Examining Bias in Machine Learning Tools for the Diagnosis of Attention-Deficit/Hyperactivity Disorder

Mahima Agrawal

Majors: Computer Science & Community Health

Center for Interdisciplinary Studies

Tufts University 2020

Acknowledgements

I would like to thank my thesis committee members Professor Fernando Ona, Professor Lenore Cowen, and Professor Michael Hughes for the countless time, support, advice and encouragement they provided me as I began to explore this topic. I would also like to thank Julie Dobrow at the Center for Interdisciplinary Studies for her assistance in developing this thesis, as well as the department as a whole for the opportunity they provide students to explore ideas in unique and diverse ways.

Table of Contents

CHAPTER 1: INTRODUCTION

The public health context of mental disorders as a whole is marred with institutional challenges, from the time invested in obtaining actual diagnosis to the magnitude of in person consultations and physician visits. Because of these struggles, the relationship between machine learning and healthcare—while still emerging—is a highly promising one, as it has the potential of reducing institutional barriers even as it magnifies the reach of medical interventions. Technology, however, is not immune to biases and concerns of equity of its own, an issue which becomes particularly important to recognize in the healthcare setting as it can cost patients certain benefits to their health and even risk their lives.

Understanding how to integrate machine learning into mental health interventions requires an understanding of how certain populations are disproportionately impacted by disease morbidity than others. Without this, the same influencers of harmful outcomes become encoded into practice, pushing disparities further. Over the past twenty five years the discipline of public health has generated a series of Social Determinants of Health (SDOH) frameworks to address the underlying causes of health disparities that are often hard to access, both in terms of knowledge and intervention. Within public health frameworks, SDOH refers to aspects of a person's environment—whether that be schooling, interpersonal relationships, legal restrictions of their state, and more—that drive health outcomes, whereas health disparities refers to differences in health outcomes, access, and attainment that are not only preventable, but in fact unjust. The World Health Organization (WHO) outlines an SDOH framework that highlights the role of structural determinants—such as legislation and cultural norms—and its interplay with individual circumstances—such as access to food and housing—as a means to understanding

influencing factors on disparities that appear downstream in healthcare settings (2010). Understanding the relationship between the development of machine learning models—a form of data analysis in which algorithms learn patterns in presented data in order to guide decision making—and these SDOH is crucial to effective and accurate public health interventions.

I approach the relationship between SDOH of mental health and machine learning through the lens of Attention Deficit HyperActivity Disorder (ADHD). ADHD is a neurodevelopmental disorder with an estimated prevalence rate of 9.4% among children (Danielson 2018). This disorder is characterized by innate differences in brain activity that typically increase an individual's difficulty with maintaining attention, managing tasks, or controlling impulses to the point that it impedes daily functioning or subverts what is considered "normative" behavior. Under current diagnostic criteria outlined by the Diagnostic and Statistical Manual (DSM), the condition may take the form of three presentations: inattention, marked predominantly by an inability to focus, hyperactivity, often described with impulsive behaviors or fidgeting, or a combined type (APA 2013). This disorder has rapidly evolved in the past two decades alone, with some studies indicating a jump in prevalence from 6% in 1998 to 10.2% in 2015 (Xu et al., 2018). ADHD provides a unique lens by which to examine the role of machine learning in health as there exists both a direct translation between several aspects of its clinical process (symptoms checklist) and model development (feature selection) as well as difficulty in diagnosing the condition due to underlying disparities.

This thesis seeks to examine and characterize the integration of machine learning and ADHD diagnosis to provide insight for future uses in public health—particularly through the lens of bias. In Chapter 2 I review the current criteria for diagnosis of ADHD as well as

document how it has changed in the most recent version of the DSM. I argue that underdiagnosis of ADHD is an equity issue due to its role as a gatekeeper to many of the accommodations and treatments that improve success and fulfillment. I then hypothesize ways in which factors such as stigma, access, and environment influence underdiagnosis in ways that may be inequitable across groups of different races, gender, and socioeconomic status. I note, however, that existing studies primarily concern ADHD under DSM-IV conditions and the extent to which conclusions hold under DSM-V must be examined further. Additionally, I survey new literature surrounding bias and equity in machine learning practices and outline different ways by which bias may be introduced into the development of a model. This provides the foundation for later chapters in which I discuss two cases in which machine learning models have been developed to assist practitioners in diagnosing ADHD.

Whereas chapter 3 outlines the methods by which the cases analyzed in later chapters were both identified and selected, chapter 4 presents an overview of the data itself, highlighting components of the model development involved in analysis. This chapter also includes novel model cards of each case in an effort to ease the model evaluation process and promote standardization of such practices across the field.

Chapter 5 codifies the two cases through a qualitative analysis, in which I thematically examine three aspects of the machine learning development process. My primary question that I seek to explore in this section is whether these models have a possibility of ameliorating or exacerbating the equity issues discussed in chapter 2—in which the outlined disparities are being perpetuated in a model of diagnosis based primarily in human interactions and exempt from machine learning interventions. Through these analyses I show that there exists risk of encoding

bias into these models, thereby teaching and even enhancing biases that already exist in the ADHD diagnosis process.

In chapter 6, I highlight limitations to my analysis in the previous chapter, posing questions regarding the impact, scope, and future of effective machine learning integration and fairness conceptualization in the problem of disparate outcomes of ADHD diagnosis. This is followed by chapter 7, in which I conclude my research by providing a series of recommendations for both machine learning researchers and public health officials who aim to continue work in this area.

