and Identification

With data on ADHD diagnosis D_i, behavioral assessment Q_i, patient gender θ_i, and conditional ADHD match signal x_i, I estimate 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 parameter estimation procedure involves two steps where the first recovers the gender-specific population mean ADHD risk parameter, μ_θ. The remaining parameters are obtained by matching

a set of moments defined by behavioral assessment rates and components of conditional diagnosis probabilities following equation (7), 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 equation (7) but also intuitively following the comparative statics discussion in Section 3.2. Behavioral assessment rates are increasing in mean risk, μ_θ, and decreasing in patient utilization costs, c_θ. At the same time, conditional diagnosis rates are increasing in mean risk, μ_θ, and decreasing in diagnostic thresholds, 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 ADHD match signals are observed for all patients, one could estimate μ_θ using gender-specific sample average of ADHD match signals, $\frac{1}{N_\theta}\sum_{i\in N_\theta} x_i$. However, x_i is only observed for the subset of patients that receive a behavioral assessment. Because patients endogenously select into behavioral assessment according to unobserved ADHD risk, the average value of *observed* signals will over-estimate the population risk mean, as shown by equation (8).

$$E[x_i|Q_i=1] = E[x_i|v_i > c_i] = \mu_\theta + \underbrace{\rho_\theta\sigma_\theta\frac{\phi(\frac{c_i-\mu_\theta}{\sigma_\theta})}{1-\Phi(\frac{c_i-\mu_\theta}{\sigma_\theta})}}_{\text{upward bias}} \tag{8}$$

To recover unbiased estimates of mean population risk for males and females, I leverage quasi-exogenous variation in the otherwise unobserved utilization costs (c_i). To build intuition for this approach, consider a set of patients who, regardless of symptom levels, do not have any constraints (and may even be nudged) to scheduling a behavioral assessment. For low enough levels of c_i, the probability of behavioral assessment is approximately 1, so the patient will schedule a behavioral assessment and thus ADHD match signals, x_i, will be observed. Further, the bias term in (8) for these patients with low c_i goes to 0, and thus sample mean of x_i for patients with low utilization

costs (or conditionally high probability of behavioral assessment) provides an unbiased estimation of population mean risk, μ_θ.

As c_i is unobserved in application, I instead estimate individual propensity to schedule a behavioral assessment 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 behavioral assessment propensity through the unobserved patient utilization costs, c_i.

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

I use selection-adjusted referral rates of primary care physicians as the source of quasi-exogenous behavioral assessment propensity 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 latter is the PCP originally seen when the 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.[21]

To see how the original PCP identifier is correlated with behavioral assessment scheduling 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 their own concerns or suggestions from teachers. However, it is likely that they first bring up these concerns with their child's primary care physician who is trained to ask about patient school performance and behavioral concerns during annual wellness visits (American Academy of Pediatrics, 2011). If warranted by the response, PCPs may encourage the parent to schedule a follow-up appointment (either with themselves, with another pediatrician, or with a psychiatrist) so that a full behav-

[21] I use the *original primary care physician* as opposed to the *diagnosing physician* as the latter 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 requirement (b).

ioral 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 (referral rates).[22] To empirically verify that the PCP identifier meaningfully influences the patient probability of scheduling a behavioral assessment, I regress patient behavioral assessment indicator, Q_i, on patient controls and original PCP fixed effects, interacted with patient gender. I test for and find strong joint significance of PCP fixed-effects, results presented in Appendix Table A2.

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 of symptom development means that parents do not selectively chose primary care physicians after observing v_i. Second, there are approximately 300 *original* primary care physicians covering the patients in my sample, but only 24 of these ever diagnose ADHD.[23] So while PCPs may differ in the number of patients they refer or encourage to seek follow-up mental health care, they generally do not diagnose ADHD themselves, suggesting that patients set up behavioral assessments with alternative physicians, again implying no relation

[22] Appendix Figure A1 shows the variation across primary care physicians, with a histogram of the residualized leave-one-out PCP referral rates for both male and female patients.

[23] In majority of cases, PCPs will instead refer patients to other pediatricians in the health system.

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. I account for this possible selection-on-observables following the standard approach in the literature (e.g., Arnold et al., 2018, 2022). Specifically, I first regress behavioral assessment indicator on patient characteristics and year fixed effects, net patient gender. I then take the residual for each patient and average across all other patients assigned to the PCP. Using the leave-one-out residual allows me to measure PCP referral rates *relative* to PCPs who see (observably) similar patient mix. I then test whether patient gender and other patient demographics are correlated with PCP referral propensity by estimating an OLS regression of leave-one-out average PCP residuals on these patient characteristics. Appendix Table A3 presents these coefficients, and a F-test for joint significance suggests no correlation between patient gender (or age, race/ethnicity, insurance status) on primary care physician referral rates, thus providing support to condition (b).[24]

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 match signals and exogenous behavioral assessment propensity. Empirically, I first conduct a probit regression of behavioral assessment, Q_i, according to equation (9) where W_i includes a set of demeaned patient controls needed for selection-adjustment and Z_i denotes original PCP identifiers.

$$P(Q_i = 1) = \Phi\left(W_i\beta + \gamma_1 Z_i + \gamma_2 Male_i Z_i\right) \tag{9}$$

Next, I obtain exogenous behavioral assessment propensity, $\widehat{P_\theta(Q_i|Z_i)}$, by predicting behavioral assessment for each patient, holding W_i at sample means. With W_i

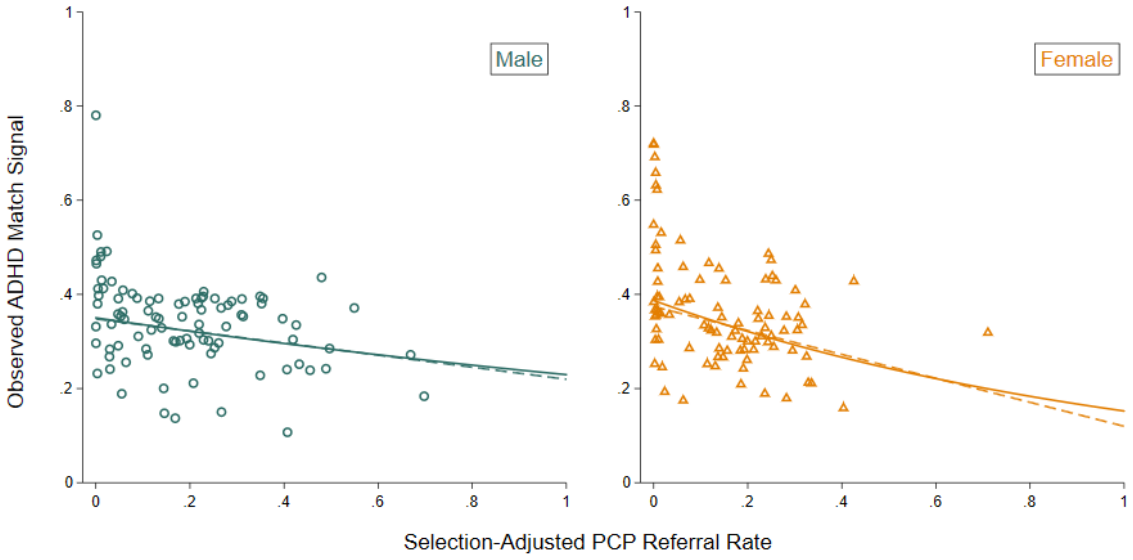
[24]There may still be concern that patients choose PCPs based on unobserved factors that are correlated with ADHD risk, leading to biased estimates of μ_θ. However, so long as these unobserved factors are independent of patient gender, the relative difference between male and female ADHD risk is unaffected. I further discuss the implications of this assumption in Appendix C.2.

demeaned, $\widehat{P_\theta(Q_i|Z_i)}$ is the selection-adjusted gender-specific PCP referral rate.

