Federal Reserve Bank of Chicago

Mis(sed) Diagnosis: Physician Decision Making and ADHD

Kelli Marquardt

REVISED
May 2025

WP 2022-23

https://doi.org/10.21033/wp-2022-23

Working papers are not edited, and all opinions and errors are the responsibility of the author(s). The views expressed do not necessarily reflect the views of the Federal Reserve Bank of Chicago or the Federal Reserve System.

Mis(sed) Diagnosis:
Physician Decision Making and ADHD

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

May 2025

Abstract

The mechanisms driving disparities in mental healthcare are not well understood. This paper develops a structural model of diagnosis for Attention Deficit Hyperactivity Disorder (ADHD), highlighting how patient and physician factors can result in disparities. Using electronic health record data, I estimate model parameters and decompose the observed male:female ADHD diagnostic difference of 2.3:1. Simulations show that only 46-55% 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 for their male patients, which I argue may have economic justifications.

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

[*] I am grateful for helpful feedback from Janet Currie, Keith Ericson, Bo Honore, Luojia Hu, Christoph Kronenberg, John Romley, Maya Rossin-Slater, Jessamyn Schaller, Molly Schnell, and conference and seminar participants at various universities and institutions. I thank Shria Holla and Aryan Safi for excellent research assistance, and I also thank my dissertation committee: Gautam Gowrisankaran, Keith Joiner, Ashley Langer, Juan Pantano, and Tiemen Woutersen. 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: kelli.r.marquardt@gmail.com

[†] The views here do not represent those of the Federal Reserve Bank of Chicago or the Federal Reserve System. Errors are my own.

1 Introduction

While overall health disparities have broadly declined in the last decade, mental health disparities show the opposite trend (Agency for Healthcare Research and Quality, nd). Mental health conditions- especially when diagnosed in childhood— can significantly shape long-run educational attainment, social program participation, and labor market outcomes (see Currie, 2025, and cites therein). Accordingly, understanding the sources of mental health diagnosis, in particular whether they come from patient-side demand factors or physician-side supply decisions, is essential for designing effective policies to address the potential for mental health disparities and their long-run consequences.

In this paper, I develop a model of mental health diagnosis and use electronic health record data to quantify the mechanisms that contribute to mental health diagnostic disparities. I focus on a common child mental health diagnosis, Attention Deficit Hyperactivity Disorder (ADHD), which has a particularly salient diagnosis rate difference by gender. In the United States, approximately 11.5% of children are diagnosed with ADHD, with males diagnosed and treated more than 2 times the rate of females. ADHD is a particularly useful condition through which to examine mechanisms of diagnostic disparities. Clinical diagnostic guidelines for ADHD are uniform and not gender-specific, raising questions about whether observed differences reflect true differences in symptom prevalence, patient behavior, or physician decision-making under uncertainty. Moreover, ADHD diagnosis is based on subjective assessments of behavior rather than objective medical testing, which may introduce substantial physician discretion.[1] While economists have studied the short and long run impacts of ADHD diagnosis and treatment (e.g., Currie and Stabile, 2006; Currie et al., 2014; Chorniy and Kitashima, 2016), documenting the factors that influence ADHD diagnostic disparities, particularly the role of the physician, remains an understudied area.

[1] A large literature studies factors that influence physician decision-making, discretion, and the potential role of guidelines in shaping their behavior (see overview in Currie et al., 2024). There is also a more recent and growing economic literature on factors that influence the treatment and impact of mental health conditions in general. See, for example, Alexander and Schnell (2019),Currie and MacLeod (2020), Biasi et al. (2021), and Cuddy and Currie (2024).

To address this open question, I begin by introducing a model that illustrates how differences in ADHD diagnosis can be explained by variation in underlying ADHD symptom prevalence, patient preferences, and factors driving physician decision-making under uncertainty. I use electronic health record data to estimate gender-specific model parameters, which allows me to quantify the male/female ADHD diagnostic disparity and isolate how each of these mechanisms contribute to differences in observed diagnosis rates.[2] Understanding the factors that contribute to clinical ADHD diagnosis decisions is important as both *missed* and *mis*-diagnosis are costly. On one hand, under-diagnosis can hinder human capital development through, for example, limited access to benefits of educational accommodations (Ballis and Heath, 2021). On the other hand, over-diagnosis may expose children to unnecessary pharmacological treatment and its potential side effects (Currie et al., 2014).[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 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 their perceived cost associated with over/under diagnosis or

[2] It remains an open question within the medical community 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.

[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).

strict guideline adherence.

Taken as a whole, the model highlights four key mechanisms of mental health diagnosis that can potentially vary by patient gender and therefore contribute differentially 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 36,000 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 2.3:1. I show that this gap persists even after controlling for a variety of patient observables and selection into care, motivating the need for a structural model to help decompose and quantify the underlying mechanisms of this diagnostic difference.

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 who 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 physician. While this physician may be chosen endogenously, I assume that the patients' choice of *initial* primary care provider is orthogonal to behavioral symptom development. I show that these base primary care providers 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 diagnosis 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 rate for each gender is a function of physician prior beliefs and their perceived tradeoff between over/under diagnosis. 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 about one-half 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 relatively 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 a little over one-half of the gap in male/female ADHD diagnosis rates.

I end by exploring mechanisms that may rationalize physicians' use of gender-specific diagnostic thresholds, despite uniform clinical guidelines. First, I note that the estimated diagnostic uncertainty is higher for males, which would justify lower diagnostic thresholds even under symmetric cost/benefit of diagnosis. However, I also argue that in the case of ADHD, there are asymmetric externalities associated with diagnosis for boys and girls, further supporting the use of gender-specific thresholds. With an additional empirical analysis, I show that diagnostic thresholds vary based on the specific sub-type of ADHD symptoms (inattentive vs hyperactive/impulsive), especially for boys. This finding is consistent with the notion that more salient or disruptive symptoms carry higher external costs (as in Aizer, 2008), which physicians may implicitly incorporate into diagnosis decisions. I also document higher rates of internalizing mental health diagnoses among females such as anxiety and depression. If treatment for these co-existing conditions mitigates ADHD-related behaviors (or if treatment for ADHD would exacerbate the other symptoms as in Currie et al., 2014), the marginal benefit of an additional ADHD diagnosis may be lower for females. Taken together, the evidence suggests that males and females face different externalities and benefits from ADHD diagnosis on the margin. While guidelines are uniform, physicians operate under uncertainty, and when facing heterogeneous externalities or relying on heuristics, physicians may deviate from clinical guidelines in ways that are economically rational, even if not guideline-consistent.

This paper contributes to several strands of literature. First, it adds to a large body of work on healthcare disparities and variation in healthcare delivery generally (e.g., Chandra and Skinner, 2003; Finkelstein et al., 2016; Cutler et al., 2019; Cuddy and Currie, 2020). Within this literature, studies document large variation due to physician-specific factors (Badinski et al., 2023), noting differences in treatment decisions across otherwise similar patients based on demographic characteristics such as race, ethnicity, and gender (e.g., Chandra and Staiger, 2010; Alsan et al., 2019; Cabral and Dillender, 2024; Corredor-Waldron et al., 2024). I find that this is also true in the context of ADHD, where physician diagnostic behavior contributes substantially to gender disparities in mental health diagnosis, even after controlling for observable differences in patient presentation.

This leads to the second strand of literature: understanding expert decision-making under uncertainty and guidelines adherence. Recent work documents substantial heterogeneity in how physicians interpret symptoms, follow clinical guidelines, and allocate scarce resources (e.g., Abaluck et al., 2016; Chan and Gruber, 2020; Currie and MacLeod, 2020; Chan et al., 2022; Mullainathan and Obermeyer, 2022; Schnell, 2022). A key question in this work is whether variation in physician decision-making reflects poor non-compliance with guidelines or appropriate expert discretion. For example, Abaluck et al. (2020) show that physicians deviate from guidelines in treating atrial fibrillation patients, despite larger benefits from guideline adherence. On the other hand, Cuddy and Currie (2024) emphasize tensions between clinical and professional guidelines in mental health prescribing, noting that some discretionary deviations may be beneficial. My setting—a subjective diagnostic process governed by uniform clinical guidelines—provides a sharp test of how adherence varies by patient group, particularly under diagnostic uncertainty and heterogeneous externalities. Taking insights from this literature, I develop a model of mental health diagnosis that incorporates both patient demand factors along with physician decision-making. Importantly, while traditional models assume that health states or true diagnoses are observed on some level, this 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.

Within the context of ADHD specifically, my paper adds to the existing economic literature exploring the potential for ADHD diagnostic errors and effects of childhood ADHD more generally.[4] For example, multiple researchers have shown that where a child's birth-

[4] Understanding ADHD diagnosis is also explored in the medical/public health literature, including meta analyses on diagnostic differences (e.g., Sciutto and Eisenberg, 2007; Hinshaw, 2018), physician/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). Broadly, these studies show that ADHD diagnosis practices vary depending on setting and find inconsistencies in the application of diagnostic criteria. However, they are limited in their potential for informing policy as sample sizes are small and

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; Schwandt and Wuppermann, 2016; Layton et al., 2018; Persson et al., 2025). The implications of these potential ADHD misidentifications are not as clear. Chorniy and Kitashima (2016) find that ADHD treatment reduces risky behavior and injuries. Aizer (2008) notes that ADHD diagnosis and treatment may have positive spillover externalities to peers in the classroom, and access to special education programs have long run benefits (Ballis and Heath, 2021). On the other hand, Currie et al. (2014) do not find evidence that ADHD drug treatment improves educational outcomes and shows that it may even have harmful consequences, especially for girls. My paper adds to this existing literature by decomposing 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.

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

mechanisms are not causally identified.

and observational comparisons. I also describe 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, parameter identification, and present the first stage estimation results. Section 6 presents the remaining model estimates and results from various supplementary analyses. This includes a model simulation decomposition exercise to isolate and quantify mechanisms of disparities, in addition to an extension of the model allowing for type-specific diagnostic thresholds. I also interpret these results and discuss both medical and economic rational for the use of physician discretion in diagnosis decisions. Finally, Section 7 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; Chorniy and Kitashima, 2016). Moreover, academic accommodations are also typically made available for children with ADHD, through 504 plans or Individualized Education Program (IEPs), which have been shown to have significant long run benefits (Ballis and Heath, 2021). On the other hand, there are also costs associated with over diagnosis of ADHD. In particular, stimulant medications used to treat ADHD are classified as Schedule II controlled substances, with documented potential for abuse and adverse side effects. These findings highlight the relevant tradeoffs and importance of accurate ADHD diagnosis for human capital development.

While the exact cause of ADHD is unknown, the medical literature agrees there is a strong heritability component. However, genetics alone do not indicate a diagnosis, and there is less consensus regarding other environmental and structural factors (Hinshaw, 2018).[5] There is no biological or medical test to determine the presence of ADHD in a given patient. Instead, an ADHD diagnosis is defined by a list of behavioral symptoms outlined in *The Diagnostic and Statistical Manual of Mental Disorders*, currently in its fifth edition (DSM-V).

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.

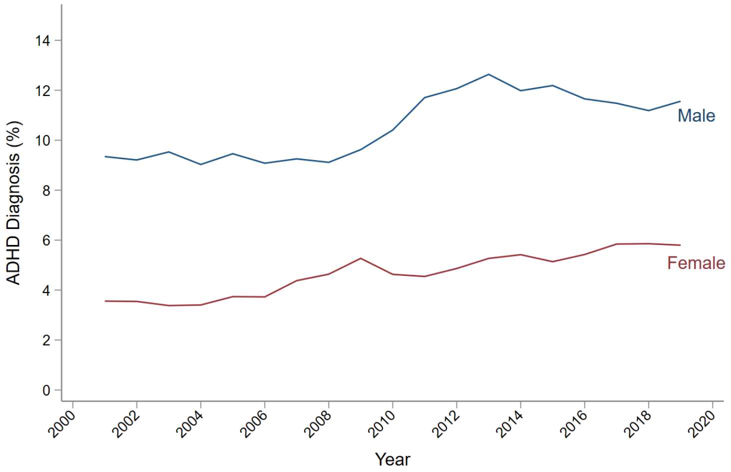
Note: This table reflects an abbreviated list of DSM-V symptoms by ADHD type. The full version is published in American Psychiatric Association (2013).

There are three possible types or presentations of ADHD: inattentive, hyperactive-impulsive, and combined type. While the combined type is most commonly diagnosed (indicating symptoms of both inattention and hyperactivity), research shows that males are relatively more likely to express hyperactive-impulsive symptoms and females are relatively more likely to express those of inattention (Hinshaw et al., 2022). Importantly, 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

[5] See Appendix D.2 for discussion on ADHD Heritability. Other risk factors mentioned in the medical literature include: low birth-weight, prenatal toxins, and exposure to lead (Kim et al., 2020). A list of more debated causes include: food additives/diet, in-utero cellphone radiation, and excess exposure to television/video games.

1.[6] 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, nor does it specify different rules based 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:

[6] For the combined type, patients need 6 or more behavioral symptoms in both the Inattentive and the Hyperactive/Impulsive symptom lists.

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 physician. Scheduling this assessment is not required for all children, but may be encouraged based on feedback from teachers, guidance counselors, or primary care providers during annual wellness checks. A patients' primary care provider (PCP) is either chosen or assigned when the child first enters the health system and is specified as the first point of contact in the health system for general health care needs. While some PCPs may diagnose ADHD themselves, given time constraints or practice authority, it is also common for them to make referrals to other pediatricians or behavioral specialists in the healthcare system. As I emphasize later in the paper, variation in the referrals from primary care providers is an important driver of behavioral assessment take-up.

According to published pediatric best-care practices, a behavioral assessment should include an interview with the patient, the parent, and a teacher or alternative caregiver. 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, 2019) outlines best-practices for ADHD diagnosis, they also admit that these guidelines are often difficult for 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 perceived costs of over/under diagnosis. 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 thresholds, 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.[7] 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.[8] Parents will schedule a behavioral assessment if the child's behavioral symptoms outweigh any mental healthcare utilization costs, 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

[7] 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.

[8] Because I focus on children as patients, I assume the parent and child make joint decisions and thus typically group them together in referring to the "patient" throughout the model.

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 providers 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.