CHAPTER 2: BACKGROUND & SIGNIFICANCE

2.1 Epidemiology of ADHD

Although concerns surrounding the increase in diagnosis rates has been met with much consideration surrounding overdiagnosis among children, research has highlighted a major concern of equity in the underdiagnosis of ADHD—particularly among racial minorities, women, and adult populations (Miller et al., 2009; Ramtekkar et al., 2010; Ginsburg et al., 2014). An examination of the social determinants of health that impact ADHD rates and diagnosis reveals these subpopulations have historically lower rates of ADHD diagnosis than others—a distinction which has been proposed to occur separate from disparities in actual, biologically salient prevalences of ADHD.

In contrast to concerns regarding overdiagnosis, which primarily surround the impact of unneeded interventions and medications (a threat to health in its own regard)—underdiagnosis poses a major equity concern because the diagnosis of ADHD itself acts as a gatekeeper to many accommodations designed to ease the daily actions of a person struggling with symptoms: for children this may include school accommodations such as extra time in testing and an adjustment to assignments or workload, and for adults this may be manifest as protection from work discrimination, social support groups, and adaptive behavior training.

Additionally, a lack of ADHD diagnosis has been shown to result in greater emotional difficulties and lowered self-esteem, which impacts social functioning, relationship building, and pursuit of daily activities (ADHD Editorial Board, 2019). People with ADHD of all age ranges also may benefit from assigned medications as a result of the diagnosis, however medication attainment marks a step in the clinical trajectory of the condition separate from that of diagnosis

and also produces unique concerns in regards to equity. As such, the following examination of literature surrounding ADHD disparities limits its scope to the diagnosis of ADHD itself and the underlying causes of such disparities that have been examined.

2.2 Diagnostic Timeline

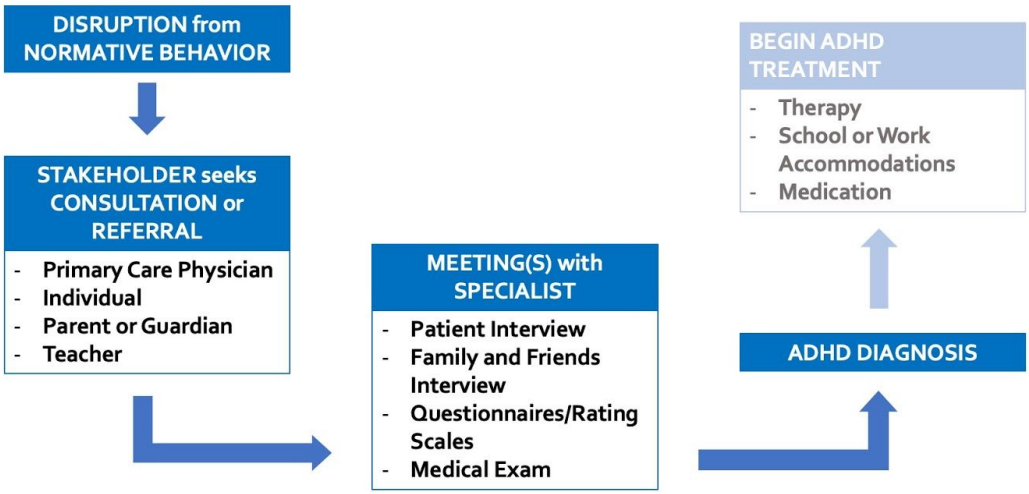


Figure 1: An Overview of the Clinical Trajectory of ADHD

Physicians Gualtieri & Johnson outline the primary structure and timeline of ADHD diagnosis under DSM-IV conditions in their assessment of its objectivity, of which the highlighted components and process are upheld under DSM-V conditions by ADDitude Magazine, an online resource for research and caregiver advice focused specifically on ADHD. Figure 1 presents a high level map of the clinical trajectory timeline, factoring in the final step of treatment into its structure as well. In examining only the process of ADHD diagnosis itself, however, several actions that remain instrumental in reaching future steps of the process are emphasized:

a) *Seeking a Consultation.* Prior to a meeting with a psychiatrist or specialist, an impairment in an individual's life that relays a need to seek intervention must be identified. Because

symptoms of ADHD are unlikely to present themselves in controlled settings such as physician visits, impairments typically present as disturbances in classrooms, home life, and relationship building (Gualtieri & Johnson, 2005). As such, while primary care physicians may be able to provide an initial screening and recommendation for consultation, it becomes primarily the responsibility of those in the individual's life to identify that they are struggling with behavioral and adaptive issues. For children with ADHD this role typically falls on teachers and family members, and for adults this expands to include the individual themselves (ADHD Editorial Board, 2019).

b) *Consultation Interview.* At this stage, either a general practitioner or a psychiatrist who specializes in ADHD conduct an evaluation on the patients by way of an in depth interview. Depending on the physician and the complexity of the case, this may range from a single one to two hour meeting with just the affected patient to a series of interviews with the patient and others in their lives including family members and caregivers (ADHD Editorial Board, 2019).

c) *Checklists and Rating Scales.* Depending on the age of the patient, checklists may be filled out by parents, teachers, or the patient themselves. These checklists are—of all the components of the diagnostic process, most directly linked to the symptom criteria outlined by DSM-V and include scoring for symptom severity, identification of presentation, and the like.

d) *Physical Exam and History.* The observing physician also conducts a full medical exam and overview of the patient's social, familial, and symptom history. These components are tools for the doctor to rule out other conditions with overlapping symptoms and are

particularly necessary for cases where ADHD diverges from the DSM-V checklist, such as has been noted for adults with ADHD (ADHD Editorial Board, 2019).