While there is significant variation in these adjusted referral rates, the maximum value is only about 0.75. In the absence of a PCP with adjusted referral rates near 1, values of μ_θ can be estimated via extrapolations of observed ADHD match signals on exogenous behavioral assessment propensity. Specifically, I fit a model of observed ADHD match signals, x_i, on $\widehat{P_\theta(Q_i|Z_i)}$ for both male and female patients, and obtain selection-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. (2022) and in line with the literature on identification in selection models (see Chamberlain, 1986; Heckman, 1990).

Figure 4 provides a visualization of the identification for mean ADHD risk by patient gender. The vertical axis plots patient ADHD match signal, x_i, for the set of patients in which it is observed ($Q_i = 1$), paired with their selection-adjusted behavioral assessment propensity on the horizontal axis.

Figure 4: Behavioral Assessment Rates and Observed ADHD Risk



Note: This figure plots gender-specific observed ADHD match signals on selection-adjusted PCP referral rates obtained from predicted behavioral assessment probabilities from equation (9), with demeaned patient controls set to 0, for the set of patients with $Q_i = 1$. The exponential and linear fits are represented by the solid and dashed line, respectively.

Consistent with the theory, observed ADHD match signals, x_i, are decreasing in exogenous behavioral assessment propensity, $\widehat{P_\theta(Q_i|Z_i)}$. A low value of $\widehat{P_\theta(Q_i|Z_i)}$ implies that child i is a patient of a PCP with generally low referral rates. Thus, these patients are ex-ante unlikely to schedule a behavioral assessment appointment. Despite this, the patient appears in the data as receiving a behavioral assessment anyway, which means that they must have high ADHD risk, v_i, consistent with high observed match 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 with conditionally high referral rates. These patients are more likely to schedule behavioral assessments regardless of true symptom risk, and thus have lower *observed* match signals on average.

The solid lines in Figure 4 represent the gender-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, and inverse weighting by the variance of the gender-specific PCP fixed effect from estimating equation (9). Table 5 presents the estimated model fit coefficients for both males and females. This table also presents the vertical intercept at $\widehat{P_\theta(Q_i|Z_i)} = 1$ of the gender-specific curves, corresponding to the unbiased estimates of male and female population mean ADHD risk, μ_θ. Figure 4 also includes the linear fit (dashed lines), with coefficients and extrapolation in Appendix Table A4.

Table 5: Male/Female Extrapolation

	Male (1)	Female (2)
$\widehat{\alpha_0}$	0.350	0.387
	(0.015)	(0.023)
$\widehat{\alpha_1}$	-0.423	-0.933
	(0.175)	(0.305)
Fitted μ_θ	0.230	0.152

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 mean observed ADHD risk signal for patients who receive behavioral assessment and X is risk-adjusted PCP referral rate. All regressions weighted by the inverse variance of PCP-gender fixed-effects from estimating equation (9). Fitted μ_θ denotes the intercept at X=1. Standard errors in parenthesis.

5.2 Second Stage: Recovering Remaining Parameters

I estimate the remaining model parameters by matching moments defined by behavioral assessment rates and coefficients from a conditional diagnosis probit obtained via maximum likelihood estimation, separately for male and female patient groups. Appendix Table A5 further details these moments with their empirical and theoretical counterparts.

With μ_θ estimated in the first stage, it is clear how remaining parameters are identified up to ADHD risk dispersion, σ_θ. Gender-specific mean utilization cost, c_θ, is identified through variation in behavioral assessment rates *conditional* on mean ADHD risk parameter μ_θ (see equation 2). Both diagnostic uncertainty, ρ_θ, and diagnostic thresholds, τ_θ, are identified in the conditional physician diagnosis probability equation (see equation 6). 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 match 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 estimate σ_θ using the moment defined by equation (10) which follows from the truncated normality of selected ADHD match signals. Thus σ_θ is identified by the difference between observed match signals and population mean risk, adjusting for selection due to different healthcare utilization costs and signal strength by patient gender.

$$\overline{x_{obs}}|\theta = E[x_i|v_i > c_i] = \mu_\theta + \rho_\theta \sigma_\theta \frac{\phi\left(\Phi^{-1}(1-\widehat{Q|\theta})\right)}{\widehat{Q|\theta}} \tag{10}$$

6 Estimates and Simulations

Table 6 presents the full set of results for male and female patients, with standard errors and 95% confidence intervals for male/female parameter differences based on bootstrapped patient samples. The sign of each parameter differences in Table 6 can be informative about which mechanisms contribute to the male/female ADHD diagnostic gap and in what direction. As discussed in Section 3.2, diagnostic differences between male and female patients can be attributed to variation in prevalence, mental healthcare utilization, diagnostic uncertainty, and diagnostic thresholds. The results in Table 6 suggest that both underlying ADHD prevalence and physician-set diagnostic thresholds play an important role explaining diagnosis rate differences.

Table 6: Model Parameter Estimates

	Male	Female	Difference
Pop. Mean Risk μ_θ	0.230	0.152	0.077
	(0.034)	(0.047)	[-0.016, 0.147]
Pop. Risk Dispersion σ_θ	0.303	0.340	-0.037
	(0.142)	(0.143)	[-0.266, 0.177]
Utilization Costs c_θ	0.477	0.499	-0.022
	(0.108)	(0.134)	[-0.205, 0.065]
Signal Quality ρ_θ	0.220	0.308	-0.088
	(0.064)	(0.076)	[-0.228, 0.102]
Diagnostic Threshold τ_θ	0.370	0.500	-0.130
	(0.052)	(0.121)	[-0.266, -0.038]

Note: 1000 bootstrapped patient samples used to obtain gender-specific model parameter standard errors (in parenthesis) and 95% confidence interval for the difference between male and female parameter estimates.

First, the population mean risk for males is higher than that for females, with a difference of 0.077 statistically different from 0 at the 10% level. This higher male ADHD prevalence will increase the ADHD diagnostic difference through both the patient selection channel (behavioral assessment scheduling) and through higher physician posterior beliefs. This result is directionally consistent with the medical literature which notes higher ADHD symptom prevalence in boys than girls (AHRQ, 2011). Second, males and females have similar mental health utilization costs suggesting patient preferences do not drive differences in ADHD diagnosis rates. I find some suggestive evidence

that physicians put more weight on female ADHD match signals ($\rho_f > \rho_m$), which by construction measures the overlap between patient symptoms and DSM-V defined symptoms. This finding, though not statistically significant, 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 more on heuristics for male patients.

Most striking is the large and significant difference in diagnostic thresholds between male and female patients. Physicians use significantly lower diagnostic threshold for male patients ($\tau_m < \tau_f$), meaning that they are more likely to diagnose a male patient than a female patient with identical posterior ADHD risk. Because the DSM-V does *not* specify gender specific diagnostic requirements, this finding suggests that physicians deviate from clinical guidelines when making the diagnosis decision in ways that contribute to an ADHD diagnostic disparity by gender. Paired with the utility model that defines these diagnostic thresholds, this key result suggests that physicians deviate from clinical guidelines because their perceived cost of *missed*-diagnosis relative to *mis*-diagnosis is higher for male patients than female patients. I discuss the interpretation and implications of these diagnostic threshold differences in Section 7.