Denoting Q_i as an indicator for behavioral assessment, I define $Q_i = \mathbb{1}(v_i > c_i)$, which reflects patient selection into mental health care either due to high underlying symptoms concerns (v_i), low utilization costs (c_i), or both. 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.[9]

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 unbiased and correlated with the true state through $\rho_\theta \in (0, 1)$. I allow correlation to vary by patient

[9] This assumption allows me to interpret the diagnostic threshold parameter τ_θ as physician preferences over diagnostic errors and/or deviations from guidelines. In Appendix D.2, I discuss the benefits of this assumption and implications if it fails.

gender as a way to capture variation in diagnostic uncertainty coming from signal quality.[10]

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

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, $D_i \in \{0, 1\}$, with the goal of matching the diagnosis to the true health state. This can be modeled as a risk-threshold decision rule where physicians diagnose ADHD to patients whose posterior risk of ADHD is above a diagnostic threshold, τ_θ.[11]

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

In Appendix D.1, I present a physician utility framework and derive this risk-threshold rule, showing how τ_θ can be interpreted as reflecting the physician's perceived tradeoff between the costs of over- and under-diagnosis.[12] Intuitively, if physicians view *misdiagnosis*

[10] 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.

[11] While the baseline model allows diagnostic thresholds to vary by patient gender, it assumes a constant threshold across ADHD sub-types within each gender. In a supplementary analysis (Section 6.2), I relax this assumption and examine its relevance in rationalizing physician diagnostic decision-making.

[12] While rare in practice, its possible that ADHD diagnosis decisions can be reversed. Therefore, τ_θ measures physicians perceived tradeoffs *at the time of behavioral assessment.* In this way, τ_θ also captures the perceived cost of delaying and/or reversing diagnosis down the road, which also could vary based on patient gender.

as costly, they are worried about diagnosing children on the margin of ADHD according to risk and will thus apply a higher diagnostic threshold than imposed by the guideline. On the other hand, if physicians view *missed diagnoses* as costly, they would prefer to diagnose children on the margin of ADHD and will thus apply a lower diagnostic threshold. While uniform clinical guidelines imply a fixed threshold, physicians may differentially deviate by patient gender if they believe the relative costs of diagnostic outcomes differ across these groups. Therefore, I allow these thresholds to vary by patient gender to capture heterogeneity in physicians' perceived diagnostic tradeoffs and in the potential benefits of discretionary deviation from uniform guidelines.[13]

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

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

3.2 Mechanisms of Diagnosis and Diagnostic Disparities

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

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

Diagnosis rates are a function of underlying prevalence, mental healthcare utilization costs, diagnostic uncertainty, and physician preferences/diagnostic thresholds. My model

[13] In related models from the physician bias literature, variation in thresholds is often interpreted as taste-based discrimination as it captures the difference in diagnosis rates for identical patients in terms of risk. However, if the perceived costs of diagnostic outcomes differ by gender, heterogeneous thresholds may be warranted and should not necessarily be viewed as "discrimination." In the model and estimation, I remain agnostic and refer to τ_θ as gender-specific diagnostic thresholds. In Section 6.2, I further consider whether these threshold differences are economically and/or medically warranted.

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.[14] 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 the latter is due to higher physician prior beliefs. On the other hand, high 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 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).[15]

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– 1. higher male prevalence ($\mu_m > \mu_f$), 2. higher signal strength for male patients ($\rho_m > \rho_f$), 3. lower utilization costs for male children ($c_m < c_f$), and/or 4. lower diagnostic thresholds applied to male patients ($\tau_m < \tau_f$).

From a 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.

[14] 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.

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

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: $(\mu_\theta, \sigma_\theta, c_\theta, \rho_\theta, \tau_\theta)$ 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-9 or 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 (EHR) provided by a large healthcare center in Arizona. I first identify the set of physicians (whether general practice or behavioral specialty) who have diagnosed ADHD at least once over the sample period of January 2014 to September 2017. I then obtain encounter-level data for all pediatric patients of this set of physicians over the same sample period.[16]

[16] The details (and implications) of the sample inclusion criteria are discussed in Appendix B.1.

Next, I 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 erroneous encounters, visit cancellations or no-shows, and patients with missing demographic information. The remaining data encompass 36,193 unique patient encounters, for 11,070 unique patients. Patient characteristics include: birth year, gender, race/ethnicity, initial primary care provider, and insurance status. Encounter characteristics include: patient age, visit provider ID, appointment date, broad appointment type descriptors, 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. While the specific symptoms differ by sub-type of ADHD, the clinical requirements for diagnosis are the same (e.g. ≥ 6 symptoms). Further, the ICD coding did not update until mid-2015, making it difficult to distinguish between specific ADHD types using diagnosis codes alone.[17] Patient-level summary statistics are presented in Table 2. 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.

Of the roughly 11,000 patients seen from 2014 to 2017, 5.2% received 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.[18] Males are diagnosed with ADHD significantly more than females. The raw

[17] The diagnosis codes for ADHD include: 314.0 and 314.1 (ICD-9) and F90.0, F90.1, F90.2, F90.9 (ICD-10). Until the updated ICD-10 codes were released in end of 2015, physicians used the ICD-9 codes for ADHD diagnosis, despite there not being a one-to-one mapping with the existing DSM-V criteria.

[18] 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). However, given that males and females in the sample are similar with respect to ethnicity and other observables (see Appendix Table A1), this does not pose a concern for the analysis which focuses on *relative* differences in diagnostic mechanisms by gender.

diagnostic difference is 2.32:1, with 7.2% of males receiving a clinical diagnosis but only 3.1% of females. On average, patients will be seen by two different physicians over an average of 3.3 appointments, a majority of which are with a provider other than their initial primary care provider.[19]

Table 2: Summary Statistics

	Mean	Std. Dev.	Min	Max
ADHD Dx.	0.052	0.221	0	1
Male Dx.	0.072	0.259	0	1
Female Dx.	0.031	0.172	0	1
Male	0.508	0.500	0	1
Age	10.316	3.551	5	18
White	0.561	0.496	0	1
Hispanic	0.494	0.500	0	1
Medicaid	0.535	0.499	0	1
# of Physicians	1.933	1.502	1	15
# of Appt.	3.269	4.111	1	85
# of Appt. (not IPCP)	2.542	3.849	0	85
# Yrs. in Sample	1.690	0.891	1	4
N Patients	11070			

Note: This table presents summary statistics for the full set of patients in the estimation sample. For each patient, age is defined as the average age across all visits in the sample. # of Physicians indicates the number of unique providers seen by the patient during the sample period. # of Appt. indicates the total number of patient appointments during the sample period, and # of Appt. (not IPCP) denotes the total number of patient appointments where the visit provider was *not* the patient's initial primary care provider. See Appendix B for additional details on data and variable construction.

Table 3 presents reduced-form ADHD diagnostic regressions that control for demographics and any other observable differences in healthcare utilization by gender.[20] In all instances, male patients are significantly more likely to be diagnosed with ADHD than female patients. This is true at baseline (column 1), and persists after controlling for differences in other demographics (column 2), general healthcare utilization (column 3), and other mental healthcare utilization more specifically (column 4). This analysis highlights the inability to explain

[19] The initial primary care provider is defined as the specified PCP during the patient's first visit in the sample. See Appendix B.1 for additional details on provider types.

[20] Appendix Table A1 presents differences in male and female patients based on age, race/ethnicity, and health insurance coverage. While male patients are on average 4 months younger than female patients and slightly less likely to be covered by Medicaid, the magnitude of these differences are small.

the male-female diagnostic difference using only directly observable information in electronic health record or claims-based datasets.

Table 3: Reduced Form ADHD Diagnostic Regressions

	(1)	(2)	(3)	(4)
Male	0.041***	0.041***	0.044***	0.036***
	(0.004)	(0.004)	(0.004)	(0.004)
Added Patient Observables:				
Demographics	N	Y	Y	Y
General Healthcare Utilization	N	N	Y	Y
Mental Healthcare Utilization	N	N	N	Y
Adj. R-squared	0.0086	0.0164	0.0502	0.1687
N	11070	11070	11070	11070

Note: This table presents the estimated coefficient on male indicator from a OLS regression of ADHD clinical diagnosis on patient-level observables. Demographic Variables: age and age-squared (averaged across all patient appointments), indicators for insurance coverage, indicators for race/ethnicity, and birth year fixed effects. General Healthcare Utilization Variables: # of doctors seen (across all patient visits), # of appointments, indicator for any wellness-related appointment descriptor, and appointment year fixed effects. Mental Healthcare Utilization Variables: indicators for having any other (non-ADHD) mental health diagnosis, any behavior-related appointment descriptor, and any visit with a psych specialty provider. See Appendix B for additional details on data and variable construction. Robust standard errors in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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 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 text 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. While appointment type descriptors are denoted in the record, the majority are either "office/return" or "urgent/acute" which may include both behavioral and non-behavioral assessment of symptoms. 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 C.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., conduct disorder), or a comorbid condition (e.g., generalized anxiety disorder) as noted by the DSM-V.[21] I then assume appointments are not behavioral assessments if (i) they have a diagnosis code that is not considered a mental health related, and (ii) the patient never receives a mental health diagnosis at any point in the sample period. The most common diagnosis categories in this negatively labeled set include disorders of the eye (ICD10 codes: H00–H60) and diseases of the respiratory system (ICD10 codes: J00–J99).

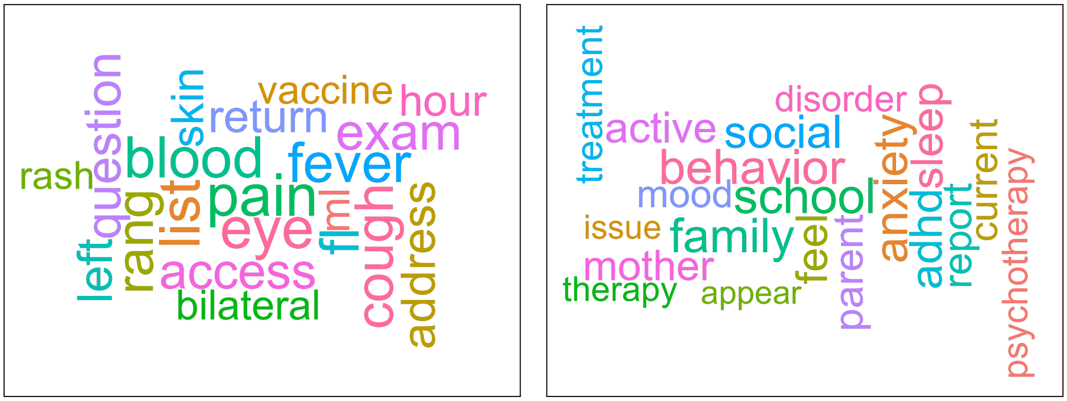
The remaining appointments are considered 'unlabeled'. These are appointments in which there is either no associated diagnosis code in the record, or those for which the diagnosis code could plausibly reflect either a mental or physical health concern, particularly when the patient has a mental health diagnosis elsewhere in their record (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

[21] Mental health conditions are often co-existing (i.e., patients meet the criteria for multiple diagnoses) and thus behavioral evaluations for one condition likely involve considerations of the symptoms associated with others. According to published clinical guidelines (American Academy of Pediatrics, 2011, 2019), " In the evaluation of a child or adolescent for ADHD, the [provider] should include a process to at least screen for comorbid conditions, including emotional or behavioral conditions (eg, anxiety, depression, oppositional defiant disorder, conduct disorders, substance use), developmental conditions (eg, learning and language disorders, autism spectrum disorders), and physical conditions (eg, tics, sleep apnea)."

representation for both negative and positive behavioral assessment labels. As expected, the positive behavioral assessment label includes words related to behavioral symptoms such as: *behavior*, *school*, *social*, 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



(a) Non-Behavioral Words

(b) Behavioral Words

Note: This figure shows "Word Clouds" of most predictive 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 C.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 20.8% of children receive a behavioral assessment. Table 4, presented 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 23.2% and 18.3% respectively. In Appendix C.1, I present results from various validation exercises, and I discuss external validity in Appendix B.2.

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).[22]

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 a natural language processing algorithm 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 C.2.

[22] In Appendix Figure C2 and Table C1, 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 D.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 mechanism decomposition, and subsequent results from supplementary analysis, remain unaffected.

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 patient 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.[23]

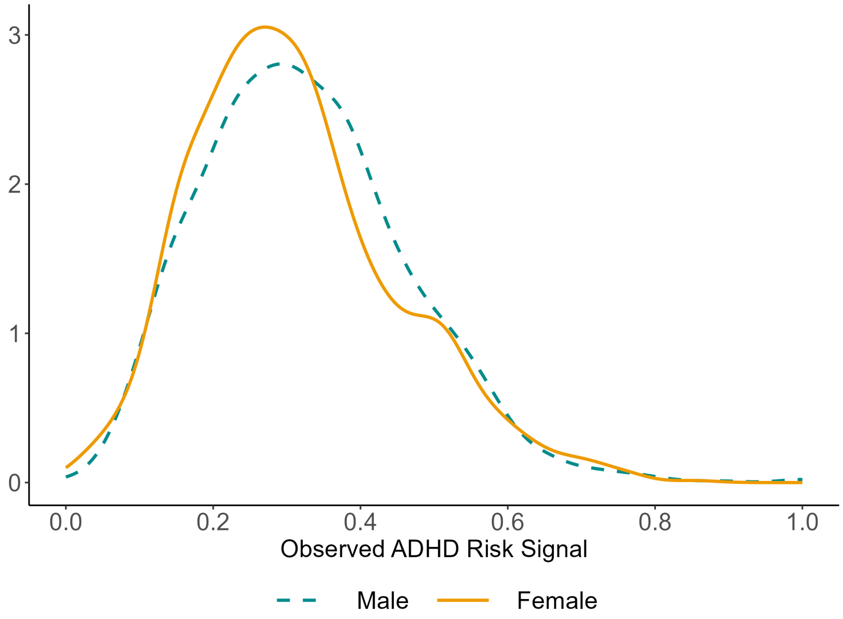
I then calculate the patient overall ADHD match signal by taking the cosine similarity measure between the patient document vector and the official ADHD symptoms listed within *The Diagnostic and Statistical Manual of Mental Disorders*, (DSM-V), rescaling to be within 0 and 1 for better interpretability. I do this for each sub-type of ADHD, with x_{i1} and x_{i2} denoting relative match between patient i note and ADHD symptoms of Inattention (Type I in Table 1) and Hyperactive/Impulsivity (Type II in Table 1), respectively. I then define $x_i = max\{x_{i1}, x_{i2}\}$. While I take the maximum match across symptom types for the main analysis, I relax this definition in Section 6.2 and discuss implications of varying symptom types and their salience/externalities.