This entire process clearly indicates a complex and—at some level—personalized process to the diagnosis of ADHD. The DSM-V checklist, while a critical component of the diagnostic process, is in practice supplemented and on occasion even subverted by the thorough interview and history overview that clinicians conduct.

2.3 Changes in Diagnostic Criteria Over Time

The DSM-V was published in May of 2013. While the changes to the ADHD diagnostic criteria may appear to be subtle and minute, the new criteria distinctly impacts what types of individuals are now able to be diagnosed due to a reduction in required symptoms, a relaxation of the levels of severity and clinical impairment necessary to achieve diagnosis, and the removal of autism spectrum disorder (ASD) as exclusionary criteria (no comorbidity was allowed in DSM-IV). Beyond this, the criteria for the condition extended its age range from the 7 year old cutoff outlined in DSM-IV to 12 years, as well as included examples of adult ADHD as an attempt to better capture the scope of ADHD across age (APA, 2013). All of these changes mark a significant increase in leniency of the condition, likely to increase the rates of diagnosis by way of expanding who falls under the umbrella of 'diagnosable', but how these relaxations will impact underdiagnosed populations—if at all—remains unclear. Because of how recently this change has occurred, much of the research in this field is based on DSM-IV criteria and, while it may inform understandings of disparities and fairness measured in modern machine learning applications for ADHD diagnosis, cannot be directly applied to make conclusions on ADHD under DSM-5. Much of the literature available in regards to ADHD—even when written after

DSM-5's publication, often uses datasets from the early 2000's, and many popular data surveys (such as the Early Childhood Longitudinal Studies Program (ECLS) conducted by the National Center for Education Statistics) contain data beginning at around the 2011-2012 mark, with the ECLS study continuing and still in progress. Some research overseeing data after the switch to DSM-5 is available (see Coker, 2016; Danielson, 2018; Xu, 2018), and indicates that there has been a distinct reduction in disparities that have been historically observed. However, the research methodology of these studies indicates a utilization of survey based prevalence data such as the Nation Survey of Children's Health (NSCH) and the National Health Interview Survey (NHIS), which simply provides a snapshot of the state of ADHD in the United States and cannot directly relate any trends to this DSM change. Longitudinal cohort studies, controlled clinical studies, and the like may be able to provide more specific insight into the impact of DSM-5, but as of now such research remains sparse.

2.4 Disparities

Race[1]

The issue of racial disparities within ADHD diagnosis has remained a prominent conversation in the mental health community since the 1990's—when actual research regarding the relationship between race and diagnosis started to appear. Historically, ADHD has been attributed as a condition more likely to occur among white individuals, particularly among children and adolescents. In an analysis of the ECLS, where children enrolled in kindergarten in

[1] Within the context of race there often exists a conflation with notions of ethnicity—particularly as they may have overlapping influence on aspects of an individual's life experience. For the purpose of the research examined, categories of race follow the standards set by the U.S. Census Bureau, in which they remain relatively distinct from ethnicity except in the case of the 'Hispanic' category. As such, the categories for health research where hispanic identity is considered typically are: Non-Hispanic White, Non-Hispanic Black, Asian, Hispanic, and Other. It is important to note that this poses a limitation in understanding the impacts of racial disparities, particularly among those who fit multiple categories such as Afrolatin(x) folk.

1998/99 were followed until eighth grade, it was found that Black/African American children had a 36% lower likelihood of being diagnosed with ADHD as compared to Caucasian children in the study, and 50% when assessed with confounding variables (Morgan, 2013). Although this study utilizes relatively old data—all of the participants of this study were diagnosed under DSM-IV conditions—these results have been observed and reproduced throughout research from the first decade of the 21st century (Miller et al., 2009). It is only recently, from 2015 onwards, that this trend has appeared to shift.

The prevalence of black children and adolescents in the U.S. increased rapidly across several different interview surveys in 2016, actually exceeding the observed prevalence for white children and adolescents. In a weighted prevalence estimate of ADHD diagnoses reported in the 2016 NSCH, it was found that 10.7% of non-Hispanic black children and adolescents had a current ADHD diagnosis, as compared to 8.4% of their non-Hispanic white counterparts (Danielson et al., 2018). This is corroborated by results found in the 2015-2016 NHIS, where 12.8% of non-Hispanic black individuals reported having ever been diagnosed with ADHD as compared to 12% of non-Hispanic white individuals (Xu et al., 2018). This is a drastic change from the 2013-2014 NHIS survey, where weighted prevalence rates were 8.8% and 1..4% for non-Hispanic black and non-Hispanic white individuals, respectively (Xu et al., 2018). Of note however, is that this rapid growth in diagnosis among black individuals is not replicated among adults, where a racial disparity still persists (Fairman et al., 2020).