6.1 Simulated Mechanisms Contribution

How do these gender-differences in ADHD diagnosis parameters contribute to the overall differences in diagnosis rates between male and female patients? In this section, I use the structural model and estimates in Table 6 to run ADHD diagnosis simulations, which allows me to isolate and quantify the role of each mechanism as motivated by the model and discussion in Section 3.2.[25]

[25] Appendix Table A6 shows how well the simulated model matches key moments of the observed data, both overall and for male and female subsets of patients. The simulated model does extremely well at matching average diagnosis rates (D), behavioral assessment rates (Q), and mean ADHD match signals ($x|Q$). It slightly underestimates conditional diagnosis rates ($D|Q$), more-so for female patients than male patients.

To show how the various mechanisms contribute to the ADHD diagnostic difference measure, I analyze simulated diagnosis rates under counterfactual scenarios that place restrictions on the source of gender-specific variation. The results of this analysis are presented numerically in Table 7 and visually in Figure 5.

Table 7: Simulated Mechanism Contribution

	Diagnostic Difference	Added Effect	Relative Contribution
No Difference	**1.00**	-	-
Prevalence Contribution			
ADHD Risk Distribution: μ_θ and σ_θ			
at Male estimates	1.49	+0.49	33%
at Female estimates	1.45	+0.45	30%
Patient Contribution			
Utilization Costs: c_θ			
at Male estimates	1.65	+0.16	11%
at Female estimates	1.59	+0.14	10%
Physician Contribution			
Signal Quality: ρ_θ			
at Male estimates	1.52	-0.13	-9%
at Female estimates	1.46	-0.13	-9%
Diagnostic Thresholds: τ_θ			
at Male estimates	2.49	+0.98	65%
at Female estimates	2.49	+1.03	69%
Overall Difference	**2.49**	+1.49	100%

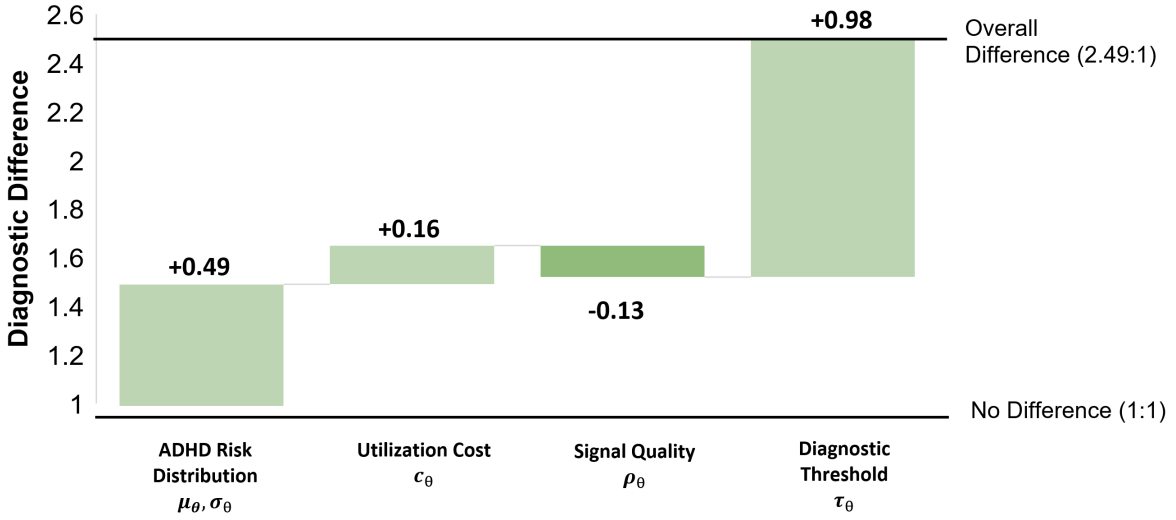
Note: This table presents results from diagnostic simulations with sequential restrictions on the model parameters. Rows show which parameters are varied, starting with no variation, and adding variation until all parameters are at estimated value. Diagnostic Difference is calculated as simulated male diagnosis rate divided by simulated female diagnosis rate. Added Effect calculates the added net diagnostic difference from the previous simulation. Relative Contribution calculated as added effect divided by total effect of 1.49.

The first row of Table 7 corresponds to no diagnostic difference (1.00:1), in which parameters are restricted to be identical for both boys and girls. The second panel shows what happens 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 underlying true ADHD symptom risk varies by patient gender, the simulated diagnostic difference increases from 1.00:1 to 1.49:1 or 1.45:1 depending on at which estimates the remaining parameters are held. This represents 33% or

30% of the observed difference in male and female diagnosis rates, suggesting that at most one-third of the male/female ADHD diagnostic difference can be attributed to differences in the underlying ADHD symptom prevalence.

When patient mental healthcare utilization costs are also allowed to vary by patient gender, diagnostic differences increase only slightly, suggesting that very little of the male/female disparity can be attributed to differences in selection into mental health care (net of true prevalence differences). 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 male/female diagnostic gap, but this is more than made up for by different diagnostic thresholds which explains between 65% to 69% of the observed diagnosis rate difference between male and female patients.

Figure 5: Cumulative Simulated Mechanism Effect



Note: This figure shows the cumulative effect of each mechanism in explaining the ADHD male/female diagnostic difference. Values come from Column 2 of Table 7, where parameter restrictions in simulations are set at male parameter values.

Figure 5 presents the mechanism decomposition visually. The first bar, which corresponds to true underlying ADHD risk, fills about one-third of the overall male/female ADHD diagnostic difference, meaning that at least some of the difference in diagnosis rates between male and female patients can be attributed to differences in true

underlying prevalence rates. However, this suggests that the remaining two-thirds of the diagnostic difference is an unwarranted disparity, at least according to the DSM-V guidelines. The final bar in Figure 5 shows that the ADHD diagnostic disparity primarily stems from physicians using different thresholds based on patient gender, a practice that suggests deviations from clinical guidelines. In the following section, I discuss implications of this deviation and whether or not this disparity is medically and even economically unwarranted.

7 Discussion and Implications

The results presented above show that male children are more likely to match ADHD diagnostic guidelines, both in the selected sample and the population more broadly. However, I also show that conditional on the true prevalence difference between boys and girls, there is still a significant diagnostic disparity, that is mostly explained by differences in physician thresholds for diagnosis. Specifically, I show that physicians use lower thresholds to diagnose their male patients, which is in contrast to the DSM-V guidelines that require the same number and severity of symptoms regardless of patient gender. The implication of these results, however, are nuanced and depend on the goals of health, education, and/or economic policy-makers.

From a purely healthcare perspective, the goal may to eliminate disparities caused by deviations from clinical guidelines. In this case, the results above suggest that policies aimed at physician diagnostic compliance can reduce the ADHD gender disparity to the estimated true prevalence difference of 1.5:1. Alternatively, it may be that physicians know more than what these DSM-V guidelines convey, in which case the goal would be to update clinical guidelines in a way that reflects how ADHD manifests differently in male and female children. In fact, it is a common consensus among psychologists that because the DSM-V definition of ADHD is outdated and/or too terse, physician discretion and variation from clinical guidelines is medically warranted (Cheyette and Cheyette, 2020).