In total, the average signal match is 0.319 with a standard deviation of 0.140. For reference, a value of $x_i = 1$ indicates that the patient had the highest overlap between their symptoms and those defined by the DSM-V, relative to other patients. Conversely, a value of $x_i = 0$ indicates the lowest match between the patient note and DSM-V defined symptoms of ADHD. The average signal for males is 0.326, which is only slightly larger than the average female signal match of 0.311 (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

[23] The "tf-idf" value is defined as $\frac{f_{ki}}{F_i} \times [1 + log(\frac{D}{D_k})]$ where f_{ki} is frequency of uni/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 uni/bi-gram k.

patients that receive a behavioral assessment.

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 presents only a truncated distribution of the true population ADHD risk.

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 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, 31.0% of males will be diagnosed with ADHD and only 16.8% 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 relatively small in magnitude. This table provides suggestive evidence that the ADHD diagnostic difference is a function of selection, prevalence, *and* physician decision-making factors. I next outline the model estimation approach which allows me to separate out the magnitude and direction of these underlying mechanisms.

Table 4: Mental Health Observational Comparisons

	Total	Male	Female	Difference
Full Sample				
ADHD Dx.	0.052	0.072	0.031	0.0413***
	(0.221)	(0.259)	(0.172)	
Behav. Appt. (Q_i)	0.208	0.232	0.183	0.049***
	(0.406)	(0.422)	(0.387)	
N	11070	5624	5446	
Behavioral Assessment Subsample ($Q_i = 1$)				
ADHD Dx.	0.248	0.310	0.168	0.1428***
	(0.432)	(0.463)	(0.374)	
ADHD Match Signal (x_i)	0.319	0.326	0.311	0.0146**
	(0.140)	(0.139)	(0.139)	
N	2302	1305	997	

Note: This table shows differences in mental health variables by patient gender. ADHD Dx. (D_i) based on ICD codes in the 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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 the structural parameters of the 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". A valid cost-shifter must be (a) correlated with behavioral assessment propensity through the unobserved patient utilization costs, c_i, and (b) independent of patient ADHD risk, v_i.

I use regression-adjusted referral rates of the patient's *initial* primary care provider (IPCP)—defined as the PCP listed at the patient's first visit in the sample—as the source of quasi-exogenous variation in behavioral assessment propensity. To see how the IPCP 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, they may also bring up these concerns with their child's primary care provider who is trained to ask about patient school performance and behavioral concerns during annual wellness visits (American Academy of Pediatrics, 2011, 2019). Therefore, primary care providers play a role in lowering patient scheduling costs, c_i, through subsequent referral. It is important to note that I cannot observe actual referral events in the data. Instead, I define the "referral rate" of an IPCP as the share of their patients that ultimately receive a behavioral assessment at some point in the sample. This rate may reflect formal referrals but may also capture indirect effects that lower patient scheduling costs such as: providing mental health education, comfortability with health system personnel, help with internal scheduling software, etc., thus satisfying the relevance condition (a).

Importantly, IPCPs 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).[24] To empirically verify that the IPCP identifier meaningfully influences the patient probability of scheduling a behavioral assessment (relevance), I regress patient behavioral assessment indicator, Q_i, on a set of patient controls and the leave-one-out IPCP residualized referral rate. These IPCP referral rates are estimated using data from all other patients of a given IPCP, and residualized measures are used to ensure the variation is coming from IPCP referral rates *relative* to IPCP who see observably similar patient mix. Panel A of Appendix Table A2 presents the result of this exercise, highlighting that these PCP referral rates are strongly predictive of actual behavioral assessments.

Condition (b) is satisfied if IPCPs 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

[24] Appendix Figure A1 shows the variation across IPCPs, with a histogram of the residualized leave-one-out IPCP referral rates for both male and female patients.

institutional details provide support for its validity. First, while a child may switch primary care providers during the sample in response to unobserved (to the econometrician) shocks, I use the identity of the *initial* primary care provider denoted in the patients health record. IPCPs 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 their IPCP after observing v_i. Second, I note that only 28% of the IPCPs covering patients in my sample ever diagnose ADHD (to any patient).[25] So, while IPCPs 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 between the IPCP and patient v_i.

Finally, while patients may not select their IPCP based on v_i directly, condition (b) would still be violated if IPCP selection is based on other factors, W_i, that are correlated with ADHD risk. To test for this, I conduct a balancing-type exercise in which I first regress patient behavioral assessment indicator, Q_i, on a set of observable patient controls. The predicted value of this regression adjusts for the variation in behavioral assessment probabilities coming from *observed* patient demographics and prior utilization controls. I then test whether the residualized IPCP referral rate is associated with the *predicted* rather than *actual* behavioral assessment probability for each patient. The results of this exercise are presented in Panel B of Appendix Table A2. I find that while IPCP referral rates are predictive of actual behavioral assessment (panel A), they are not significantly associated with predicted behavioral assessment probabilities (panel B), providing support for the independence assumption.[26]

[25] In many cases, the physician who ultimately conducts the full behavioral assessment and diagnosis is not the patient's IPCP. These "referrals" may reflect actual formal provider hand-offs, but could also be a result of PCP switches, provider capacity/availability, patient scheduling conflicts, etc.

[26] There may still be concern that patients choose their IPCP 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 D.2.

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 stemming from IPCP referral rates. Empirically, I first estimate a probit regression of behavioral assessment, Q_i, according to equation (9) where W_i includes a set of demeaned patient controls (to net out any IPCP selection based on observables) and γ_j^θ denotes IPCP-by-gender fixed effects.[27]

$$P(Q_i = 1) = \Phi\left(W_i'\beta + \gamma_j^m \mathbb{1}_{(\theta_i=m)} + \gamma_j^f \mathbb{1}_{(\theta_i=f)}\right) \tag{9}$$

With W_i demeaned, $\widehat{\gamma}_j^\theta = (\widehat{\gamma}_j^m, \widehat{\gamma}_j^f)$ is the regression-adjusted IPCP referral rate for male and female patients, respectively. In other words, $\widehat{\gamma}_j^\theta$ measures IPCP j's propensity to refer the *average* patient. While there is significant variation in these adjusted referral rates, the maximum value is only about 0.75. In the absence of an IPCP with regression-adjusted referral rates near 1, values of μ_θ can be estimated via extrapolations of observed ADHD match signals on exogenous behavioral assessment propensity. 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).

Specifically, after obtaining IPCP-by-gender referral rates $(\widehat{\gamma}_j^\theta)$ from equation (9), I then obtain the gender-specific average observed ADHD risk for that IPCP via the analogous OLS regression in equation (10) where x_i is the observed ADHD risk signal for those with $Q_i = 1$, and δ_j^θ are again IPCP-by-gender fixed effects.

$$x_i = \delta_j^m \mathbb{1}_{(\theta_i=m)} + \delta_j^f \mathbb{1}_{(\theta_i=f)} + u_i \tag{10}$$

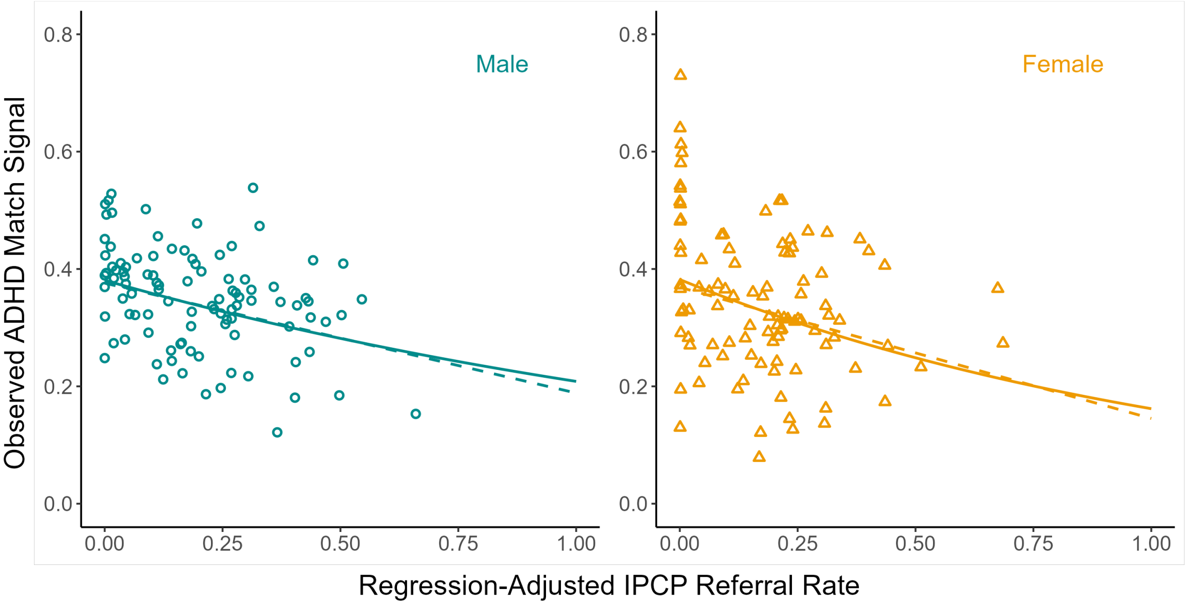
With this, I fit a model of average observed ADHD match signals among patients referred

[27] W_i' includes indicators for patient race, ethnicity, insurance type, and year of birth fixed effects. It also includes the following based on the set of patient visits prior to the first visit t with $Q_{it} = 1$: average age and age squared, total number of appointments by year, number of unique visit providers, indicator for general wellness appointment descriptor, indicator behavior-related appointment descriptor, and indicator for any visit with a psych specialty provider.

by a given IPCP, $\widehat{\delta}_j^\theta$, on the IPCP's referral rate propensity, $\widehat{\gamma}_j^\theta$, separately for $\theta \in \{m, f\}$, inverse weighting by the variance of the observed ADHD risk signal for patients of that IPCP. The resulting selection-adjusted values of μ_m and μ_f are determined by evaluating the fitted model at $\widehat{\gamma}_j^\theta = 1$ for $\theta \in \{m, f\}$, respectively.

Figure 4 provides a visualization of the identification for mean ADHD risk by patient gender. Each dot represents a different IPCP, where the horizontal axis denotes their regression-adjusted referral rate propensity ($\widehat{\gamma}_j^\theta$) and the vertical axis denotes the average observed ADHD match signal for their referred patients ($\widehat{\delta}_j^\theta$).

Figure 4: Behavioral Assessment Rates and Observed ADHD Risk



Note: This figure plots IPCP-by-gender average observed ADHD match signals ($\widehat{\delta}_j^\theta$ from equation 10) on regression-adjusted IPCP referral rates ($\widehat{\gamma}_j^\theta$ from equation 9). This figure also shows the gender-specific model fit, obtained from IPCP-by-gender regressions that inverse weight by the variance of observed ADHD match signals among the IPCPs referred patient set. The exponential and linear model fits are represented by the solid and dashed line, respectively,

Consistent with the theory, observed average ADHD match signals, x_i, are decreasing in regression-adjusted IPCP referral rates, $\widehat{\gamma}_j^\theta$. An IPCP with low $\widehat{\gamma}_j^\theta$ implies that IPCP has a low propensity to refer the average patient. Thus, patients of said IPCP 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 average match signal, x_i. On the other hand, a large value of $\widehat{\gamma}_j^{\theta}$ implies the child is a patient of an IPCP 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.[28]

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 observed ADHD match signals among IPCP's referred patients. Table 5 presents the estimated model fit coefficients for both males and females. This table also presents the vertical intercept at $\widehat{\gamma}_j^{\theta} = 1$ 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)
$\widehat{\alpha_0}$	0.381	0.382
	(0.015)	(0.021)
$\widehat{\alpha_1}$	-0.602	-0.859
	(0.172)	(0.264)
Fitted μ_{θ}	0.208	0.162

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 average observed ADHD risk signal among referred patients of the IPCP ($\widehat{\delta}_j^{\theta}$), and X is regression-adjusted IPCP referral rate ($\widehat{\gamma}_j^{\theta}$). All regressions weighted by the inverse variance of observed ADHD match signals among the IPCPs referred patient set. Fitted μ_{θ} denotes the intercept at X=1. Standard errors in parenthesis.

[28] Appendix Table A3 compares IPCPs by tercile of regression-adjusted referral rates. IPCPs in the top and bottom terciles have fewer patients than those in the middle tercile, resulting in greater variance in their fixed effect estimates from Equation 10 and therefore they are given less weight in the extrapolation procedure. IPCPs do not differ meaningfully by credential, specialty, or training. Those in the top tercile tend to be less experienced (i.e., fewer years since medical school graduation year), but this pattern holds similarly for both male and female patient sets.

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 (11) 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.

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

6 Results and Supplementary Analyses

Table 6 presents the full set of estimated model parameters and the male/female parameter differences. The sign of each parameter difference 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 diagnostic uncertainty/thresholds play an important role explaining diagnosis rate differences.

Table 6: Model Parameter Estimates

	Male	Female	Difference
Pop. Mean Risk μ_θ	0.208	0.162	0.047
	(0.032)	(0.043)	
Pop. Risk Dispersion σ_θ	0.328	0.310	0.018
	(0.086)	(0.101)	
Utilization Costs c_θ	0.448	0.442	0.007
	(0.051)	(0.079)	
Signal Quality ρ_θ	0.272	0.332	-0.060
	(0.058)	(0.071)	
Diagnostic Threshold τ_θ	0.397	0.497	-0.099
	(0.033)	(0.084)	

Note: This table presents the full model parameter estimates along with the difference between the male point estimate and female point estimate. Row 1 parameters come from the estimation procedure described in Section 5.1. Remaining rows correspond to the estimation procedure described in Section 5.2 and Appendix Table A5. Standard errors (in parenthesis) based on replicating the procedures for 1000 bootstrapped patient samples.