While this reduction in black-white racial disparity among children provides significant insight into the trajectory of prevalence rates for the future and as such requires further study into influencing factors, it does not explicitly disprove the historically researched and identified

sources of racial bias that have persisted in ADHD diagnosis. As such, it remains necessary for continued research in this field to examine the sources of these biases and explore mitigating factors.

Among these two surveys, it is important to note that Hispanic individuals had by far the lowest rates of ADHD diagnosis prevalence, with a weighted lifetime diagnosis prevalence of 6.7% in the 2016 NSCH (Danielson et al., 2018) and 6.1% lifetime diagnosis from 15-16 NHIS (Xu et al., 2018). While this highlights a very clear issue of underdiagnosis among hispanic communities, research into disparities into this subgroup remains and relatively under examined as compared to race through the lens of African American/black and Caucasian/white subgroups.

Gender[2]

Recent publications of prevalence data have indicated a male to female ratio in ADHD diagnosis of about 2.29:1 (Danielson et al., 2018)—with some reporting higher rates near 2.5 (Mowlem et al., 2019). Fairman et al. indicates an even stronger ratio, with logistic regressions providing a weight of 2.88:1 among youth 19 years old and below and 3.02:1 in adults (2020). While the relationship between differing ADHD diagnosis rates and gender has been suggested to be in part caused by biological differences in prevalence and presentation, there is a lack of research corroborating this idea. Further, in research that seeks to examine biological factors in ADHD diagnosis across gender, the extent to which these factors influence overall prevalence disparities remains unclear.

[2] It must be noted that in much of the literature reviewed, there does not exist a clear definition of gender—in much of society, policy, and even health care settings notions of gender remain both restricted to a binary classification of male vs. female as well as a conflation of gender identity, gender presentation, and sex assigned at birth. The latter definition is most utilized in research (particularly where medical records are involved), although this once again may be muddled by the role of self reporting demographic data in many of the studies examined. For the purpose of this thesis, gender is limited to a definition of sex assigned at birth.

For example, Ramtekkar et al., found that, in their study assessing for DSM IV like ADHD symptoms among families in Missouri, the ratio of male children of the families who presented with ADHD symptoms when compared to female children was 2.28 to 1 (2010). While this ratio reaches the lower bounds of what has been reported to be today's male to female relationship, Ramtekkar et al. highlights that it is significantly lower than clinic based studies—which are considered among researchers to more accurately represent diagnosis rates—published around the time of this paper (2010). What is particularly notable in this study is that—rather than survey the presence of an ADHD diagnosis itself—the study sought to identify symptomatology as a method to distinguish between actual prevalence and diagnostic prevalence in its analysis.

It is important to note several severe limitations to this study: some populations were excluded based on family size, African American participants totaled only 1.54% of the sample size, and difference in language used among telephones screeners may have impacted symptom disclosure. Even so, it provides an early example of a case in which biological factors alone could not account for the reported rates of disparity in that time period (2010). It remains unclear, particularly with a relaxation of diagnostic criteria through the DSM-V, how much of the currently observed gender disparities can be attributed to innate biological differences, and how much are exacerbated by other underlying causes.

Socioeconomic Status[3]

[3] SES intends to outline and define an individual or subgroup's standing in society, although the boundaries of this definition remain unclear. Several attributes to categorize and examine this SDOH are currently in use, including measures of income against the Federal Poverty Line (FPL), separation of income by quartile, educational attainment, housing status, and insurance status (Russell 2015).

Upon a first examination of the data it may appear that there exists no concern of underdiagnosis of ADHD across socioeconomic status (SES), particularly for those of low SES. Rates of diagnosis are highest for those below the federal poverty line (FPL) when compared to other income levels—12.9% among this group as compared to 10.2%, 10.0%, and 9.2% for those at or below twice, twice or below four times, and at or greater than 4 times the FPL respectively (Xu et al., 2018).

When examined through a different marker of SES separate from income, however, another dimension becomes clear. A data brief produced by the National Center for Health Statistics (NCHS) found that the prevalence of ADHD was highest among those with public insurance such as Medicaid (11.7%) when compared to both those with private insurance (8.6%) and those without insurance at all (5.7%)—where rates were the lowest (Pastor & Hawkins, 2015). Thus, while this does not indicate an underdiagnosis of low income individuals specifically, SES examined through the lens of insurance coverage highlights that those without insurance remain underdiagnosed compared to both the publicly and privately insured. Although much has changed regarding insurance coverage since this publication (the Affordable Care Act was enacted in 2010, and later policy expansion mandated the coverage of mental health services by both federally funded and private insurers), 28.2 million people reported a lack of insurance to the 2018 Census, 3.9 of whom are children (KFF, 2018). This only emphasizes how crucial understanding this disparity becomes in mitigating bias surrounding ADHD diagnosis.

2.5 Structural and Root Causes of Disparities

The examination of prevalence and morbidity of ADHD amongst the listed subpopulations highlight something quite unique: while much historical data of the previous two

decades has corroborated high levels of underdiagnosis for african american communities and women in particular, these gaps have rapidly been closing in the past five years—some even surpassing the other groups they were compared against. While the reason for this may come from a variety of sources—changes in criteria produced by DSM-V, a modification to how the surveys were conducted, higher rates of solicitation of a diagnosis—examining the structural underpinnings of these historically noted disparities still serves beneficial for two reasons: the development of a machine learning model—the initial premise of evaluating such biases in ADHD diagnosis—must take care to understand where trends in data may stem from in order to adjust accordingly, and a more even distribution of ADHD across subpopulations does not immediately indicate equitable diagnosis. This would only be true if rates of actual ADHD persist evenly across variables such as race and socioeconomic status, when research into risk factors of ADHD suggest otherwise[4] (Marshall et al., 2020).