Another argument is that the disparity caused by physician diagnostic thresholds

can be attributed to non-monetary social and/or educational costs of diagnostic errors. Recall that a utility model defines gender-specific diagnostic thresholds, and a lower threshold corresponds to higher costs of under-diagnosis (or conversely, a lower cost of over-diagnosis). Therefore, given a male and female patient with identical ADHD risk, physicians diagnose as if it costlier to under-diagnose the male patient but less costly to under-diagnose the female patient. This diagnosing behavior may in fact be warranted given the way in which ADHD presents differently in male and female children.

For example, female patients with ADHD are more likely to experience internalizing comorbid mental health conditions like anxiety or depression (Hinshaw et al., 2022; Gershon and Gershon, 2002). It is possible that these symptoms crowd-out ADHD diagnosis, even for patients who meet the DSM-V criteria for ADHD. It could also be the case that treatment for these other co-existing conditions (e.g. behavioral therapy) can help alleviate ADHD specific symptoms, making it less costly to under-diagnose females with ADHD, consistent with the use of higher female thresholds. At the same time, male patients with ADHD are more likely to experience the Hyperactive/Impulsive subtype (Type II) which is associated with externalizing behaviors whereas females are more likely to have the Inattentive subtype (Type I) with internalized behaviors (Hinshaw et al., 2022; Quinn and Madhoo, 2014). While all symptoms can hinder learning and child development, it is clear that Type II symptoms are more likely to cause external disruptions and spillover 'costs' to other students, siblings, or caregivers. Physicians may consider external costs (or be influenced by parent/teacher demands) when evaluating their patients, and in turn use lower male thresholds to avoid these higher costs of under-diagnosis.

Given these differences in ADHD symptom presentation and potential for co-existing conditions between male and female patients, it is reasonable to argue that male and female patients experience different costs of ADHD diagnostic errors. The implication of the above argument is that even if male and female patients have identical ADHD risk according to DSM-V criteria, physicians are economically and potentially medically warranted in their use of different thresholds. Future research is needed to

quantify these external costs/benefits of diagnostic errors on the margin and whether cost differences by gender can justify the estimated diagnostic threshold differences presented in this paper.

7.1 Mis(sed) Diagnosis?

The results and discussion so far have focused on *differences* in diagnosis rates between male and female children. I show that at least two-thirds of the difference in diagnosis rates cannot be explained by underlying ADHD symptom prevalence. Thus far, I have remained agnostic about whether this is due to over-diagnosis of male patients, under-diagnosis of female patients, or a combination of both. As a final exercise, I make an additional assumption about what defines true ADHD prevalence, and show how my model can be used to simulate and examine variation in ADHD diagnostic errors.

For purposes of classifying ADHD diagnostic inaccuracies, I refer back to the DSM-V guidelines for ADHD. These clinical guidelines state that regardless of sub-type, a patient must experience 6 (or more) of the 9 specified ADHD symptoms (see Table 1), implying a guideline-defined threshold of $\overline{v} = \frac{6}{9} = 0.66$. 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 gender. Combining this with the full diagnosis model allows me to simulate the extent of over/under diagnosis for both boys and girls.

This simulation exercise results in a male ADHD prevalence rate of 7.8% and a female ADHD prevalence rate of 6.8%. Comparing these to diagnosis rates observed in the data, it would appear that both males and females are under-diagnosed. However, this does not account for the heterogeneity coming from risk dispersion, patient selection into care, or diagnostic uncertainty. With $\overline{v} = 0.66$, the diagnostic simulation finds that 2.3% of males and 0.6% of females are over-diagnosed, and 2.6% of males and 3.1% of females are under-diagnosed.

It is important to note here that the results of this simulation exercise are likely sample-dependent. Compared to the national average, the empirical sample has a

lower ADHD diagnosis rate of 5%. This may be explained by the higher than national-average proportion of patients of Hispanic ethnicity, and pediatric medical research documents a significantly lower ADHD diagnosis rate for this population group coming from cultural biases (Morgan et al., 2013). This might bias the parameter estimate *levels*, leading to higher rates of under-diagnosis than what would be expected from a more nationally represented sample. However, given the similar ethnicity composition across male and female sub-samples, the male/female parameter estimate *differences* and corresponding decomposition analysis in Section 6.1 remain unbiased.

This final simulation exercise is limited in that it is sample-dependent and makes a strong assumption about how the DSM-V definition of ADHD maps to $\overline{v} = 0.66$ in the structural model. However, it does help illustrate two key points. First, comparing observed diagnosis rates to (perhaps arbitrarily known) prevalence rates masks important heterogeneity of both missed and mis-diagnoses. Second, without relying too much of exact quantities, there is still suggestive evidence that male patients are more likely than female patients to be misdiagnosed, and females are more likely to be missed.

8 Conclusion

Reducing mental health disparities is a national priority, yet quantifying these disparities and isolating their contributing mechanisms is difficult in practice given the subjective nature of mental health diagnosis. This paper presents a new theoretical framework and empirical approach which helps contribute to our understanding of the magnitude and sources of mental health diagnostic disparities.

The model and empirical analysis are motivated by the large gender-specific difference in diagnosis rates for childhood Attention Deficit Hyperactivity Disorder. Male children are 2.5 times more likely to be diagnosed with ADHD than female children, a diagnostic disparity that cannot be explained by prevalence rates alone. I develop a model of ADHD diagnosis, composed of three distinct 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 gender-specific risk distribution, which accounts for variation in true ADHD prevalence between male and female children.

I use electronic health record data to estimate the gender-specific model parameters. First, I construct the necessary variables by applying machine learning and a novel natural language processing algorithm to clinical doctor note text. I then estimate male and female population mean ADHD risk using selection-adjusted primary care physician referral rates. The remaining model parameters are recovered using a method of moments approach leveraging variation in behavioral assessment rates and gender-specific conditional ADHD diagnosis probit. I find that males have higher ADHD prevalence, higher diagnostic uncertainty, and lower diagnostic thresholds than their female counterparts.

The overall ADHD male-to-female diagnostic difference is 2.5:1. A model simulation exercise using parameter estimates show that about one-third of this diagnostic difference can be explained by differences in true underlying ADHD symptom prevalence. The remaining difference is largely driven by variation in physician decision-making based on patient gender, specifically lower diagnostic thresholds used for male patients. Paired with an underlying utility framework, these threshold estimates imply that physicians diagnose as if a missed diagnosis is costlier than a misdiagnosis, especially for their male patients. I discuss the implications of this disparity source in detail, and argue that while these different thresholds suggest non-compliance to medical guidelines, they may be economically and even medically warranted. The clinical support for these heterogeneous costs should be explored further, and perhaps even call for 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.

While this paper addresses the male/female diagnostic difference for ADHD, the general framework can be applied to other mental health conditions and/or population-groups. Mental health conditions are costly to both the individual and society. Thus, identifying disparities and associated mechanisms across socioeconomic status, age, race/ethnicity, residence, etc., is an important goal for 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., and Hull, P. (2022). Measuring racial discrimination in bail decisions. *American Economic Review*, 112(9):2992–3038.

Arnold, D., Dobbie, W., and Yang, C. S. (2018). Racial bias in bail decisions. *The Quarterly Journal of Economics*, 133(4):1885–1932.

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.

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. (2022). Selection with variation in diagnostic skill: Evidence from radiologists. *The Quarterly Journal of Economics*, 137(2):729–783.

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. NBER Working Paper 16382, 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. NBER Working Paper 26679, National Bureau of Economic Research.