First, the population mean risk for males is higher than that for females, with a difference of 0.047. 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). The risk dispersion estimates are noisier, but point estimates suggest that symptom prevalence has higher variation among boys. The utilization cost estimate is slightly higher for male children, though the difference is small in comparison to overall magnitudes and male-female difference is not statistically significant. This suggests that, conditional on symptom prevalence, other patient/parent factors are not the main driver of differences in ADHD diagnosis rates.

Next, both signal quality and diagnostic thresholds are higher for females than for males.

This implies 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 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 relative difference in diagnostic thresholds between male and female patients. Physicians use much lower diagnostic thresholds 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 further examine and discuss the interpretation and implications of these diagnostic threshold differences in the following section.

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.[29]

To show how the various mechanisms contribute to the ADHD diagnostic difference measure, I analyze simulated diagnosis rates under counterfactual scenarios that place re-

[29] Appendix Table A6 compares the simulated model moments to those observed in the raw 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 overestimates conditional diagnosis rates ($D|Q$), relatively more-so for female patients than male patients, but differences are not large.

strictions on the source of gender-specific variation. The results of this analysis are presented numerically in Table 7 and visually in Figure 5. 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 path of added variation in this simulation decomposition exercise then follows the natural structure of the model, starting with underlying gender-specific ADHD risk parameters (μ_θ, σ_θ), then patient mental healthcare utilization costs (c_θ), followed by physician learning (ρ_θ) and finally physician diagnosis decisions (τ_θ).[30]

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	1.59	+0.59	46%
at Female estimates	1.70	+0.70	55%
Patient Contribution			
Utilization Costs: c_θ			
at Male estimates	1.55	-0.04	-3%
at Female estimates	1.66	-0.04	-3%
Physician Contribution			
Signal Quality: ρ_θ			
at Male estimates	1.46	-0.09	-7%
at Female estimates	1.57	-0.09	-7%
Diagnostic Thresholds: τ_θ			
at Male estimates	2.28	+0.82	64%
at Female estimates	2.28	+0.71	55%
Overall Difference	**2.28**	+1.28	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 1.28.

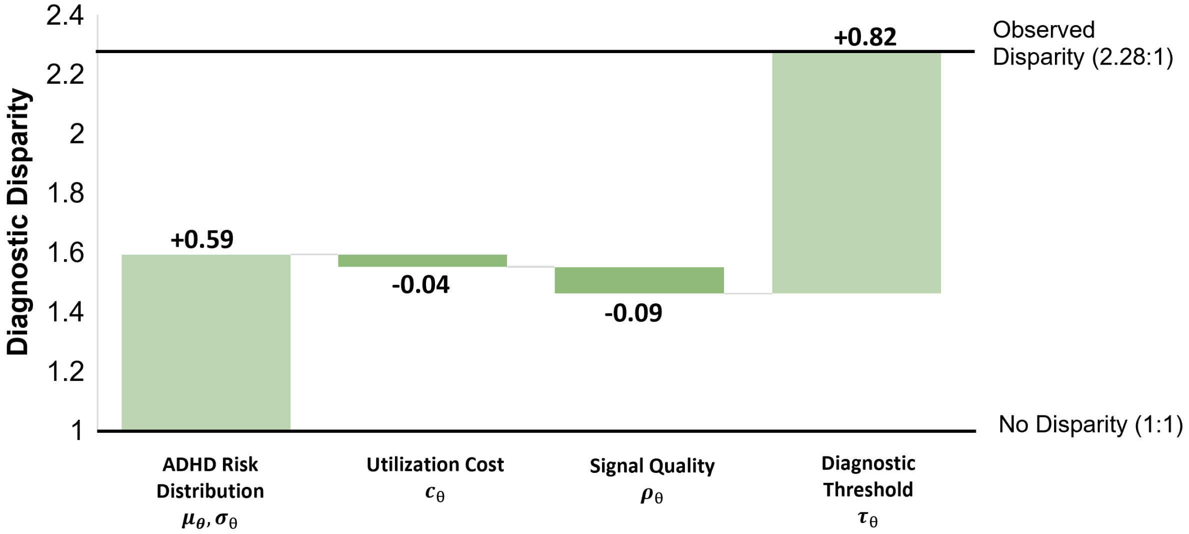
Prevalence Contribution in Table A6 shows what happens when only ADHD risk distribu-

[30] See Appendix Table A7 for a complementary exercise, showing the simulated diagnostic disparity with a 'one-at-a-time' approach, in which all but one of the model parameters are fixed. Consistent with Table 7, this table shows that both population mean risk μ_θ and physician diagnostic threshold τ_θ contribute most to the observed diagnostic disparity.

tion 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 1.59:1 or 1.70:1 depending on at which estimates the remaining parameters are held. This represents 46% or 55% of the observed difference in male and female diagnosis rates, suggesting that about one-half 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 decrease only slightly (3%), 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 reduce the male/female diagnostic gap, but this is more than made up for by different diagnostic thresholds which explain between 55% to 64% 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-half 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 50% 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.

6.2 Physician Diagnostic Thresholds

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. In this subsection, I examine potential mechanisms that may rationalize these heterogeneous thresholds and explore how they relate to clinical presentation, potential for negative externalities, and the prevalence of other mental heatlh comorbidities. I end with a brief discussion of the broader welfare implications of these results.

Table 6 shows that physicians apply lower diagnostic thresholds to male patients than to female patients, despite uniform DSM-V guidelines that do not differentiate by gender. To further quantify the role of these heterogeneous thresholds, I re-simulate the model, restricting diagnostic thresholds to be identical for both boys and girls. Specifically I restrict $\tau_m = \tau_f$, but allow all other model parameters to be at their estimated values in Table 6. Results are presented in Table 8, with the first two rows provided for comparison to the raw data and baseline simulated outcomes.

If physicians applied the lower male diagnostic threshold to female patients as in row 3 of Table 8, the female ADHD diagnosis rate would increase by two percentage points (to 5.1%), and the diagnostic gender gap would decline to 1.46:1. Conversely, applying the higher female threshold to male patients (row 4 of Table 8) would reduce male diagnosis rates and yield a gap of 1.57:1. In both cases, holding thresholds constant significantly narrows the diagnostic disparity, confirming that threshold heterogeneity is a major driver of the observed gender gap in ADHD diagnosis.

Table 8: Diagnostic Gender Gap - Using Same Thresholds

	Male Diagnosis Rate	Female Diagnosis Rate	Diagnostic Gender Gap
Data	7.2%	3.1%	2.32
Baseline	7.4%	3.2%	2.28
$\tau_f = \widehat{\tau_m} = 0.397$	7.4%	5.1%	1.46
$\tau_m = \widehat{\tau_f} = 0.497$	5.1%	3.2%	1.57

Note: This table compares male diagnosis rate, female diagnosis rate, and the ratio (diagnostic gender gap) under various scenarios. Data corresponds to rates observed in the raw data. Baseline corresponds to rates from simulated model at gender-specific parameter estimates from Table 6. The following two rows correspond to simulations holding τ fixed at the male estimate and the female estimate, respectively. All other parameters are held at their gender-specific estimated value.

The results in Table 8 are suggestive of there being more potential for over-diagnosis in males and more potential for under-diagnosis in females, at least according to the DSM-V formal criteria. However, while these clinical guidelines denote that diagnostic thresholds be the same for males and females, the discretion in physician thresholds may be warranted (and even clinically appropriate). For example, one reason why physicians may use lower diagnostic thresholds for male patients is because there is more diagnostic uncertainty ($\rho_m < \rho_f$) in their symptom signals. This could rationalize different diagnostic thresholds even if the perceived cost of misdiagnosis (or benefits of guideline deviation) is the same for boys and girls.

However, in the specific case of ADHD, it is also possible that there are indeed differences in the relative cost of over/under diagnosis or strict guidelines adherence across gender. For example, clinical research shows that male patients are relatively more likely to exhibit the hyperactive/impulsive (Type II) ADHD symptoms whereas females are relatively more likely to exhibit the inattention (Type I) ADHD symptoms (Hinshaw et al., 2022). Notably, the hyperactive/impulsive symptoms are typically more salient and behaviorally disruptive, thus potentially associated with negative externalities such as spillovers to classroom peers (Aizer, 2008). Physicians may internalize these external costs (or be influenced by parent/teacher demands) when evaluating their patients, and in turn apply lower thresholds to avoid under-diagnosing more visibly disruptive boys.

To examine whether this mechanism helps explain the diagnostic threshold differences, I re-estimate the model allowing thresholds to vary not only by gender but also by ADHD symptom sub-type. Specifically, I determine whether the patient's ADHD symptom signal, x_i, is larger for the match with inattentive symptoms (Type I in Table 1) or for hyperactive/impulsive symptoms (Type II in Table 1). I then adjust the diagnosis rule from equation (5) as follows, where x_{i1} and x_{i2} are the extracted signals from Type I ADHD symptom list and Type II ADHD symptom list, respectively:

$$D_i \mid x_i, \theta = \mathbb{1}\big[v_i \mid x_i \geq \tau_{1\theta}\mathbb{1}_{(x_{i1}>x_{i2})} + \tau_{2\theta}\mathbb{1}_{(x_{i2}>x_{i1})}\big] \tag{12}$$

Here, $\tau_{1\theta}$ corresponds to the diagnostic threshold used for patients whose symptoms are predominately inattentive, and $\tau_{2\theta}$ corresponds to the diagnostic threshold used for patients whose symptoms are predominately hyperactive and impulsive. The results from this alternative model are presented in Table 9.

Table 9: Model Parameter Estimates- Type Specific Thresholds

	Male	Female	Difference
Pop. Mean Risk μ_θ	0.208	0.162	0.047
	(0.031)	(0.041)	
Pop. Risk Dispersion σ_θ	0.334	0.309	0.025
	(0.112)	(0.071)	
Utilization Costs c_θ	0.453	0.441	0.012
	(0.073)	(0.049)	
Signal Quality ρ_θ	0.266	0.333	-0.067
	(0.056)	(0.072)	
Diagnostic Threshold, Type 1 $\tau_{1\theta}$	0.449	0.487	-0.038
	(0.092)	(0.069)	
Diagnostic Threshold, Type 2 $\tau_{2\theta}$	0.394	0.497	-0.103
	(0.042)	(0.050)	

Note: This table presents the model parameter estimates along with the difference between the male point estimate and female point estimate. Row 1 parameters come from the estimation procedure described in Section 5.1. Other parameters based on the estimation procedure described in Section 5.2, allowing thresholds to vary according to equation (12). Standard errors (in parenthesis) based on replicating the procedures for 100 bootstrapped patient samples.

Reassuringly, the non-threshold parameter estimates are similar to those from the baseline model in Table 6, indicating robustness. The type-specific diagnostic threshold estimates, however, reveal important new patterns.

First, among male patients, physicians apply substantially lower thresholds for Type II symptoms than for Type I symptoms (0.394 vs 0.449). This suggests they view male hyperactive/impulsive symptoms as costlier than inattentive symptoms, and deviate from clinical guidelines accordingly. This aligns with the notion that disruptive behaviors are more costly to leave untreated. I note, however, that among female patients, thresholds are slightly higher for Type II than Type I (0.497 vs. 0.487), though the difference is small and not statistically significant. This asymmetry suggests that while physicians adjust thresholds downward for boys with more salient/disruptive symptoms, they do not necessarily exhibit the same behavior for girls.[31]

Further, although diagnostic thresholds for males are lower than females for both ADHD symptom sub-types, the gender difference is especially pronounced for Type II symptoms. The female threshold for hyperactive/impulsive symptoms is 26% higher than that for the male threshold, while for inattentive symptoms the diagnostic threshold gap is only 8%. This confirms that the diagnostic threshold disparity observed in the baseline model is largely driven by the use of significantly lower thresholds for hyperactive/impulsive symptoms in boys.

Taken together, these results suggest that physician diagnostic thresholds vary not only by gender, but also by symptom sub-type. The fact that the most pronounced threshold differences are observed among male patients with hyperactive/impulsive symptoms is consistent with the idea that physicians are especially responsive to more salient and potentially disruptive symptoms that may impose negative spillovers to classroom peers or caregivers. This pattern supports the hypothesis that diagnostic threshold variation reflects asymmetric externalities associated with ADHD-related symptoms by gender.

[31] One possible explanation is that even within the same ADHD subtype, symptom expression may differ by gender. Hyperactivity in boys may be perceived as more disruptive than in girls, either due to behavioral norms or contextual expectations, leading physicians to treat similar symptoms differently across gender (Mowlem et al., 2019).

Finally, another potential mechanism behind these threshold differences is the presence of co-existing mental health conditions and the associated spillovers from treatment decisions. For example, female patients are more likely than males to be diagnosed with internalizing conditions such as anxiety and depression.[32] These co-existing conditions may "crowd out" ADHD diagnosis, even for patients who meet the DSM-V criteria. Higher diagnostic thresholds for females could be optimal if treatment for the comorbid condition (e.g., behavioral therapy) also mitigates ADHD-related symptoms, thus reducing the marginal benefit of ADHD diagnosis. In fact, there may even be added marginal costs of diagnosis if stimulant medications used to treat ADHD exacerbate internalizing symptoms, consistent with findings in Currie et al. (2014).

Collectively, the results show that physician diagnostic thresholds vary systematically by patient gender and symptom subtype, and that these differences explain a substantial share of the observed ADHD diagnostic disparity. While these threshold differences represent deviations from clinical guidelines, they may reflect rational responses to the mechanisms discussed above. The welfare implications of this behavior, however, are nuanced and depend on the goals of health, education, and/or economic policy-makers.

For example, I find that the differences in diagnostic thresholds might be driven by negative symptoms externalities, specifically hyperactive boys. While this would rationalize observed physician behavior, it may also reinforce disparities. If boys are over-diagnosed because their symptoms are more socially disruptive, and girls are under-diagnosed because their symptoms are less visible, this could exacerbate inequalities in treatment, access to beneficial educational accommodations, and long-term outcomes.