Stigma

One of the researched theories as to the historically lowered rates of African American/black individuals with an ADHD diagnosis has been the role of stigma surrounding the condition among African American communities. In a series of interviews asking black caregivers of adolescent boys about their perceptions of ADHD, many interviewees voiced concerns regarding being both judged as a parent and the stigma that their child may experience as a result of the ADHD diagnosis (Evans, 2019). This in part stems from the individuals' beliefs about the root cause of ADHD—the condition's link to pre and post-natal care appeared to

[4] A prime example of this is the water crisis of Flint, Michigan. Flint—with a poverty rate of 40% and in which half of its residents are African American—experienced and continue to experience devastating effects of lead poisoning. In relation to the environmental risk factors of ADHD, the existence of environmental racism in the U.S. aggregates these risk factors in communities that are disproportionately low income and people of color.

indicate a reluctance to seek a diagnosis that may be perceived both as preventable and a reflection of the caregiver's actions as a parent (2019).

Mistrust of Doctors

The same series of interviews also highlighted another source of hesitation from these black caregivers: a lack of belief that ADHD diagnosis would result in the best treatment for their children. Many of the interviewees noted a significant fear of the medications associated with ADHD in a two-fold response: first, what the impact of these medications may be on the affected child's social functioning, their future risk of substance use, and its inability to solve rather than mask the problem and second, whether the doctor themselves is taking the proper measures to seek out alternative therapies and solutions. For this second point, interviewees seemed to disproportionately believe that clinicians would not put in the same amount of effort for children of color in the treatment process as they would for white children (Evan, 2019). This mistrust of clinicians among the black community—separate from whether this difference in treatment is actualized or not—is rooted deep in the history of African American's social and cultural experiences in the United States: from the use of black slaves in medical experiments and surgeries to the Tuskegee Study that spanned nearly half of the 20thcentury, American healthcare practices have born a massive amount of trauma unto the African American community and shapes this communities interactions with the current system.

Perception of Symptoms

It has been speculated that underdiagnosis of ADHD in girls can be somewhat attributed to adaptive responses to ADHD symptoms; research has indicated that girls are more predisposed to developing internalizing behaviors as a response to ADHD related classroom

difficulties. Internalizing behaviors, defined as negative behaviors turned on oneself rather than enacted on others, can include social withdrawal, self-harming actions such as cutting, irritability and nervousness, among others (Littman, 2012). Because these behaviors do not typically result in disruptions in the classrooms in the manner that externalizing behaviors (conflicts among peers, vandalism, angered outburst) do they are more likely to go unnoticed by teachers and parents—who act as the primary solicitors of ADHD diagnosis for children in their lives. Further, outlined examples of what may present as symptomology for ADHD does not explicitly state cases of internalizing behaviors as they relate to hyperactivity or inattention, and as such even in the instances that they are noticed by those same first level screeners, they may not be recognized for what they are—indicators of an underlying ADHD diagnosis. This is further supported by findings that indicate externalizing symptoms of ADHD among girls was a strong predictor of subsequent diagnosis—stronger even than between boys and externalizing behaviors (Mowlem et al., 2018). This has been suggested to stem from the fact that externalizing behaviors appear in stark contrast to normative, internalizing behaviors of girls. (Mowlem et al, 2018).

This same study also noted that while an increase in severity in symptoms correlated to a higher likelihood of ADHD diagnosis, the odds ratio of such a relationship was slightly higher among the female population of the study than the male population—indicating that there is a higher threshold of severity that girls are implicitly subjected to in order to receive ADHD diagnosis (Mowlem et al., 2018). Although this study was conducted on a Swedish population of children and adolescents, these raised expectations of impairment and deviance from the norm for female populations may provide an explanation for similar disparities seen in the United

States, suggesting that girls with both internalizing and less severe cases of ADHD are less likely to be diagnosed to their male counterparts (Mowlem et al., 2018).

Cost

The problem that many ADHD diagnosis models seek to solve is the very same one which introduces a major barrier to a highly affected population; the time spent in physician and psychiatrist observations, the cost of such visits, and the overall duration of the diagnosis process (which can take months for some) present a series of hurdles that—while those with higher income and health insurance are likely able to pass—are increasingly difficult, if not nearly impossible, for those who are uninsured or low income to move beyond. What this results in is a distinct lack of uninsured populations in the general prevalence of ADHD as well as within selected populations for clinical studies.

Gualtieri & Johnson describe a common outcome in the clinical field where—due to the diagnostic complexity of ADHD—primary care physicians often feel unable to diagnose the condition in patients when presented with the symptoms, referring the patient to a psychiatrist or other specialist for accurate diagnosis. While this likely results in an accuracy in diagnosis due to the specialist's expertise in the field, this specialist referral is unlikely to be a cheaply accessible endeavor; Gualtieri & Johnson state that in their neighborhoods of Chapel Hill, North Carolina and Charlotte, North Carolina "comprehensive psychological evaluations for ADHD can cost between $800 and $2,000" (Gualtieri & Johnson, 2005). This cost, particularly for those of lower socioeconomic status (SES) and/or those with either insurance that does not cover mental health specialists or no insurance at all, may be incredibly cost-prohibitive in soliciting and obtaining the ADHD diagnosis. The out of pocket cost of diagnosis alone—not considering

medications or treatment upon discovery of the condition—for an uninsured individual may blocker to even the initial stage of diagnosis solicitation.