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. NBER Working Paper 27711, National Bureau of Economic Research.

Cuddy, E. and Currie, J. (2020). Treatment of mental illness in american adolescents varies widely within and across areas. *Proceedings of the National Academy of Sciences*, 117(39):24039–24046.

Currie, J., Kleven, H., and Zwiers, E. (2020). Technology and big data are changing economics: Mining text to track methods. In *AEA Papers and Proceedings*, volume 110, pages 42–48. American Economic Association.

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.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–74.

Gershon, J. and Gershon, J. (2002). A meta-analytic review of gender differences in adhd. *Journal of attention disorders*, 5(3):143–154.

Gowrisankaran, G., Joiner, K., and Léger, P. T. (2022). Physician practice style and healthcare costs: Evidence from emergency departments. *Management Science*.

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.

Hinshaw, S. P., Nguyen, P. T., O'Grady, S. M., and Rosenthal, E. A. (2022). Annual research review: Attention-deficit/hyperactivity disorder in girls and women: underrepresentation, longitudinal processes, and key directions. *Journal of Child Psychology and Psychiatry*, 63(4):484–496.

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. (2022). Physician practice style for mental health conditions: The case of adhd. *FRB of Chicago Working Paper*.

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.

Persson, P., Rossin-Slater, M., and Qiu, X. (2021). Family spillover effects of misdiagnosis: The case of adhd. NBER Working Paper 28334, National Bureau of Economic Research.

Quinn, P. O. and Madhoo, M. (2014). A review of attention-deficit/hyperactivity disorder in women and girls: uncovering this hidden diagnosis. *The primary care companion for CNS disorders*, 16(3):27250.

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.

Data: The data were purchased using funds awarded via the University of Arizona Graduate and Professional Student Council Research and Project Grant 2019. Data provided by The University of Arizona Center for Biomedical Informatics & Biostatistics- Department of Biomedical Informatics Services.

Online Appendix

Mis(sed) Diagnosis: Physician Decision Making and ADHD

Marquardt (2023)

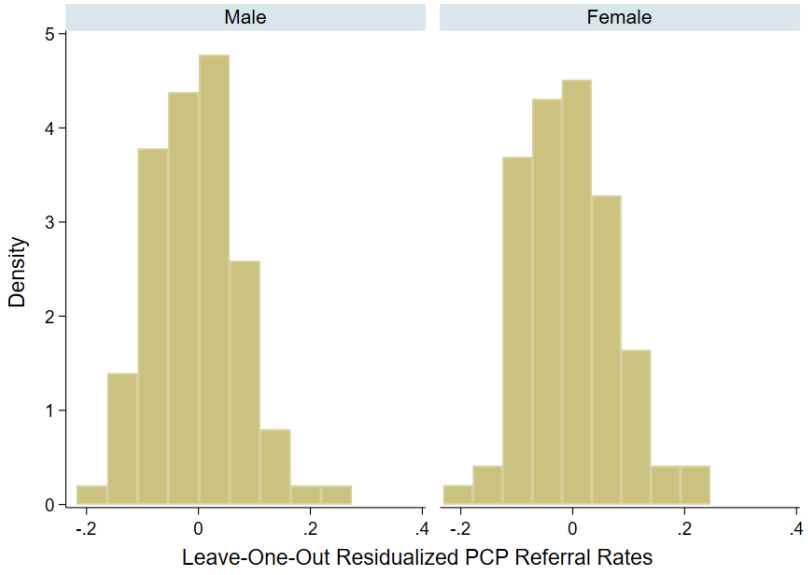
A Additional Tables and Figures

Table A1: Male/Female Difference in Observables

	Male	Female	Difference
Full Sample			
Age	10.57	10.92	-0.343***
Medicaid	0.530	0.546	-0.015
Private Ins.	0.425	0.415	0.011
White-Non Hispanic	0.347	0.347	0.001
Non-White Hispanic	0.282	0.283	-0.001
N	5,554	5,396	
Behavioral Assessment Sample			
Age	10.83	12.53	-1.701***
Medicaid	0.514	0.499	0.016
Private Ins.	0.444	0.472	-0.028
White-Non Hispanic	0.420	0.444	-0.024
Non-White Hispanic	0.237	0.233	0.004
N	1,155	832	

Note: This table presents gender-specific means and difference in means for full sample and Behavioral Assessment subsample ($Q_i = 1$). Significance based on two-sample T-test with * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure A1: *Original PCP* Referral Rate Distribution



Note: This figure plots the gender-specific histograms of *original primary care physicians* leave-one-out referral rates. Referral rate residuals are determined by first regressing behavioral assessment indicator on patient demographic controls, prior utilization controls, and year fixed effects to ensure measure captures referral rate relative to PCPs with the same patient mix. Leave-one-out PCP referral rates calculated using the average residual of all other same-gender patients of the individual's original PCP.

Table A2: Test of First Stage PCP Relevance

Wald Test for PCP Fixed-Effect Significance			
	Total	Male	Female
	(1)	(2)	(3)
Wald Chi-Squared Test Statistic	2484***	1118***	1227.5***
Degrees of Freedom	184	92	92
Patients	5400	2734	2666
Mean Behavioral Assessment Rates	0.197	0.223	0.171

Note: This table shows results from Wald Chi-squared joint test of significance on original PCP fixed effects in a probit regression of patient behavioral assessment indicator on set of patient controls and PCP fixed effects. Patient controls include: Age, Psych Referral, Medicaid, Private Ins., Hispanic, White, Appt. Type, # of Physicians, #of Appts., Year FE. Results shown for three separate regressions based on total sample, male sample, and female sample, respectively. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A3: Test of PCP Selection Balance

	Total (1)	Male (2)	Female (3)
Male	0.001		
	(0.003)		
Medicaid	0.006	0.009	0.002
	(0.004)	(0.005)	(0.007)
Private Ins.	-0.002	0.001	-0.005
	(0.006)	(0.007)	(0.008)
Hispanic	0.001	0.000	0.003
	(0.003)	(0.004)	(0.004)
White	-0.000	-0.001	0.001
	(0.002)	(0.003)	(0.002)
Age	0.000	0.001	0.000
	(0.000)	(0.001)	(0.000)
Year FE	Y	Y	Y
Healthcare Utilization	Y	Y	Y
Observations	5400	2734	2666
Joint Significance? (p-value)	0.541	0.338	0.716

Note: This table presents results from patient level regression of leave-one-out residualized PCP referral rates on observed patient demographics. Outcome is the leave-one-out average referral rate residual among all other patients of patients original PCP. Referral rate residuals determined by first regressing behavioral assessment indicator on patient demographic controls, prior utilization controls, and year fixed effects to ensure measure captures referral rate relative to PCPs with the same patient mix. Robust standard errors in parenthesis, clustered at the PCP level. The table also reports the p-value associated with a F-test of joint significance for patient demographic variables.

Table A4: Linear Extrapolation

	Male (1)	Female (2)
$\widehat{\alpha_0}$	0.348	0.374
	(0.136)	(0.018)
$\widehat{\alpha_1}$	-0.128	-0.254
	(0.053)	(0.086)
Fitted μ_θ	0.220	0.120

Note: This table shows coefficients from weighted OLS regression with linear functional form: $Y = \alpha_0 + \alpha_1 X$ where Y is the mean observed ADHD risk signal for patients who receive behavioral assessment and X is risk-adjusted PCP referral rate. All regressions weighted by the inverse variance of PCP-gender fixed-effects from estimating equation (9). Fitted μ_θ denotes the intercept at X=1. Standard errors in parenthesis.