From a purely healthcare perspective, the goal may be 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 by limiting over-diagnosis of male patients and under-diagnosis of female patients. Alternatively, it may be

[32] See Appendix Table A8 for the in-sample diagnosis rates for other internalizing and other externalizing mental health conditions by gender.

that physicians are responding to information not fully captured by these DSM-V guidelines, in which case the more appropriate response 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).

Ultimately, the findings raise important questions about the adequacy of uniform diagnostic criteria in contexts where the costs (whether clinical or social) of deviating from guidelines differ across patient groups. If male and female patients experience different consequences from *mis*diagnosis or *missed* diagnosis, then heterogeneous thresholds may be economically—and potentially medically—warranted. Future research is needed to quantify these differential costs and benefits to assess the value of clinical guidelines and expert discretion in the case of ADHD, and in and mental health diagnosis more broadly.

7 Conclusion

Mental health disparities are a national concern, yet quantifying these disparities and isolating their contributing mechanisms— while essential for effective policy design— is challenging in practice due to 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 2.3 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 regression-adjusted primary care provider 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.

Model estimates reveal that while approximately half of the observed male/female diagnostic gap can be attributed to underlying prevalence differences, the remainder is driven by differences in physician decision-making, particularly the use of lower diagnostic thresholds for male patients. Paired with an underlying utility framework, these threshold estimates imply that physicians diagnose as if a missed diagnosis is relatively costlier than a misdiagnosis, especially for their male patients.

While the ADHD clinical guidelines are uniform with respect to patient gender, I present supplementary analysis and discussion of mechanisms that explain why physicians deviate from these clinical guidelines. Specifically, when I allow thresholds to vary by symptom subtype, I find that the use of low diagnostic thresholds for male patients is driven by those with hyperactive/impulsive symptoms, which are typically those associated with more salient and disruptive behaviors, and hence carry social externalities, making it marginally costlier to under- diagnose. Further, I note that female patients are more likely to exhibit internalizing mental health conditions (e.g., anxiety or depression) that may already be treated or would otherwise be exacerbated by ADHD pharmacological treatment, hence making them relatively costlier to over-diagnose. Thus, given diagnostic uncertainty, physicians may rationally incorporate these asymmetric tradeoffs of diagnosis, and deviate from uniform clinical guidelines accordingly.

However, even if rational, its possible that these physician deviations from guidelines may in fact reinforce disparities. For example, if physicians are more likely to under-diagnose female patients simply because their symptoms are less salient or not associated with negative social externalities, this could exacerbate the diagnostic disparity and disadvantage female patients in access to beneficial treatment and/or school-based support. Physicians diagnose under uncertainty, and when heuristics are present or externalities are asymmetric, they may deviate from guidelines in ways that are rational but not necessarily clinically (or even socially) optimal. The clinical foundations underlying the mechanisms discussed in this paper 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.

Mental health conditions are costly to both the individual and society. Identifying the mechanisms underlying mental health disparities is an essential first step in designing effective policies to address them. While this paper focuses on the male/female diagnostic difference for ADHD, the general framework can be applied to other mental health conditions and/or population-groups, motivating an importation direction for future research.

References

Abaluck, J., Agha, L., Chan Jr, D. C., Singer, D., and Zhu, D. (2020). Fixing misallocation with guidelines: Awareness vs. adherence. NBER Working Paper 27467, National Bureau of Economic Research.

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.

Agency for Healthcare Research and Quality (n.d.). National healthcare quality and disparities reports. https://www.ahrq.gov/research/findings/nhqrdr/index.html.

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.

Aizer, A. (2008). Peer effects and human capital accumulation: The externalities of add. NBER Working Paper 14354, National Bureau of Economic Research.

Alexander, D. and Schnell, M. (2019). Just what the nurse practitioner ordered: Independent prescriptive authority and population mental health. *Journal of Health Economics*, 66:145–162.

Alsan, M., Garrick, O., and Graziani, G. (2019). Does diversity matter for health? experimental evidence from oakland. *American Economic Review*, 109(12):4071–4111.

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 Academy of Pediatrics (2019). Adhd: clinical practice guideline for the diagnosis, evaluation, and treatment of attention-deficit/hyperactivity disorder in children and adolescents. SUBCOMMITTEE ON CHILDREN AND ADOLESCENTS WITH ATTENTION-DEFICIT/HYPERACTIVE DISORDER.

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.

Badinski, I., Finkelstein, A., Gentzkow, M., and Hull, P. (2023). Geographic variation in healthcare utilization: The role of physicians. NBER Working Paper 31749, National Bureau of Economic Research.

Ballis, B. and Heath, K. (2021). The long-run impacts of special education. *American Economic Journal: Economic Policy*, 13(4):72–111.

Biasi, B., Dahl, M. S., and Moser, P. (2021). Career effects of mental health. NBER Working Paper 29031, National Bureau of Economic Research.

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.

Cabral, M. and Dillender, M. (2024). Gender differences in medical evaluations: Evidence from randomly assigned doctors. *American Economic Review*, 114(2):462–499.

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, D. C. and Gruber, J. (2020). Provider discretion and variation in resource allocation: The case of triage decisions. *AEA papers and proceedings*, 110:279–283.

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 Skinner, J. S. (2003). Geography and racial health disparities. NBER Working Paper 9513, National bureau of economic research.

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*.

Child and Adolescent Health Measurement Initiative (2022). 2022 national survey of children's health (nsch) data query. `https://www.childhealthdata.org`. Data Resource Center for Child and Adolescent Health supported by the U.S. Department of Health and Human Services, Health Resources and Services Administration (HRSA), Maternal and Child Health Bureau (MCHB). Retrieved [05/11/2025].

Chorniy, A. and Kitashima, L. (2016). Sex, drugs, and adhd: The effects of adhd pharmacological treatment on teens' risky behaviors. *Labour Economics*, 43:87–105.

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.

Corredor-Waldron, A., Currie, J., and Schnell, M. (2024). Drivers of racial differences in c-sections. NBER Working Paper 32891, 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.

Cuddy, E. and Currie, J. (2024). Rules vs. discretion: Treatment of mental illness in us adolescents. Accepted at Journal of Political Economy. NBER Working Paper 27890.

Currie, J. (2025). Investing in children to address the child mental health crisis. NBER Working Paper 33632, National Bureau of Economic Research.

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

Currie, J., MacLeod, W. B., and Musen, K. (2024). First do no harm? doctor decision making and patient outcomes. NBER Working Paper 32788, National Bureau of Economic Research.

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.

Currie, J., Stabile, M., and Jones, L. (2014). Do stimulant medications improve educational and behavioral outcomes for children with adhd? *Journal of health economics*, 37:58–69.

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

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.

Demontis, D., Walters, G. B., Athanasiadis, G., Walters, R., Therrien, K., Nielsen, T. T., Farajzadeh, L., Voloudakis, G., Bendl, J., Zeng, B., et al. (2023). Genome-wide analyses of adhd identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains. *Nature genetics*, 55(2):198–208.

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.

Finkelstein, A., Gentzkow, M., and Williams, H. (2016). Sources of geographic variation in health care: Evidence from patient migration. *The Quarterly Journal of Economics*, 131(4):1681–1726.

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.

Grimm, O., Kranz, T. M., and Reif, A. (2020). Genetics of adhd: what should the clinician know? *Current psychiatry reports*, 22:1–8.

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.

Kim, J. H., Kim, J. Y., Lee, J., Jeong, G. H., Lee, E., Lee, S., Lee, K. H., Kronbichler, A., Stubbs, B., Solmi, M., et al. (2020). Environmental risk factors, protective factors, and peripheral biomarkers for adhd: an umbrella review. *The Lancet Psychiatry*, 7(11):955–970.

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.

Levy, F., Hay, D. A., McSTEPHEN, M., Wood, C., and Waldman, I. (1997). Attention-deficit hyperactivity disorder: a category or a continuum? genetic analysis of a large-scale twin study. *Journal of the American Academy of Child & Adolescent Psychiatry*, 36(6):737–744.

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.

Mowlem, F., Agnew-Blais, J., Taylor, E., and Asherson, P. (2019). Do different factors influence whether girls versus boys meet adhd diagnostic criteria? sex differences among children with high adhd symptoms. *Psychiatry research*, 272:765–773.

Mullainathan, S. and Obermeyer, Z. (2022). Diagnosing physician error: A machine learning approach to low-value health care. *The Quarterly Journal of Economics*, 137(2):679–727.

Nikolas, M. A. and Burt, S. A. (2010). Genetic and environmental influences on adhd symptom dimensions of inattention and hyperactivity: a meta-analysis. *Journal of abnormal psychology*, 119(1):1.

Persson, P., Qiu, X., and Rossin-Slater, M. (2025). Family spillover effects of marginal diagnoses: The case of adhd. *American Economic Journal: Applied Economics*, 17(2):225–256.

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.

Schnell, M. (2022). Physician behavior in the presence of a secondary market: The case of prescription opioids. Conditionally accepted, Econometrica.

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

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

Thapar, A., Cooper, M., Eyre, O., and Langley, K. (2013). Practitioner review: what have we learnt about the causes of adhd? *Journal of Child Psychology and Psychiatry*, 54(1):3–16.

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.

Appendix

Mis(sed) Diagnosis: Physician Decision Making and ADHD

Marquardt (2025)

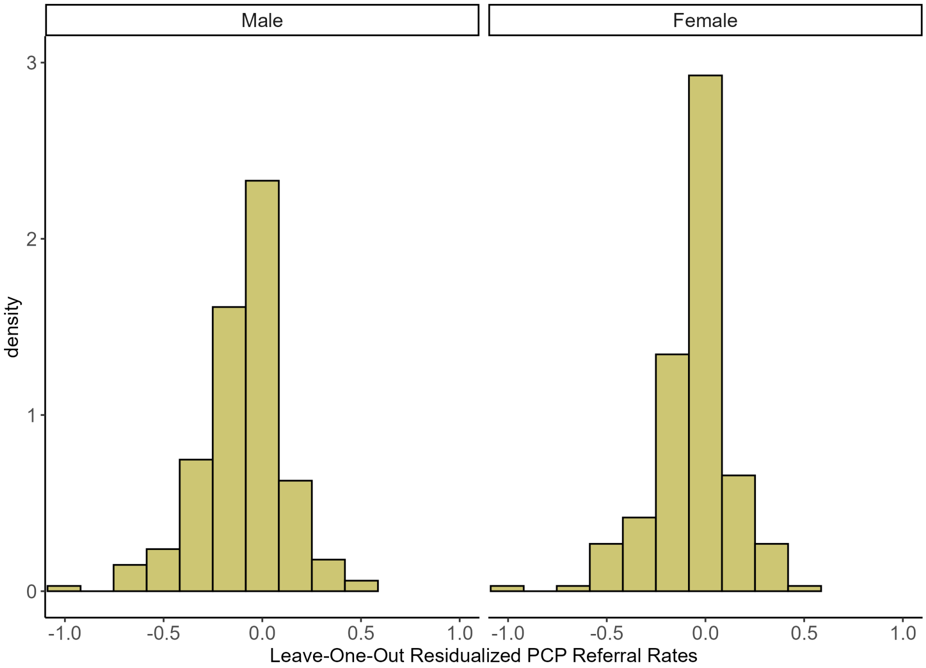
A Additional Tables and Figures

Table A1: Male/Female Difference in Observables

	Male	Female	Difference
Full Sample			
Age	10.167	10.470	-0.304***
Non-Hispanic White	0.348	0.347	0
Non-White Hispanic	0.281	0.281	-0.001
Medicaid	0.527	0.544	-0.017*
Private Ins.	0.426	0.416	0.01
N	5624	5446	
Behavioral Assessment Sample			
Age	10.312	11.690	-1.378***
Non-Hispanic White	0.413	0.429	-0.016
Non-White Hispanic	0.238	0.231	0.007
Medicaid	0.525	0.526	-0.001
Private Ins.	0.431	0.445	-0.014
N	1305	997	

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: Initial Primary Care Provider (IPCP) Referral Rate Distribution



Note: This figure plots the gender-specific histograms of IPCP leave-one-out referral rates. Referral rate residuals are determined by first regressing behavioral assessment indicator on patient controls, W_i, to ensure measure captures referral rate relative to IPCP with the same patient mix. Leave-one-out IPCP referral rates calculated using the average residual of all other same-gender patients of the individual's IPCP.

Table A2: Test of IPCP Relevance and Independence

	Total (1)	Male (2)	Female (3)
Panel A: *Actual* Behavioral Assessment Indicator			
Male IPCP Referral Rate	0.654	0.649	-
	(0.048)	(0.049)	
Female IPCP Referral Rate	0.767	-	0.777
	(0.056)		(0.058)
Panel B: *Predicted* Behavioral Assessment Indicator			
Male IPCP Referral Rate	0.000	-0.005	-
	(0.008)	(0.008)	
Female IPCP Referral Rate	0.009	-	0.012
	(0.009)		(0.009)
Patient Demographics	Y	Y	Y
Healthcare Utilization	Y	Y	Y
Observations	9118	4634	4484

Note: This table presents results from tests of IPCP referral rate relevance (Panel A) and independence (Panel B). The outcome variable in Panel A is the actual behavioral assessment indicator, Q_i. The outcome variable in Panel B is the predicted behavioral assessment probability from an initial regression of Q_i on patient controls, W_i. In both panels, the main regressor is the leave-one-out average regression-adjusted rate residual among all other patients of the given patients' IPCP. Referral rates are regression-adjusted to ensure measure captures referral rate relative to IPCPs with the same patient mix. Patient demographics include: mean age and age-squared, indicators for insurance type, race, and ethnicity, and birth year fixed effects. Healthcare utilization controls include indicators for ever having a wellness-related appointment descriptor, a behavior-related appointment descriptor, and visit with a psychiatrist, number of appointments total and by year, and number of unique physicians seen. Robust standard errors in parenthesis.