Localization of Impairments

Another contributing factor to the disproportionate lack of uninsured individuals within ADHD data may be what is known as the "frog-pond effect". The frog-pond effect describes the social phenomenon in which an individual's excellence or failing is noticed at a much more distinct level when it differs greatly to the individual's relative environment. In the context of ADHD, this theory notes that a student struggling with academic achievement and attention issues is more likely to be identified as performing against normative behaviors in a school situated in a high income neighborhood as compared to a lower income neighborhood, where underfunding combined with an aggregation of poor health outcomes would allow for a student showing symptoms of the condition to go unnoticed. This provides further source of concern as several of the risk factors associated with ADHD development—including lead exposure—have been associated with lower SES (Marshall, 2020) and indicates that a disproportionately impacted subgroup likely remains underdiagnosed, as a result lacking in robust data representing such a population.

2.6 Implications

It is important to note that much of the causes listed above interrupt the diagnosis timeline as early as step one; structural and social barriers of stigma and cost, among others prevent individuals from reaching a point where they are seeking out a clinician opinion in the first place. This is true even where aspects of checklists and ratings scales have been noted as a source of disparities—while the personalized manner of diagnosis places a responsibility on the

clinician to view the context beyond what the rating scales may provide, those same checklists guide parents and teachers on what symptoms and severities to be aware of. Thus, they act as a preliminary screening tool that may prohibit certain populations from achieving eventual diagnosis.

One of the primary concerns surrounding ADHD diagnosis remains rooted in its subjectivity. While the use and provision of the discrete diagnostic criteria provided by the DSM-V encourages a standardization of the diagnosis process as well as guidance on the markers of ADHD, the symptoms which are described—including "often does not seem to listen when spoken to directly" and "often talks excessively"—simply cannot be measured objectively. Whether through an assessment by family members or parents, the patient themselves, or even a clinical psychologist, perceptions of what constitutes non-normative behavior may be and has been proven to be influenced by the environment in which the behavior is observed, the symptom identifier's own expectations of what is considered excessive, and perceptions on the need and legitimacy of the condition itself.

2.7 Bias in Machine Learning

Although the term "bias" holds several different meanings in the field of machine learning, I chiefly examine bias as it is defined under frameworks of Fairness—a newly emerging field aimed at measuring and ensuring equality across an algorithm. Where fairness in machine learning may be defined as an absence of discrimination—whether against an individual or a subgroup of a population—bias is thusly identified as a driving factor behind "unfairness", or the existence of such discrimination (Mehrabi et al., 2019). In doing so, understandings of fairness and bias closely align with prevailing definitions of health disparities—where

differences in outcomes become problematized based on the production of harm for a particular subgroup rather than the inherent presence of such a difference.

It is crucial to note, however, that the goal of any machine learning algorithm is to be able to take in data and 'learn' upon it; that is, identify patterns between characteristics of the data—referred to as features—and utilize those to assign an outcome. As such each and every machine learning algorithm fundamentally relies on the existence of some form of discriminatory factor—we must be able to infer some relationship between a feature and a classification in order to assign some outcome, an inference that is (ideally) generalizable, accurate, and consistent. Thus, when discussions surrounding the ethics of machine learning arise, we must first accept the necessity of at least some sets of assumptions.

Notions of fairness push against exactly what may be allowed within these sets of assumptions, and how the inclusion of certain biases—or the exclusion of others—has the ability to produce consistently harmful results for a particular subgroup of individuals (Chouldechava & Roth, 2018). Within the field of fairness, the following questions are posed to any algorithm: is this inferred bias accurate for all variations of the data? Is there, categorically, a group which faces an outcome disproportionate to what is expected? Or, even when it is expected, is it an outcome that is ethical to replicate?