Table A5: Empirical and Theoretical Moment Descriptions- by Gender

Description	**Empirical Value**	**Theoretical Value**
Behavioral assessment rate: $\widehat{Q_i\|\theta}$	$\frac{1}{N_\theta}\sum_{i\in\theta} Q_i$	$\Phi\left(\frac{\mu_\theta - \boldsymbol{c_\theta}}{\sqrt{1+\sigma_\theta^2}}\right)$
Match coefficient in conditional diagnosis probit: $D_i\|_{Q_i=1,\theta} = \Phi(\alpha + \boldsymbol{\beta} X_i)$	$\hat{\boldsymbol{\beta}} = \frac{\sum_{i\in\theta,Q_i=1}\left((x_i-\overline{x})(\Phi^{-1}(D_i)-\overline{\Phi^{-1}(D)})\right)}{\sum_{i\in\theta,Q_i=1}((x_i-\overline{x})^2)}$	$\frac{\boldsymbol{\rho_\theta}}{\sigma_\theta\sqrt{1-\boldsymbol{\rho_\theta}^2}}$
Constant term in conditional diagnosis probit: $D_i\|_{Q_i=1,\theta} = \Phi(\boldsymbol{\alpha} + \beta X_i)$	$\hat{\boldsymbol{\alpha}} = \frac{\sum_{i\in\theta,Q_i=1}\Phi^{-1}(D_i) - \hat{\beta}\sum_{i\in\theta,Q_i=1} x_i}{N_{Q_i=1,\theta}}$	$\frac{(1-\rho_\theta)\mu_\theta - \boldsymbol{\tau_\theta}}{\sqrt{1-\rho_\theta^2}}$
Observed ADHD signal mean: $\overline{x_{obs}}\|\theta$	$\frac{1}{N_{Q_i=1,\theta}}\sum_{i\in\theta,Q_i=1} x_i$	$\mu_\theta + \rho_\theta\boldsymbol{\sigma_\theta}\frac{\phi\left(\Phi^{-1}(1-\widehat{Q_i\|\theta})\right)}{\widehat{Q_i\|\theta}}$

Note: This table describes the four gender-specific moments (eight in total) used to identify model parameters: $c_\theta, \rho_\theta, \tau_\theta,$ andσ_θ for $\theta = m, f$. Theoretical Values come directly from the structural model described in Section 3.1 and Empirical Values are functions of data only.

Table A6: Observed verses Simulated Rates

	Observed			Simulated		
	Total	Male	Female	Total	Male	Female
ADHD Dx. (D)	0.050	0.071	0.028	0.050	0.071	0.029
Behavioral Appt. (Q)	0.181	0.208	0.154	0.178	0.206	0.153
ADHD match ($x\|Q$)	0.318	0.322	0.314	0.318	0.322	0.314
Cond. Dx. ($D\|Q$)	0.275	0.343	0.180	0.266	0.345	0.187

Note: This table presents average values across patients of ADHD diagnosis, behavioral assessment, ADHD risk signals, and conditional diagnosis. Means are calculated for full set, and subset of male/female patients. 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 6 and model outlined in Section 3.1.

Table A7: Independent Simulated Mechanism Effects

	Diagnosis Rates		Diagnostic
	Male	Female	Difference
Baseline Differences	**0.071**	**0.029**	**2.49**
Panel A: Prevalence			
ADHD Risk Distribution: μ_θ *and* σ_θ			
at Male estimates	0.071	0.041	1.72
at Female estimates	0.48	0.029	1.67
Panel B: Patient Contribution			
Utilization Costs: c_θ			
at Male estimates	0.071	0.031	2.28
at Female estimates	0.065	0.029	2.26
Panel c: Physician Contribution			
Signal Quality: ρ_θ			
at Male estimates	0.071	0.026	2.73
at Female estimates	0.076	0.029	2.62
Diagnostic Thresholds: τ_θ			
at Male estimates	0.071	0.047	1.52
at Female estimates	0.042	0.029	1.46

Note: This table reflects diagnosis rates from a model simulation exercise that restricts variation in only one set of model parameters. The simulated gender-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 6. I include diagnosis rates when equalization is based on male estimate and female estimate. Panel B restricts variation in patient utilization costs, and Panel C restricts variation in physician parameters, signal quality and diagnostic thresholds, respectively.

B Variable Construction using Clinical Texts

B.1 Behavioral Assessment: Q_i

In this appendix, I present the Machine Learning Algorithm used to construct a proxy for the behavioral assessment 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 B1). 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 B1). 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.[26]

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 B1: ICD-9 Labeled Dataset Codes

[26]These sample sizes are larger than the main estimation sample as I do not to make any sample restrictions in building the machine learning algorithm.

I next prepare the doctor notes for feature extraction. This includes traditional text pre-processing procedures: replace contractions, remove special characters and stop words, conversion 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.[27]

- 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=348), number of variables at node split (mtry=2), and maximum number of observations in terminal nodes (nodesize=6).

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 0.02801.

	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 speci-*

[27] 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.

ficity, 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 behavioral assessment 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 correctly 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 behavioral assessment. Approximately 9% of appointments receive a positive predicted label. Results at the patient level are shown in text Table 4.

B.2 ADHD Match Signal: x_i

I next construct the ADHD match signal, x_i, by calculating the "closeness" between the patient's expressed symptoms and the ADHD-specific symptoms defined by the *The Diagnostic and Statistical Manual of Mental Disorders*, (DSM-V). In this appendix, I present the Natural Language Processing Algorithm I use to construct x_i for all patients with $Q_i = 1$. See appendix in Marquardt (2022) for a simplified example.

I first construct vectors to represent each subtype of ADHD by processing the ADHD-type symptom text taken directly from *The Diagnostic and Statistical Manual of Mental Disorders*, (DSM-V). That is, I combine the DSM-V ADHD diagnosis text into three documents corresponding to Inattention (type 1), hyperactive/impulsive (type 2), and all symptoms combined together for the Combined sub-type (type 3).

To ensure that similar words all map to the same meaning, I run each document through a Part-of-Speech tagger and use WordNet to replace each word with it's most common synonym. To further allow for variation in natural language, I aslo obtain each word's "closest" relative word using pre-trained word embeddings from GloVe (Global Vectors for Word Embeddings). I then remove all stop words that are not negation-based, stem all remaining words, and tokenize each document using bi-grams (grouping of two words next to each other in the document).

I then conduct a similar process to create vectors for each patient document, after first combining encounter notes into a single document for each patient. I combine only encounters that were labeled as $Q_{ij} = 1$ by the machine learning prediction described in the previous section. For patients with an eventual ADHD diagnosis code, I include the encounter associated with the first appearance of ADHD diagnosis and behavioral notes from earlier encounters. I also include encounter 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.[28]

With the behavioral assessment notes combined into one document per patient, I then pre-process the text using the standard text cleaning procedures in addition to spell check and abbreviation replacement using a medical dictionary. As with the DSM-V text, I remove stop words (net negation terms), I stem each word, and tokenize documents using bi-grams. To allow for semantic mapping (rather than direct word match), I also replace each stem with its most common synonym and/or word embeddings from the DSM-V processed vectors.