Table A3: IPCP Characteristics, by Referral Rate Tercile

	For Male Patients			**For Female Patients**		
	Top Tercile	Middle Tercile	Bottom Tercile	Top Tercile	Middle Tercile	Bottom Tercile
Referral Rate: $\widehat{\gamma}_j^\theta$	0.37	0.18	0.03	0.33	0.17	0.02
ADHD Match Average: $\widehat{\delta}_j^\theta$	0.32	0.33	0.39	0.32	0.31	0.40
ADHD Match Precision: $\text{se}(\widehat{\delta}_j^\theta)$	0.07	0.05	0.08	0.08	0.06	0.10
# of Patients	47.21	80.55	11.24	29.32	97.36	8.30
# of Patients ($Q_i = 1$)	12.18	16.06	2.82	6.82	16.85	1.79
Credentials						
MD/DO	1.00	0.97	0.97	1.00	0.94	1.00
Other	0.00	0.03	0.03	0.00	0.06	0.00
Specialty						
Psych	0.00	0.03	0.00	0.00	0.03	0.00
General Medicine	0.97	0.94	1.00	1.00	0.94	0.97
Experience						
$\leq$ 5 Years	0.29	0.03	0.03	0.29	0.06	0.00
$>$ 15 Years	0.41	0.79	0.73	0.50	0.64	0.79
Education						
US Medical School	0.88	0.85	0.85	0.88	0.94	0.76
N	34	33	33	34	33	33

Note: This table presents the average characteristics of initial primary care providers by tercile of their male and female risk-adjusted referral rate (Figure 4). Referral Rate corresponds to the value on the male or female x-axis of Figure 4 and ADHD Match- Average corresponds to the associated y-axis value. ADHD Match Precision is used to inverse weight the IPCP estimate in the extrapolation described in Section 5.1. Credentials, Specialty, Experience, and Education based on hand-collected information using physician name. See Appendix B.1 for additional details.

Table A4: Linear Extrapolation

	Male (1)	Female (2)
$\widehat{\alpha_0}$	0.376	0.369
	(0.014)	(0.018)
$\widehat{\alpha_1}$	-0.187	-0.224
	(0.056)	(0.080)
Fitted μ_θ	0.189	0.145

Note: This table shows coefficients from weighted OLS regression with linear functional form: $Y = \alpha_0 + \alpha_1 X$ where Y is the average observed ADHD risk signal among referred patients of the IPCP ($\widehat{\delta}_j^\theta$), and X is regression-adjusted IPCP referral rate ($\widehat{\gamma}_j^\theta$). All regressions weighted by the inverse variance of observed ADHD match signals among the IPCPs referred patient set. 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: $c_\theta, \rho_\theta, \tau_\theta, \text{and}\sigma_\theta$ 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)	0.052	0.072	0.031	0.053	0.074	0.032
Behavioral Appt. (Q)	0.208	0.232	0.183	0.206	0.231	0.182
ADHD match ($x\|Q$)	0.319	0.326	0.311	0.319	0.324	0.312
Cond. Dx. ($D\|Q$)	0.248	0.310	0.168	0.258	0.320	0.178

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 Difference	**0.074**	**0.032**	**2.28**
Prevalence Contribution			
ADHD Risk Distribution: μ_θ *and* σ_θ			
at Male estimates	0.074	0.055	1.34
at Female estimates	0.046	0.032	1.43
Patient Contribution			
Utilization Costs: c_θ			
at Male estimates	0.074	0.032	2.34
at Female estimates	0.076	0.032	2.33
Physician Contribution			
Signal Quality: ρ_θ			
at Male estimates	0.074	0.030	2.45
at Female estimates	0.078	0.032	2.40
Diagnostic Thresholds: τ_θ			
at Male estimates	0.074	0.051	1.46
at Female estimates	0.051	0.032	1.57

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, Prevalence Contribution Panel 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 either the male estimate or the female estimate. Patient Contribution Panel restricts variation in patient utilization costs, and Physician Contribution Panel restricts variation in physician parameters, signal quality and diagnostic thresholds, respectively.

Table A8: Other Mental Health Diagnosis- Externalizing and Internalizing

	Total	Male	Female
Full Sample			
ADHD	0.052	0.072	0.031
Other External	0.006	0.008	0.003
Other Internal	0.045	0.035	0.055
Behavioral Assessment Subsample ($Q_i = 1$)			
ADHD	0.248	0.310	0.168
Other External	0.026	0.032	0.018
Other Internal	0.215	0.150	0.301

Note: This table presents diagnosis rates of ADHD, Other External mental health conditions (conduct disorder), and Other Internal mental health conditions (anxiety and/or depression). Rates are presented for the Full Sample and the Subsample with $Q_i = 1$, overall and by patient gender.

B Data Appendix

B.1 Sample Construction & Inclusion Criteria

In this appendix, I describe the sample inclusion criteria and the potential implications of sample selection on model estimates. I also note the various physician types and the definitions based on how they are determined by the data. While I only use the identity of the patient's *initial* primary care provider in the empirical analysis, I provide additional details here on other physician types as they relate to sample construction and assumptions for identification.

The data are derived from de-identified electronic health records provided by a large healthcare system in Arizona. Sample inclusion follows a multi-step procedure described below.

First, I identify the visit provider ID attached to all pediatric patient encounters (those under 18 years of age) that are associated with either a primary or secondary ADHD diagnosis code during the sample period (January 2014 to September 2017). The list of these visit provider IDs determine the set of "sample inclusion physicians" (SIP). By construction, these are all providers who issued at least one primary or secondary ADHD diagnosis to a pediatric patient during the sample period. There are 227 sample inclusion physicians (see Table B1 for summary statistics).

Next, I identify all patient encounters in which the visit provider is one of the sample inclusion physicians, regardless of whether the patient was ever diagnosed with ADHD or not. I extract all encounters that meet this criteria from January 2014 to September 2017. Finally, I restrict the sample further by dropping encounters where the patient is under 5 years of age or over 18. I also drop visits that were canceled/patient no-show, along with visits with missing patient demographic information. The remaining data encompass 36,193 unique patient encounters for 11,070 unique patients.

Encounter characteristics include: appointment date, age of the patient, associated primary and secondary diagnosis codes (if any), name and ID of the visit provider, and name and ID of the patient's primary care provider at the time of the visit. The data also include over 100 unique appointment type descriptors, though these vary in consistency. The most common descriptors are "return/office" or "urgent/acute." Where possible, I hand-classify descriptors into either *wellness-related* (e.g., "well child care," "peds family visit," "physical") or *behavior-related* (e.g., "return child psych," "behavioral visit," "child psych"). However, many descriptors are idiosyncratic and cannot be reliably categorized. Importantly, each encounter also includes a de-identified free-text clinical doctor note summarizing the visit.

Finally, I also observe at the patient level: gender, year of birth, race, ethnicity, and health insurance coverage (Medicaid, Commercial, or other).

Physician Type Definitions

Given the nuances of both the sample inclusion criteria and the source of identifying variation in step 1 of the estimation procedure (see Section 5.1), this section is devoted to defining the various physician types referenced in the analysis.

First, I note that there are two providers that are listed in the raw data for a given patient encounter. One is the *visit provider ID*, corresponding to the provider the patient met with for that specific visit. The other in the *primary care provider (PCP) ID*, which denotes the patient's primary care provider listed at the time of the visit.

Notably, only a subset of visits in the data are with the patient's listed primary care provider (i.e., where visit provider ID is the same as the PCP provider ID). While all PCPs practice general pediatric or family medicine, not all physicians practicing in those areas are recorded as a patient's designated PCP. In many cases, patients are scheduled with pediatricians other than their PCP, even for general wellness, potentially due to availability constraints of the listed PCP or scheduling constraints on the part of the patient or parent.

The list below provides the definition of the physician types that I reference in the analysis. The two most relevant are the selection inclusion physicians (as they are used to build the full sample) and the initial primary care provider (as they are used as identifying variation in the first stage). The diagnosing physicians are a subset of sample inclusion physicians. These are included for comparison but are not used for model estimation or identification purposes.

- Sample Inclusion Physician (SIP): Visit providers who issued either a primary or secondary ADHD diagnosis to any pediatric patient during the sample period.

- Initial Primary Care Provider (IPCP): The primary care provider listed by the patient at their *first* visit in the sample.

- Diagnosing Physician (DP): For patients diagnosed with ADHD, the visit provider most frequently associated with the patient's ADHD-related encounters.

Table B1 shows the counts and cross comparisons of these various physician types. This table also includes physician characteristics. Physician credentials, specialty, experience, and education are based on hand-collected information using physician name (and practice location where necessary) and coded as "Unknown" when not identifiable.

Table B1: Physician Types

	Sample Inclusion Physician (SIP)	Initial Primary Care Provider (IPCP)	Diagnosing Physician (DP)
SIP	**1.00**	0.32	1.00
IPCP	0.14	**1.00**	0.18
DP	0.70	0.28	**1.00**
Any PCP	0.52	1.00	0.55
Credentials			
MD/DO	0.767	0.980	0.811
Other	0.145	0.020	0.151
Unknown	0.088	0.000	0.038
Specialty			
Psych	0.238	0.010	0.270
General Medicine	0.652	0.970	0.717
Other/Unknown	0.110	0.020	0.013
Experience			
$\leq$ 5 Years	0.568	0.120	0.623
$>$ 15 Years	0.145	0.640	0.145
Unknown	0.141	0.010	0.094
Education			
US Medical School	0.736	0.860	0.767
Unknown	0.106	0.000	0.050
N	227	100	159

Note: This table presents summary statistics for the various physician types described in Appendix B.1. The first three rows show the percentage of each physician type that is also included one of the other physician types. Any PCP indicates whether the physician was ever listed as the primary care provider for any patient-visit in the sample (not necessarily the initial primary care provider). Remaining characteristics are based on hand-collected information using physician name. Experience is based on years since medical school graduation when available.

B.2 Generalizability and Robustness

One potential concern is that because all Sample Inclusion Physicians (SIPs) have issued an ADHD diagnosis at some point, the patient sample may be selected toward individuals who had a visit with a physician specializing in ADHD or mental health more broadly. However, as Table B1 shows, a substantial share of SIPs practice general medicine rather than specialize in psychiatric care, and over half were listed as a primary care provider ("Any PCP") at some point in the sample. This alleviates some of the concern that the sample is selected to include only those that see a behavioral specialist.

Nonetheless, approximately 24% of SIPs have a psychiatry specialty, and it is possible that their patients differ systematically in a way that might bias the key results. To assess robustness to this concern, I re-estimate the model using a restricted sample of patients who were included in the data based on having a visit with a SIP who is also a primary care provider (though not necessarily the patient's own PCP), and thus are likely to be more representative of the general pediatric population. This subsample represents 52% of the SIPs identified in Table B1, though notably includes almost 90% of the full sample of patients. This suggests that while a quarter of the SIPs are psychiatric specialty, their patient set is not a large fraction of the overall sample.

I re-estimate both stages of the model using this more general patient subsample. Results are presented in Table B2. The parameter estimates are similar in magnitude and direction to the baseline (in text Table 6). While the male and female difference for risk dispersion flips sign, the difference is not statistically significant in either sample. Thus, the relative differences between male and female patients remain largely robust to this sample inclusion criteria.

Table B2: Model Parameter Estimates - Robust Subsample

	Male	Female	Difference
Pop. Mean Risk μ_θ	0.220	0.127	0.093
	(0.037)	(0.053)	
Pop. Risk Dispersion σ_θ	0.327	0.348	-0.021
	(0.133)	(0.137)	
Utilization Costs c_θ	0.469	0.448	0.021
	(0.081)	(0.112)	
Signal Quality ρ_θ	0.215	0.349	-0.133
	(0.066)	(0.083)	
Diagnostic Threshold τ_θ	0.395	0.506	-0.111
	(0.045)	(0.110)	

Note: This table presents the full model parameter estimates using the subset of patients included in sample based on a visit with one of the Sample Inclusion Physician (SIP) that are also primary care provider (52% of 227 SIP in Table B1). See Table 6 and Appendix B.1 for additional details. Standard errors (in parenthesis) based on replicating the estimation procedure for 100 bootstrapped patient samples.

In addition to the sample inclusion criteria discussed above, there are two other potential sources of sample selection that warrant consideration.

First, a child will be missing from my sample if they do not have *any* visits in the health system during the sample period. The direction of potential selection bias in this case is unclear. On one hand, children who do not seek care may be generally healthy and less likely to need behavioral assessment, suggesting the assessment rate and diagnosis rate observed in my sample are an upper bound relative to the general population. On the other hand, if barriers to seeking health care in general are high (and correlated with ADHD-risk), this would impact the estimates in the other direction. However, the main concern with this sample selection bias is whether it varies by patient gender. Statistics from the National Health Interview Survey (NHIS) suggest that this is unlikely the case. As shown in the left-side panel of Table B3, 95% of children have some sort of general health care visit per year, 99.8% report having healthcare visit at least once in the last 4 years (the length of my sample), and these rates do not statistically differ based on gender.

Second, a child will a child will be missing from my sample if they are in the health system but never had an appointment with one of the 227 visit providers mentioned above. By definition of sample inclusion, these children will not be diagnosed with ADHD. Further, it is also likely that these children do not receive a behavioral assessment, thus suggesting that the behavioral assessment rate that I observe in the sample is an upper bound on the population behavioral assessment rate. While this would impact the *level* of parameter estimates, I note that the primary focus of the analysis is on *relative* differences between male and female patients. Therefore, the concern is whether this sample selection differs by gender. However, as I show in Table A1, there are a similar number of boys and girls in the full sample and they do not meaningfully differ in other observable demographic characteristics.

To further assess generalizability, I compare overall rates and male/female ratios from my empirical sample to analogous measures from a nationally representative dataset. Specifically, I analyze responses from the National Health Interview Survey (NHIS), focusing on children aged 5–17 interviewed between 2014 and 2017. While the NHIS questions do not map directly to the measures I consider for the empirical analysis, the overall rates and male/female ratios provide a useful benchmark for evaluating the representativeness of the sample and text analysis methods discussed in Section 4.1, 4.2 and Appendix C.

The left-side panel of Table B3 presents the NHIS Sample statistics, and the right-side panel represents the analogous statistics constructed from the electronic health record sample used for empirical analysis in this paper.

Table B3: Sample Comparisons- National Health Interview Survey

NHIS Sample	Overall	Male:Female Ratio	**EHR Sample**	Overall	Male:Female Ratio
Visit in 4 Years	0.998	0.999	-	-	-
Annual Visit	0.950	0.998	-	-	-
Behavioral Visit | Annual	0.124	1.29	$Q_i = 1$	0.208	1.27
ADHD (ever)	0.106	2.22	$D_i = 1$	0.052	2.32
ADHD (ever) | Behavioral Visit	0.471	1.68	$D_i = 1 \mid Q_i = 1$	0.248	1.84

Note: This table compares behavioral assessment and diagnosis rates from the empirical sample to similar rates from The National Health Interview Survey. Survey questions used for NHIS variable comparison described in Appendix B.2. NHIS Sample based on children aged 5-17 whereas the EHR Sample includes children 5-18. Both samples cover years 2014-2017.