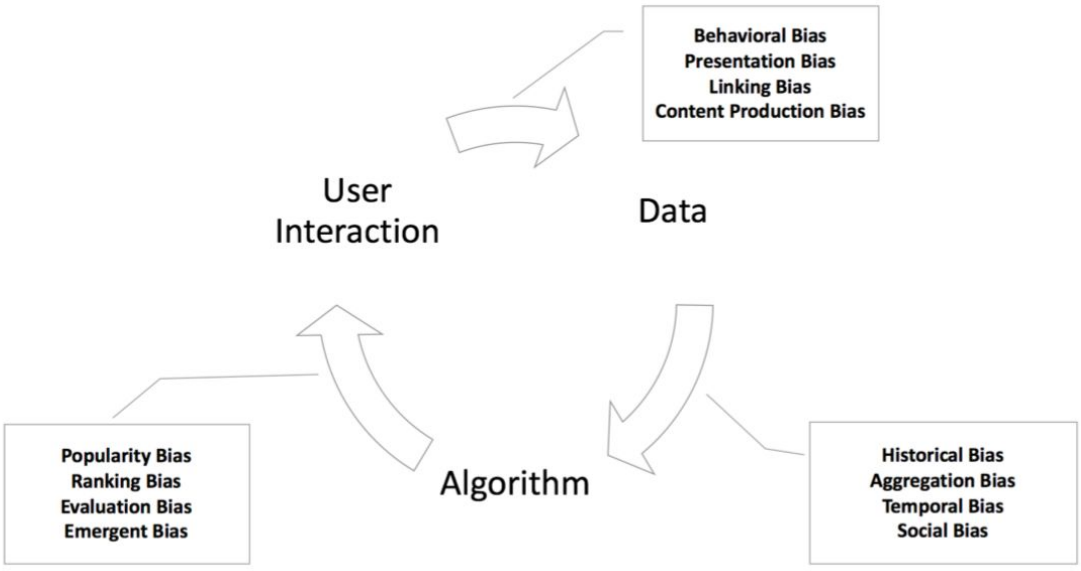


Figure 2: The placement of various biases in a feedback loop depicting the relationship between data, algorithm, and user (Mehrabi et al., 2019).

One crucial piece of the puzzle is understanding what exactly constitutes a harmful bias, and at which points in model development it may be introduced. As depicted in Figure 2, Merhabi et al. (2019) proposes categorizing key definitions of bias into the following three subgroups:

a) *Data to Algorithm.* This refers to sources of bias that may be introduced due to manners in which the accrued data upon which the model expects to be trained, tested, and validated impacts the decision making process. Within this scope, the method by which data is retrieved, the implications of generalizing a dataset to other populations, and the accuracy of the data itself may drive outcomes that disproportionately favor a particular subpopulation over another (Merhabi et al., 2019). For example, the utilization of self reported information as data may introduce bias if the communities that are asked to

provide such data are not representative, or if there exists room for subjectivity in response (such as product reviews or pain ranking scales).

b) *Algorithm to User*. This refers to instances of bias encoded through key facets of the algorithm itself, both in the selection of particular design decisions and in its interaction with general populations after development. This form of bias can occur for a number of reasons; for threshold based algorithms, where outcomes are assigned based on what score a piece of input achieves based on the values of its attributes, something as seemingly minute as a threshold cutoff set too high may lead to this form of bias. Additionally, the use of a proxy—referring to a feature that is included in the data in place of another because it is less expensive to attain—may introduce new biases if it is either not a strong enough indicator of the desired attribute or too strong of an indicator of another attribute as well, one which is now linked to the outcome that the proxy infers.

c) *User to Data*. This categorization highlights bias born out of the impact that a decision or outcome produced by a model holds in its effective application, and how those outcomes may subsequently change the environment in which the model is utilized. A commonly cited example of such is the role of crime prediction tools in policing, where areas identified as high crime are subsequently assigned greater police presence, further pushing up rates of arrest. This further integrates data which corroborates this area as high crime, even when it may have the same level of crime as other areas where police presence and subsequent arrests are low (Chouldechava & Roth, 2018).

While many of these identified sources of bias appear to be encoded into machine learning use cases through decisions made in the process of model development itself, the role and impact of

historical bias, which Mehrabi et al. (2019) classifies as data driven, constitutes a different form of harmful bias—one which sheds greater light onto expectations that the field of fairness has set for machine learning. Historical bias, defined as the "already existing bias and socio-technical issues" that may drive a model's decision making, highlights that even if one were able to ensure that methods of data aggregation, population sampling, feature selection, and evaluation minimized the introduction of bias, inherent inequities in the context of the model's use would be replicated by the algorithm.

One primary example of such a phenomenon is the use of algorithms in recidivism predictions. Chouldechava (2016) highlights ways in which a tool used to assess for the likelihood of recidivism disproportionately categorizes black individuals as high risk of recidivism despite satisfying the condition of predictive parity—where the probability of an individual being correctly assigned their outcome was the same across all subgroups of the sensitive attribute. Chouldechava (2016) uses this basis as an examination of other fairness metrics, however it must be noted that there is some truth to the algorithm's decision making process. Black individuals (specifically black men) have higher rates of recidivism than white men due to systems of over-policing that aggressively target black communities: in 2017 African American folk made up 33% of incarcerated individuals in the U.S., compared to 12% of adults in the United States (Gramlich, 2019). Additionally, officer reports on the inmate's attitude during arrest and completion of their sentence is a key feature in the data supplied to the recidivism algorithm despite encoded racial biases from officer's that it may contain, further driving up recidivism rates for black folk (Chouldechava, 2016). And yet it is clear that, despite the integration of such a bias into the system for which the algorithm is designed, this style of

decision making is inherently unethical, and the world of fairness holds the expectation that a truly just model would not replicate it.

The acknowledgement of historical bias and its impact on modeling highlights that—as model development grows to greater include fairness evaluation as a benchmark for a successful algorithm—we are not only asking for such algorithms to avoid encoding their own form of bias, but also that they seek to avoid perpetuating existing sources of prejudice, even at the cost of accuracy (Merhabi et al., 2019).

2.8 Potential Sources of Bias in Model Development

Biased Data

Perhaps one of the most important acknowledgements to make in the development of machine learning models for healthcare is that first and foremost, the data these models must be built on already contains a depth of biases, all of which play a role in the disparate impacts currently present in public health. With this, a greater ethical consideration comes to light—if we sit with the knowledge the data that we receive reflects the outcomes of a society that disproportionately prevents black communities, queer communities, homeless populations and multiple other protected classes from achieving positive health outcomes, what is the impact of training machine learning models with this data, particularly given that the expectation of an "accurate" model indicates predicting these same outcomes on similarly classified data?