Using these tokenized documents, I build the adjusted Bag-of-Words (BOW) matrix

[28]Of the children that are diagnosed with ADHD in my sample, 33% have a behavioral assessment within 30 days of the initial diagnosis and 42% have a behavioral assessment appointment within 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.

where rows (i) represent documents, columns (k) represent bi-grams of word groups, and matrix elements (i,k) are the "tf-idf" values indicating the relative frequency and importance of bi-gram k in document i.[29] 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.

Finally, 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. Since 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 gender-specific distribution of these constructed values are plotted in text, Figure 3, with mean values in text Table 4.

Table B2: Behavioral Assessment Note Length

	Total	Male	Female
Overall	427	431	415
with ADHD Diagnosis	349	357	340
No ADHD Diagnosis	451	458	443

Note: This table presents the note length for patients with behavioral assessments (i.e., those with $Q_i = 1$). Note length is determined after notes are pre-processed and words replaced with most common synonym. Values correspond to median length across notes, but mean note length produces similar patterns.

Table B2 highlights the similarity in note length for male and female patients who receive a behavioral assessment, both overall and by ADHD diagnosis. Table B3 shows that notes are also similar in content. This table includes lists of "predictive words" to give a sense of how well the NLP algorithm does at identifying ADHD-related words and insights into why boys and girls may be diagnosed differently. For reference, the

[29]The "tf-idf" value is defined as $\frac{f_{ki}}{F_i} \times log(\frac{D}{D_k})$ where f_{ki} is frequency of bi-gram k in document i, F_i is length of document i, D is number of documents, and D_k is number of documents with bi-gram k.

most common words in behavioral assessment notes are: *plan, patient, assess, school, normal*. Those specific to male notes are: *football, airplane, built, monkey, baseball* , and those specific to female notes are: *menstrual, daughter, antidepressant, irregular, purge*.

Table B3: Outcome-Specific Predictive Words

Patients with High ADHD Match ($x_i > 0.4$)	
Overall	*task, impulsive, perception, interview, attitude*
Male	*task, impulsive, irritable, opposition, relax*
Female	*insight, peer, distract, impulsive, task*
Patients with ADHD Diagnosis ($D_i == 1$)	
Overall	*organize, checklist, subtype, careless, stimuli*
Male	*predominant, schoolwork, random, friend, organize*
Female	*predominant, defiance, comorbid, fidget, disorganize*
Patients with ADHD Diagnosis given High ADHD Match ($D_i == 1 \vert x_i > 0.4$)	
Overall	*predominant, checklist, organize, careless, stimuli*
Male	*schoolwork, digit, interpersonal, friend, careless*
Female	*defiance, fidget, friend, inhibition, disorganize*
Patients with ADHD Diagnosis given Low ADHD Match ($D_i == 1 \vert x_i < 0.2$)	
Overall	*inattention, questionnaire, forget, scale, redirect*
Male	*inattention, stimuli, gallop, disrupt, mildly*
Female	*inattention, math, suffer, disrupt, fairly*

Note: This table presents the most frequent occurring words that are specific to given outcome but are not in the top 25% of most frequent words in all other notes. For example, *task* is the most common word in notes that belong to patients with high x_i value, but it is not in the top 25% of words that are used in notes of patients that have a low x_i value. Words are extracted after text processing described in text and conditional on identified as behavioral assessment (i.e., those with $Q_i = 1$). High and low ADHD match correspond to values in the top third and bottom third of estimated match, respectively.

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.[30]

Let physician utility be defined by:

$$u_i|\theta = \begin{cases} -1 & \text{if } D_i = 0,\ S_i = 1 \\ -\beta_\theta & \text{if } D_i = 1,\ S_i = 0 \\ 0 & \text{otherwise} \end{cases} \tag{C1}$$

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 gender-specific disutility of misdiagnosis *relative* to missed diagnoses.

The physician chooses $D_i = 0$ or $D_i = 1$ in order to maximize their 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 (C2), and follows from posterior ADHD risk in (4) and the DSM-V defined minimum diagnostic requirement, $\overline{v}$.

$$p(x, \theta) = Pr(v_i|x > \overline{v}) = \Phi\left(\frac{\rho_\theta x + (1 - \rho_\theta)\mu_\theta - \overline{v}}{\sigma_\theta\sqrt{1 - \rho_\theta^2}}\right) \tag{C2}$$

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 (C1), $E[u_i|D_i = 1, \theta] = -\beta_\theta(1 - p(x, \theta)) + 0(p(x, \theta))$ and $E[u_i|D_i = 0, \theta] = -1(p(x, \theta)) + 0(1 - p(x, \theta))$.

[30] This is similar to the utility in Chan et al. (2022), but with variation in cost across patient gender as opposed to variation across physicians.

Assuming misdiagnoses are costly (i.e., $\beta_\theta > 0$), 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 (C2) 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 gender-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 (C1).

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 (C1).

C.2 Modeling Assumptions and Implications

In this appendix, I discuss, the key assumptions made throughout the main text. While I cannot empirically test for the validity of each assumption, I discuss what would happen if the assumption fails, and in most cases determine the direction of the resulting estimation bias.

Full Documentation Assumption

In Section 4.2, I show how ADHD match signal x_i can be constructed using clinical doctor note text. This relies on the assumption that physicians accurately document behavioral symptoms in their notes. There are two situations in which this assumption might fail. First, it may be the case physicians do not conduct a thorough behavioral assessment and thus do not learn about all the symptoms that the patient is experiencing. Alternatively, it may be the case that the physician does learn about the patient symptoms, but does not write these down in the note. In both cases, x_i is a downward biased proxy of individual symptoms such that $x_i^{true} = x_i^{obs} + \zeta_i$ where $\zeta_i > 0$. While ζ_i is unobserved to only the physician in the first case but to the econometrician in both, the implications of the assumption are similar.

Without full documentation, $x_i^{true} > x_i^{obs}$ and therefore $\mu_\theta^{true} > \widehat{\mu}_\theta$. In other words, I underestimate mean ADHD risk in the first stage of estimation. As a result, I also underestimate mental healthcare utilization costs. However, in Appendix Tables B2 and B3, I show that male and female patients have similar doctor notes in terms of both note length and words predictive of high ADHD match. Therefore, it is reasonable to assume that if the full documentation assumption fails, then it fails for both male and female patients. In this case, $\hat{\mu}_\theta < \mu_\theta$ and $\hat{c}_\theta < c_\theta$ for $\theta \in \{m, f\}$.

The other model parameters are unlikely to be impacted by this assumption as they are identified in the second estimation stage using data on physician diagnosis decisions. In the first case, physicians do not know ζ_i and therefore use x_i^{obs} and $\hat{\mu}_\theta$ in the decision-making process, which means $\hat{\rho}_\theta = \rho_\theta$ and $\hat{\tau}_\theta = \tau_\theta$. In the second case, physicians know ζ_i and will use $x_i^{true} = x_i^{obs} + \zeta_i$ in their decision-making process instead of x_i^{obs}. The ADHD diagnosis probit slope, which identifies ρ_θ, remains unchanged with respect to x_i^{obs}, therefore $\hat{\rho}_\theta = \rho_\theta$. The diagnostic threshold estimate becomes, $\hat{\tau}_\theta = (1 - \rho_\theta)\hat{\mu}_\theta + \rho_\theta\overline{\zeta} - k_\theta$ for known gender-specific constant k_θ. Because physicians know ζ_i, it is reasonable to assume that they will replace $\hat{\mu}_\theta$ with $\mu_\theta = \hat{\mu}_\theta + \overline{\zeta}$ as their prior belief, thus cancelling out the unobserved mean $\overline{\zeta}$ and leaving $\hat{\tau}_\theta = \tau_\theta$.