In the left-panel of Table B3, *Visit in 4 Years* is an indicator based on responses to the question: "Interval since last doctor visit." *Annual Visit* is based on responses to a combination of survey questions, including: "Saw or talked to a general doctor in the past 12 months," "Total office visits in the past 12 months," and "Had a checkup in the past 12 months." *Behavioral Visit* is based on the responses to the following: "Saw/talked to mental health professional, past 12 months" and "Saw/talked to doctor for emotional/behavioral problem" which is only asked of those that responded in the positive to "Saw/talked to general doctor, past 12 months". Finally, *ADHD (ever)* is an indicator for whether the child was "ever told [they] had ADHD or ADD by a medical professional" by the interview date. The NHIS sample includes children ages 5-17 for survey years 2014-2017, and all values are weighted means using the NHIS sample person weights.

There are two important differences between the variables defined in the NHIS and those constructed from the electronic health record sample for my empirical analysis. First, the indicator for a behavioral visit in the NHIS is defined at the yearly level whereas $Q_i = 1$ in the empirical sample reflects a behavioral assessment at any point over the four-year sample period. Moreover, the NHIS behavioral visit measure is limited to respondents who reported either speaking with a mental health professional or discussing emotional or behavioral problems with a general doctor. This likely omits other pathways through which children may receive behavioral assessments in clinical settings. Therefore, the NHIS average behavioral visit rate of 12.4% is likely a lower bound on what the true behavioral assessment rate is nationwide. Indeed, according to another national survey (Child and Adolescent Health Measurement Initiative, 2022), over 25% of children aged 3-17 have "a mental, emotional, developmental or behavioral problem" which presumably follow a behavior-related clinical visit. Further, a (now outdated) pediatric clinical guideline report 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), likely a lower bound on the rate assessed today given the overall increasing trends in mental health diagnoses among children and adolescence since the early 2000s. These differences may explain why the average behavioral assessment rate is higher in my empirical sample (20.8%) than in the NHIS sample (12.4%). Reassuringly, however, the male-to-female ratio of behavioral assessment rates is nearly identical across samples (1.29 in the NHIS and 1.27 in my empirical sample), suggesting that any measurement or sample selection bias is unlikely to differ meaningfully by gender in ways that impact the main implications and interpretation of the results in this paper.

Second, the indicator for ADHD diagnosis in the NHIS is based on ever being diagnosed with ADHD whereas the $D_i = 1$ in my sample is based on being diagnosed during the four-year sample period. As before, this likely explains why the level of diagnosis rates vary across the two samples. Another explanation for the difference in levels stems from the fact that my EHR-based sample has a 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, potentially coming from cultural biases (Morgan et al., 2013). However, the male-to-female diagnosis ratio is again highly similar, both overall (2.22 in the NHIS and 2.32 in my empirical sample) and conditional on those that have a behavioral visit (1.68 in the NHIS and 1.84 in my empirical sample).

C Text Analysis Appendix

C.1 Behavioral Assessment: Q_i

In this appendix, I present the Machine Learning (ML) algorithm used to construct a proxy for the behavioral assessment indicator, Q_i, introduced in text Section 4.1. This closely follows the *Text Analysis Appendix* in Clemens and Rogers (2020).

I first separate the appointment level data into a labeled and unlabeled subsets, where i denotes patient and j denotes appointment. The labeled set is determined by ICD-9 or ICD-10 diagnosis codes where an appointment receives a positive label ($Q_{ij} = 1$) if it is associated with one of the following mental health diagnosis codes: ICD9 codes 209-320, or ICD10 codes F01-F99. An appointment receives a negative label ($Q_{ij} = 0$) if the associated diagnosis code is not among the mental health codes listed above *and* the patient never received a mental health diagnosis at any during the sample period. The most common diagnosis code groupings associated with a negative label include disorder of the eye (ICD10 H00-H60) and diseases of respiratory system (ICD10 J00-J99).

The resulting unlabeled set then contains all remaining appointments. This includes those with no reported diagnosis code, or visits for which a diagnosis code is present but cannot be definitively ruled out as mental health related as the patient receives a mental health diagnosis elsewhere in the record.

Using the labeled dataset, 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.[33]

[33] A document term matrix consists of documents d as rows, words w as columns, and matrix elements t_{dw} representing frequency of word w in document d. The tf-idf value is defined as $\frac{t_{wd}}{T_d} log(\frac{D}{D_w})$ where T_d

- Positive-label word stems: *school, mother, parent, behavior, report, famili, anxieti, current, treatment, disord, social, adhd, psychotherapi, sleep, feel, activ, therapi, issu, appear, mood*
- Negative-label word stems: *pain, eye, blood, list, fever, fl, cough, rang, exam , access, question, left, address, return, skin, hour, ml, bilater, vaccin, rash*

For cross-validation, I split the labeled data into a training and test set using 90-10 split. With 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=314), number of variables at node split (mtry=3), and maximum number of observations in terminal nodes (nodesize=10).

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 0.0345.

	Predicted-0	Predicted-1
True-0	2813	28
True-1	135	783

Before analyzing the final model predictions, I look for issues with *context specificity*, or "limitations on a model's validity outside of its training set" (Clemens and Rogers, 2020). To do so, I take a random sample of 96 notes from the unlabeled dataset, read the unprocessed de-identified notes, and determine the appropriate hand label for behavioral assessment based on my own judgment. I then apply the trained random forest algorithm to this sample, specifying a prediction threshold of 0.5. The resulting confusion matrix is reported in the table below. Of the 96 notes, 88 were correctly classified by the random forest algorithm. Seven notes were incorrectly specified, with only one non-mental health related appointment receiving a positive label.

denotes the number of terms in document d, D denotes the total number of documents, and D_w denotes the number of documents with term w.

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

For a second validation exercise, I classify appointments as behavior-related using two elements of the electronic health record that are not directly used in the machine learning procedure: the reported specialty of the visit provider and the appointment type descriptor. I do not use these elements in the machine learning algorithm as they exhibit greater variation and less consistency than the ICD-based diagnosis codes. Nevertheless, these elements can serve as a useful basis for an external validation exercise.

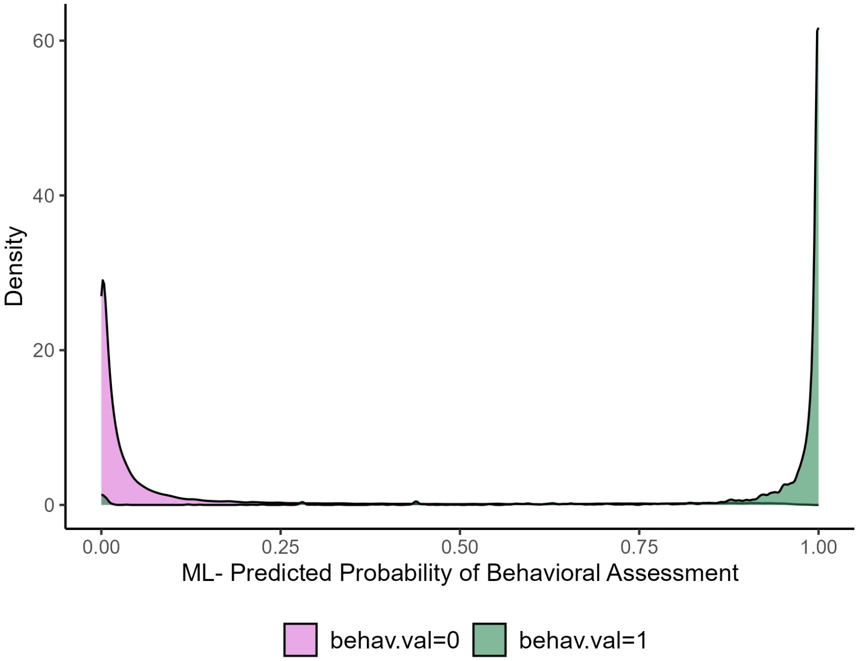
Of the 19 unique provider specialties noted in the health record, I hand-classify three as behavioral in nature: "child and adolescent psychiatry," "developmental and behavioral pediatrics," and "psychiatry." Of the 117 unique appointment type descriptors, I hand-classify 15 as behavior-related based on keywords such as "psy" or "behavioral."[34] For this exercise, I classify a visit j as $behav.val_j = 1$ if either the provider specialty or the appointment descriptor is mapped to a behavioral category, and $behav.val_j = 0$ otherwise. Since the subset $behav.val_j = 1$ reflects appointments that are behavior-related with certainty, I expect the ML algorithm described above to produce higher predicted probability of behavioral assessments for this group than for those with $behav.val_j = 0$ which may or may not include behavioral symptom discussion.

Figure C1 presents the results of this validation exercise. Specifically, for those with $behav.val = 1$ and $behav.val = 0$, I plot the distribution of the predicted behavioral assessment probabilities produced by the trained ML algorithm described above. Reassuringly, the ML-based predictions of behavioral assessment are very high (mean of 0.950) for notes associated with either a psychiatric-specialty visit provider or a behavior-related appointment

[34] There are other appointment type descriptors that may also include behavior related visits, but these tend to be vague and could also include non-behavioral visits as well. For example, the three most common appointment type descriptors are "return/ office", "urgent/acute" and "return peds" which likely include both behavioral and non-behavioral encounters. In fact, among encounters with an associated ADHD diagnosis, only 48% are behavior-related appointment type descriptors with the remainder being more general (e.g. "office visit").

descriptor ($behav.val = 1$). Conversely, the ML-based predictions of behavioral assessment are very low (mean of 0.101) for those that are not ($behav.val = 0$).

Figure C1: $\widehat{Q_{ij}}$- Validation Exercise



Note: This figure presents results from the validation exercise described in Appendix C.1. The distribution of ML algorithm predicted probability of behavioral assessment is shown for both the set of appointments with $beahv.val = 1$ (green) and for the set of appointments with $beahv.val = 0$ (purple).

Given the strong performance of the algorithm in both internal, cross, and external validation exercises, I consider the model sufficiently reliable for application to the full unlabeled set of appointments. Thus, I apply the trained random forest algorithm to generate predicted behavioral assessment indicators for all encounters. Using a prediction threshold of 0.5, I define an indicator for behavioral assessment accordingly. Approximately 9% of appointments in the unlabeled set receive a positive predicted label ($\widehat{Q}_{ij} = 1$). Aggregated results at the patient level are shown in text Table 4.

C.2 ADHD Match Signal: x_i

In this appendix, I present the Natural Language Processing (NLP) algorithm used to construct the ADHD match signal, x_i, introduced in text Section 4.2. This variable quantifies relative similarity between the patient's expressed symptoms and the ADHD-specific symptoms defined by the *The Diagnostic and Statistical Manual of Mental Disorders*, (DSM-V). See appendix in Marquardt (2022) for a simplified example.

I first construct reference vector documents for each subtype of ADHD by processing the ADHD sub-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 two documents corresponding to either Inattention (Type I) or Hyperactive/Impulsive (Type 2).[35]

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 also obtain each word's "closest" relative word using pre-trained word embeddings from GloVe (Global Vectors for Word Embeddings). Finally, I remove all stop words that are not negation-based, stem all remaining words, and tokenize each document into a vector of uni-grams (single words) and 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.[36] However, I show that ADHD match values are robust to alternative visit inclusion

[35] DSM-V also denotes a "combined" type which is the combination of both Type 1 and Type 2 symptoms. By construction, the similarity between the patient note and the combined type vector will be a convex combination of the similarity between the Type I and Type II vectors. Therefore, I do not use the combined type in the NLP algorithm as it does not provide additional empirical information.

[36] 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

windows (see Table C2).

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 symptom text, I remove stop words (net negation terms), I stem each word, and tokenize documents using uni-grams and 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 where rows (i) represent patient documents, columns (k) represent uni-gram or bi-gram, and matrix elements (i,k) are the "tf-idf" values indicating the relative frequency and importance of uni/bi-gram k in document i.[37]

Next, patient-by-type specific match values, x_{is}, are calculated by taking the cosine similarity measure between the BOW row vector for patient i and the reference vector for ADHD Type $s = \{1, 2\}$. The cosine similarity measures the overlap between patient symptom words and ADHD defined symptom lists, weighting both by tf-idf (which down-weights highly common words) and by note length (so that longer notes do not artificially receive higher similarity due to chance word overlap).

Finally, for interpretability purposes, I rescale the resulting match values by subtracting the minimum and dividing by the range, normalizing values to be between 0 and 1. A value of 1(0) denotes the patient has the highest(lowest) similarity between their noted symptoms and the DSM-V defined symptom lists. I define the overall patient ADHD match signal as the maximum relative similarity across types: $x_i = max\{x_{i1}, x_{i2}\}$. The gender-specific distribution of these constructed values are plotted in text, Figure 3, with mean values

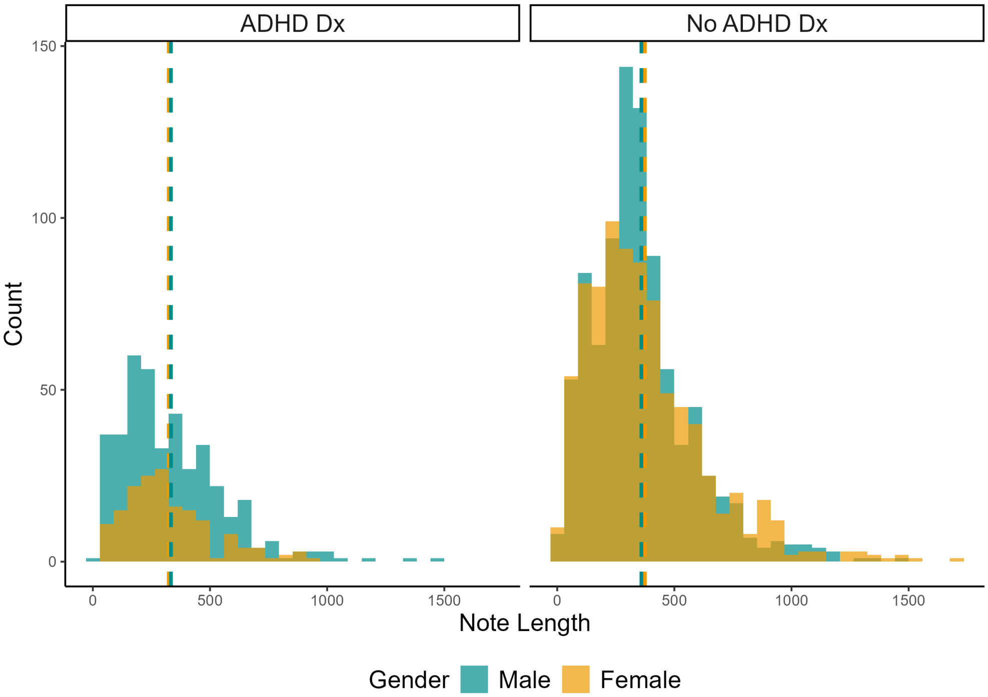
initial diagnosis. This suggests that physicians may be breaking up behavioral assessments into multiple visits and assigning ADHD diagnosis codes slightly before the assessment is fully complete.

[37] The "tf-idf" value is defined as $\frac{f_{ki}}{F_i} \times [1 + log(\frac{D}{D_k})]$ where f_{ki} is frequency of uni-gram or 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 uni-gram or bi-gram k.

in text Table 4. I note that while I use x_i in the main model estimation and results, I also incorporate the type-specific match values x_{i1} and x_{i2} in the supplementary analysis in Section 6.2.

Figure C2 highlights the similarity in note length for male and female patients who receive a behavioral assessment, both overall and by ADHD diagnosis. While behavioral assessment notes associated with an ADHD diagnosis are shorter than those without an ADHD diagnosis on average, the distribution for males and females in each category are similar.

Figure C2: Behavioral Assessment Note Length Distributions



Note: This figure presents the distribution of note length for patients with behavioral assessments (i.e., those with $Q_i = 1$), separately for those with an ADHD diagnosis ($D_i = 1$) and for those without an ADHD diagnosis ($D_i = 0$). Note length is determined after notes are pre-processed and words replaced with most common synonym. Male and female distributions denoted by blue and orange colors, respectively. Dashed vertical lines correspond to the mean note length for that gender-diagnosis category.

Table C1 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 most common words in behavioral assessment notes are: *plan, patient, assess, conversation, normal.* Words specific to male notes are: *fine, routine, game, sport, task* , and those specific to female notes are: *stressor, girl, period, worry, hair.*

Table C1: Outcome-Specific Predictive Words

Patients with ADHD Diagnosis	
Overall	*inattention, distract, insight, judgment, impulsive*
Male	*inattention, homework, distract, interfere, impulsive*
Female	*inattention, homework, task, interfere, distract*
Patients with High ADHD Match	
Overall	*task, impulsive, perception, interview, attitude*
Male	*task, impulsive, irritable, opposition, relax*
Female	*insight, peer, distract, impulsive, task*

Note: This table shows examples of the most frequent occurring words that are specific to given outcome but are not in the top 5% 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 5% 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 ADHD match correspond to values in the top tercile of ADHD match signals, x_i.

To assess the robustness to alternative encounter inclusion windows, I also construct ADHD match values for each patient based on: (i) all notes with $Q_{ij} = 1$ up to and including the initial ADHD diagnosis, (ii) all notes with $Q_{ij} = 1$ up to and including 30 days post the initial diagnosis, (iii) all notes with $Q_{ij} = 1$ up to and including 60 days post the initial diagnosis (the baseline), and (iv) all notes with $Q_{ij} = 1$ regardless of ADHD diagnosis or timing of visit. Table C2 displays the within-patient correlation across these different measures. Correlations are uniformly high, indicting that the ADHD match signal measure is robust to these alternative note inclusion widows.

Table C2: Patient ADHD Match Correlations

	<= Initial Dx	<= 30 days post	<= 60 days post	All Visits
<= Initial Dx	1.000			
<= 30 days post	0.994	1.000		
<= 60 days post	0.986	0.993	1.000	
All Visits	0.910	0.923	0.934	1.000

Note: This table shows the correlation across patient ADHD match values extracted from their behavioral assessment notes (or specified subset of notes). Measures include match values based on all $Q_{ij} = 1$ visits up to and including initial ADHD diagnosis (if diagnosed), up to 30 days post initial diagnosis (if diagnosed), up to 60 days post initial diagnosis (if diagnosed), and all visits with $Q_{ij} = 1$.

As a complementary validation exercise, I determine whether patients receive multiple assessments from different physicians, and if so, whether the match values are consistent across those assessments. To conduct this exercise, I first subset to patients with an ADHD diagnosis and select visits with $Q_{ij} = 1$ within three months prior to and two most post the initial visit with an ADHD diagnosis code. I then identify whether the patient had multiple assessments during this period, and if so, whether there were multiple visit providers in the record.

Among diagnosed patients, the median number physician assessments is 1. In fact, there are only 95 patients in the sample with an ADHD diagnosis that had more than one physician provide as assessment. Table C3 shows that, even in this more limited sample, ADHD match values are highly correlated across inclusion windows (and thus across assessments conducted by different physicians). As expected, the correlations are smaller than those in Table C2 as the vast majority of those in the full sample include only a single physician per assessment. That said, the correlations are still very high, ranging from 0.921 to 0.962, despite the fact that they reflect multiple assessments by different providers. This provides further evidence that the match signal consistently captures ADHD symptom match, and is not overly sensitive to variation in who conducts the assessment or the number of assessments included in the match.

Table C3: Patient ADHD Match Correlations | Multiple Assessments

	<= Initial Dx	<= 30 days post	<= 60 days post
<= Initial Dx	1.000		
<= 30 days post	0.950	1.000	
<= 60 days post	0.921	0.962	1.00

Note: This table shows the correlation across patient ADHD match values extracted from their behavioral assessment notes (or specified subset of notes) for the set of patients with multiple assessments and multiple providers. Measures include match values based on all $Q_{ij} = 1$ visits up to and including initial ADHD diagnosis, up to 30 days post initial diagnosis, and up to 60 days post initial diagnosis, where the later two measures include, by construction, assessments done by different physicians.

D Econometric Appendix

D.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 the physician's perceived tradeoff between the costs/benefits of over- and under-diagnosis (or, alternatively, the costs/benefits of deviating from uniform clinical guidelines).[38]

Let physician utility be defined by equation (D1), where S_i denotes the true ADHD health state as defined by the DSM-V criteria. Throughout, I refer to both *missed diagnosis* ($D_i = 0, S_i = 1$) and *misdiagnosis* ($D_i = 1, S_i = 0$) as "deviations from diagnostic guidelines" rather than "diagnostic errors," as the latter implies physician mistakes, whereas the former allows for the possibility of rational deviations.

$$u_i|\theta = \begin{cases} -1 & \text{if } D_i = 0,\, S_i = 1 \\ -\beta_\theta & \text{if } D_i = 1,\, S_i = 0 \\ 0 & \text{otherwise} \end{cases} \tag{D1}$$

The utility of "correct" diagnoses (i.e., in compliance with DSM-V guidelines) is normalized to 0 so that the model focuses on measuring factors that influence deviations from diagnostic guidelines. With the utility of missed diagnoses ($D_i = 0, S_i = 1$) standardized to -1, β_θ captures the physician's perceived (dis)utility from misdiagnosis *relative* to missed diagnosis. Importantly, this decision is made at the point of assessment, and therefore β_θ may also capture perceived dis(utility) from reversing a diagnosis down the road relative to waiting to diagnose later. This parameter is allowed to vary by patient gender, reflecting heterogeneity in these perceived diagnostic tradeoffs.

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 (D2), and follows from posterior ADHD risk in equation (4) and the DSM-V defined minimum diagnostic requirement, $\overline{v}$.

$$p(x,\theta) = Pr(v_i|x > \overline{v}) = \Phi\left(\frac{\rho_\theta x + (1-\rho_\theta)\mu_\theta - \overline{v}}{\sigma_\theta\sqrt{1-\rho_\theta^2}}\right) \tag{D2}$$

[38] 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.

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 (D1), $E[u_i|D_i = 1, \theta] = -\beta_\theta(1 - p(x, \theta)) + 0(p(x, \theta))$ and $E[u_i|D_i = 0, \theta] = -1(p(x, \theta)) + 0(1 - p(x, \theta))$.

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 (D2) 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 that defined by the DSM-V so that they diagnose patients on the margin of meeting ADHD diagnostic criteria. Intuitively, this suggests that physicians perceive that the cost of a missed diagnosis (or waiting to diagnose) is higher than the cost of a misdiagnosis, which is consistent with $\beta_\theta \in (0, 1)$ in (D1).

On the other hand, $\beta_\theta > 1$ implies $\tau_\theta > \overline{v}$. In this case, physicians will use higher thresholds that that defined by the DSM-V and will *not* diagnose patients on the margin of meeting ADHD diagnostic criteria. This suggests that physicians view misdiagnosis as costlier than missed diagnosis, which is consistent with $\beta_\theta > 1$ in (D1).

D.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. Throughout, I emphasize that while violations of these assumptions may bias the *level* of the model parameter estimates, so long as the violations are not gender-specific, the estimated *relative* differences in model parameters remain unaffected. As a result, the mechanism decomposition analysis in Section 6.1 and the implications of the Supplementary Analysis in Section 6.2 are not meaningfully impacted.

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 that 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 Figure C2 and Table C1, 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 both $\theta \in \{m, f\}$.

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 canceling 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 along with implications of the Supplementary Analysis results in Section 6.2 are not meaningfully impacted.

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: $v_i \sim N(\mu_\theta, \sigma_\theta^2)$.

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 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 multifaceted 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 guideline discretion 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 (D3), where ξ determines the deviation from accurate prior mean.

$$v_i \sim N(\mu + \xi, \sigma^2) \tag{D3}$$

If $\xi > 0$, physicians overestimate population mean ADHD risk, and $\xi < 0$ implies physicians underestimate population mean ADHD risk. 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., $\xi \neq 0$), 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:

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

Using the same utility framework, and letting $k = \frac{1}{\sigma\sqrt{1-\rho^2}}$, the new conditional diagnosis rate is defined by equation (D4), where $\tilde{\tau} = \tau - (1-\rho)\xi$.

$$\begin{aligned} P(D_i = 1 \mid Q_i = 1, x_i) &= \Phi(k\rho x_i + k(1-\rho)(\mu+\xi) - k\tau) \\ &= \Phi(k\rho x_i + k(1-\rho)\mu - k\tilde{\tau}) \end{aligned} \tag{D4}$$

The diagnostic uncertainty parameter, ρ, is also unaffected by ξ 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 $\xi > 0$, then $\tilde{\tau} < \tau$, meaning that my estimates of the perceived relative 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 physicians' inaccurate priors and their perceived cost of diagnostic errors/guideline discretion. 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 initial primary care provider (IPCP), *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 IPCP referral rate propensity.

There may still be concern that patients choose IPCPs based on unobserved factors that are correlated with ADHD risk.[39] 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 IPCP (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 IPCPs such that $Cov(v_i, \gamma_j^\theta) > 0$, 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 IPCPs (high x-axis value) have higher than average ADHD risk (high y-axis value), leading to a biased upwards extrapolation point at $\widehat{\gamma}_j^\theta = 1$. 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 IPCP such that $Cov(v_i, \gamma_j^\theta) < 0$, then my estimates of mean ADHD risk and utilization costs are biased downwards.

Similar to the full documentation assumption, if the IPCP selection assumption fails *equally* for both male patients and female patients such that $Cov(v_i, \gamma_j^m) = Cov(v_i, \gamma_j^f)$, then the gender parameter *differences* (column 3 in Table 6) still hold, and the mechanism decomposition analysis in Section 6.1 along with implications of the Supplementary Analysis results in Section 6.2 remain unaffected.

However, if there is a gender difference in the correlation between v_i and IPCP selection that cannot be controlled for with observables, then both parameter estimate *levels* and

[39] For example, those with familial history of ADHD might choose the same IPCP for their children/sibling. See ADHD Heritability discussion in the following section.

differences will be impacted. The direction of this bias depends on the sign and magnitude of this unobserved gender-specific selection, which is theoretically ambiguous and empirically untestable. Primary care provider choice and how it relates to the mental health referral process and child mental healthcare utilization are outside the scope of this paper, but are important topics for future research.

ADHD Heritability

Twin studies (which compare trait similarity between identical and fraternal twins) estimate the heritability of ADHD to be about 75% (Levy et al., 1997; Nikolas and Burt, 2010). This does not mean, however, that 75% of an individual's ADHD is caused by genetics and the remainder explained by environmental factors. Rather, heritability describes population level variability *within* a given environment. In other words, the results from twin studies imply that about 75% of the variation in ADHD prevalence observed *across* individuals can be explained by genetic variations, assuming environment fixed.

The exact genetic variations that are associated with ADHD are not fully determined. While genome-wide association studies (GWAS) have made progress in identifying genetic loci and single nucleotide polymorphisms (SNPs) commonly associated with ADHD, these only account for a fraction of ADHD heritability, suggesting rare variants and/or gene-environment interactions also play an important role in ADHD development (Grimm et al., 2020; Demontis et al., 2023).[40] Further, these studies also find that a large fraction of the genetic variants associated with ADHD are also shared with other psychiatric disorders, making it more difficult to isolate those associated with a single mental health diagnosis (Demontis et al., 2023).

Importantly, current evidence suggests that the heritability and genetic component of ADHD are very similar for males and females, indicating no substantial gender differences in the variability across individuals or the within-individual type of genetic influences contributing to ADHD risk (Nikolas and Burt, 2010). Therefore, if familial history of ADHD

[40] See also Thapar et al. (2013) and Kim et al. (2020) for overviews of this literature.

influences decisions of patients (or physicians) in the model described in Section 3, it likely does so equally for both male and females. Consequently, this does not alter the interpretation of the results in Section 6, which examine mechanisms underlying *relative* differences in ADHD diagnosis across gender.