In sum, if the full documentation assumption fails for both boys and girls, then

I underestimate mean ADHD risk and mean utilization costs, with no effect on the other parameter estimates. If the full documentation assumption fails equally for both male and female patients, then the gender parameter *differences* (column 3 in Table 6) are unaffected, and the mechanism decomposition analysis in Section 6.1 remains unbiased.

Physician Prior Assumption

In Section 3, I present a model of ADHD diagnosis that incorporates both patient selection and physician decision-making under uncertainty. In the second stage, physicians learn about patient ADHD risk and update their prior beliefs. The key assumption here is that physicians have unbiased and normally distributed prior beliefs for both males and females: $v_i \sim N(\mu_\theta, \sigma^2_\theta)$.

I make this assumption for two reasons. First, the normality of the prior allows for computational ease and clearer interpretation of the model parameters. One could argue that a more mathematically complete theoretical model would have physicians update their beliefs twice: once after patient selection but before behavioral assessment, and then again after patient assessment. This complicates estimation as it would now require twice-updating where the second prior has a truncated normal distribution, with an unknown truncation point for each patient c_i. It is still possible to recover the model parameters via simulated maximum likelihood estimation, but it would require another assumption that physicians know the distribution of patient mental healthcare utilization costs for males and females, c_θ, which is likely fails in practice. Therefore, I argue that a normally distributed prior belief with single updating is well suited for this application, and the computation and interpretation benefits outweigh the costs of a more complicated physician learning model.

Second, the accuracy of the prior mean is necessary for parameter identification. As is common with these types of decision-making under uncertainty models, it is not possible to separately identify both the agent's prior beliefs *and* the agent's preferences without having additional survey data. Therefore, I assume that physicians know the

gender-specific ADHD risk parameter μ_θ (which is identified and estimated in the selection first stage) in order to separate out the diagnostic threshold parameter, τ_θ, in the conditional diagnosis equation (6).

While the accuracy of the prior distribution is a common assumption, it is likely not satisfied in practice. In what follows, I show that if physicians have inaccurate (albeit normally distributed) prior beliefs, this will only impact the bias of one model parameter, τ_θ, which measures the perceived cost of misdiagnosis relative to missed diagnosis. The estimated diagnostic threshold will now contain both physician perceived cost of diagnostic errors and/or their inaccurate priors. Policy implications will depend on this distinction, but the main results presented in the paper are unaffected.

Suppose physician prior beliefs follow the distributed defined by equation (C3), where γ determines the deviation from accurate prior mean.

$$v_i \sim N(\mu + \gamma, \sigma^2) \tag{C3}$$

If $\gamma > 0$, physicians overestimate population mean ADHD risk, and $\gamma < 0$ implies physicians underestimate population mean ADHD. I drop the θ subscript without loss as parameters are estimated separately for both males and females, so the thought experiment holds for both samples.

Recall that the true ADHD risk distribution parameters, μ and σ, and patient mental health utilization costs, c, are estimated in a first stage patient selection model (see Section 5.1), which does not depend on the physician decision-making process or their prior beliefs. Therefore, these parameters are accurately identified regardless of the physician prior assumption. If physicians have inaccurate priors (i.e., $\gamma \neq 0$), this can only impact parameters that are identified in the conditional ADHD diagnosis, in text equation (6).

After receiving the signal x_i, physicians update beliefs resulting in posterior distribution:

$$v_i \mid x_i \sim N\left(\left(\rho x_i + (1-\rho)\left(\mu + \gamma\right)\right), \sigma^2(1-\rho^2)\right)$$

Using the same utility framework, and letting $k = \frac{1}{\sigma\sqrt{1-\rho^2}}$, the new conditional diagnosis rate is defined by equation (C4), where $\tilde{\tau} = \tau - (1-\rho)\gamma$.

$$\begin{aligned} P(D_i = 1 \mid Q_i = 1, x_i) &= \Phi(k\rho x_i + k(1-\rho)(\mu+\gamma) - k\tau) \\ &= \Phi(k\rho x_i + k(1-\rho)\mu - k\tilde{\tau}) \end{aligned} \tag{C4}$$

The diagnostic uncertainty parameter, ρ, is also unaffected by γ as it is identified by the slope coefficient measuring correlation between diagnosis decision and patient signal, x_i. Therefore, the only parameter that is impacted by inaccurate physician priors is the diagnostic threshold τ, and the bias of the estimate depends on whether physicians over or under-estimate mean ADHD risk in their priors. If physicians over-estimate mean ADHD risk with $\gamma > 0$, then $\tilde{\tau} < \tau$, meaning that my estimates of the perceived costs associated with misdiagnosis are biased downwards. On the other hand, if physicians behave as if ADHD risk is lower than true risk, then $\tilde{\tau} > \tau$, and I over-estimate the perceived cost of a misdiagnosis.

Because the model parameters are identified and estimated separately for boys and girls, it is possible for the direction of the bias on τ to differ by sub-group. However, regardless of the inaccuracy in physician prior beliefs, it is still the case that estimated diagnostic thresholds for male patients are lower than diagnostic thresholds for female patients, i.e., $\tilde{\tau_m} < \tilde{\tau_f}$. The only implication is how to interpret these diagnostic thresholds, as they now contain both physician perceived cost of diagnostic errors and/or their inaccurate priors. Distinguishing between the two is outside the scope of this paper.

PCP Selection Assumption

The mean ADHD risk parameters, μ_θ, are estimated using a selection model approach described in Section 5.1. Identification relies on the independence between patient risk, v_i, and their chosen or assigned primary care physician, *conditional* on observables.

The main text argues for this assumption and provides empirical tests showing that once referral rates are adjusted for selection-on-observables, there is no evidence of male/female differences in PCP referral rate propensity.

There may still be concern that patients choose PCPs based on unobserved factors that are correlated with ADHD risk. This will only impact the parameters estimated in the first selection stage (μ_θ and c_θ) as this assumption does not change the decision-making process of the diagnosing physician, who is usually not the original PCP (as noted in the main text).

The direction of the bias depends on the direction of unobserved correlation, which can theoretically be either positive or negative. If patients with high ADHD risk select into high referring PCPs, then my estimates of mean ADHD risk, μ_θ, are biased upwards. This can be seen visually in Figure 4. Under positive risk-referring selection, the patients who see high referring physicians (high x-axis value) have higher than average ADHD risk (high y-axis value), leading to a biased upwards extrapolation point at $\widehat{P_\theta(Q_i|Z_i)} = 1$. Because utilization costs are identified off of mean risk, then estimates of c_θ are also biased upwards. Alternatively, if patients with high ADHD risk select into low referring PCPs, then my estimates of mean ADHD risk and utilization costs are biased downwards.

Similar to the full documentation assumption, if the PCP selection assumption fails equally for both male patients and female patients, then the gender parameter *differences* (column 3 in Table 6) are unaffected, and the mechanism decomposition analysis in Section 6.1 remains unbiased. However, if there is a gender difference in the correlation between v_i and PCP selection that cannot be controlled for with observables, then both parameter estimate *levels* and *differences* will be impacted, which in turn will bias the mechanism decomposition analysis. The direction of this bias depends on sign and magnitude of this unobserved gender-specific selection, which is theoretically ambiguous and empirically untestable. Primary care physician choice and how it relates to the mental health referral process and child mental healthcare are outside the scope of this paper, but are important topics for future research.