Missing Data

One common approach in the evaluation of a model is to blindly select a test set from the original data source. While this is the standard approach for model development in cases where ensuring fairness and bias mitigation is not the identified goal, it was also reaffirmed by a

machine learning researcher from Flatiron Health in a presentation regarding the company's use of machine learning in cancer clinical trials (He, 2019). Although this is effective when the sample data that the test set is selected from is itself both very large and representative of all populations, what of models built with limited access to this data? This blind selection, in that instance, serves to actually limit the presence of subpopulations and underrepresented groups in the data, and further biases the model towards producing incorrect—or at the very least, non-evidence based—results upon these underrepresented groups.

Temporal Data

One of the crucial things about health is its malleability—not only an individual's oscillation between well and unwell, but an entire population's ability to shift health impact in one direction or the other over time. In fact, entire studies and careers have been dedicated to shifting disease burden and disparate impact to improve health outcomes. The social determinants of health reemphasize that in many cases—with cancers, infectious diseases, mental illness, even asthma—a person's level of risk and potential to heal cannot be solely explained by biology. Therefore, not only is health constantly shifting, it is in the fact goal of public health to permanently shift these for the better. In turn, in order to maintain a learning model that continuously produces results accurate to the data of the time period it is applied to, the model must in itself also be malleable—able to be retrained with new data to reflect current trends and expected outcomes in health.

The question then becomes: When does one retrain a learning model? How often must this process be done? What is the cost, over time, of continuously training such a model? From a

public health perspective, we also face the questions: how do we respond to the lag between changing trends in health and the data that proves such a change?

Missing or Expensive Features

While some features that may indicate a person's relationship to specific social determinants of health are mandated to be included in EHR data (such as race, and gender), there are several characteristics to a patient's identity that may play a role in disparate health outcomes but are not reported in EHRs. Examples include sexual orientation, transportation and housing access, and varying levels of social support. Even in the development of algorithms where EHR is not the only source of data—such as survey questionnaires, consumer data, and the like—many underlying causes to the outcomes we see in society remain obscured even to the individual experiencing them, and as such cannot be retrieved so simply. As Beutel asserts, identifying sensitive attributes as they relate to protected groups proves difficult as it not only may be increasingly expensive (in this context, 'expensive' can refer not only to a financial burden but time or processing power as well), but the attribute itself may not be present in the data at all (2017). An inability to identify and ensure the attributes presence in data presents barriers in de-biasing a given model.

Interdependent Features

In direct contrast to the previous point, models may also become biased by the presence of this sensitive attribute in the data, particularly where it is not accounted for. One common case study speaks to algorithms which blind the model to the race feature of the training data—the ultimate goal of this being to prevent the model from predicting outcomes such as recidivism or employability by known discriminatory factors. This racial bias, however, still becomes

introduced into the model due to the inclusion of zip code as a data feature which—due to the United States' historical institutionalization of slavery, housing discrimination, and segregation—acts as an indicator of race and socioeconomic status. This highlights how crucial of a role the understanding of the social determinants of health plays in the development of fair machine learning models.

CHAPTER 3: METHODS

The following sections of this paper will examine two cases of machine learning use for ADHD diagnosis. These two cases were found through a search of the Google Scholar database with the keywords “Machine Learning” “ADHD” and “diagnosis”. Of the cases that this search produced, the scope was limited to models which did not include classification systems for medication and treatment, solely focusing on diagnosis screening itself.

The purpose of these case studies was to examine how existing disparities and bias in diagnosis may transfer to algorithmic use, and as such models which utilized brain imaging or other biomarker data as features were removed. The two case studies were selected for the following reasons: the features of each case related directly to a rating scale or scoring system already in practice, each model resulted in a binary classification (this reduced the complexity of the model for the purposes of analysis), and both cases were developed after the transition to DSM-V—each case either explicitly discusses or is impacted by this change.

Upon collection of the data, I utilized the qualitative analysis practice of thematic coding—with a focus on narrative, content, and discourse analysis—to understand the researchers’ intended use case of their developed model along with elements of robustness of the model itself. Through this method of analysis I developed a high level summary of each case study, presented in chapter 4, as well as an analysis of identified themes, presented in chapter 5.

Furthermore, the analysis deliberately avoids evaluating these two cases through the lens of a comparative analysis. Each case highlights a complexity of ADHD through a chosen subproblem limited and unique in its scope: the first case examines the distinction between ADHD and overlapping conditions of ASD, while the second case assesses for ADHD diagnosis

in adult populations specifically. While this does not explicitly serve to benefit the specific research goals of this paper, it emphasizes the complexity and variance that exists within ADHD diagnosis and prompts deeper consideration to attempts at generalizing such a problem.

CHAPTER 4: DATA

Mitchell et al., have proposed the use of model cards in the assessment and reporting of models so as to provide an easily accessible and comprehensive summary of new and emerging research (2019). I followed their approach to build model cards for the two chosen case studies, listed at the end of this section.

4.1 Case One: Distinct Diagnosis of ASD and ADHD

In the years 2016 and 2017, Duda et al., published two papers in Translational Psychiatry discussing their efforts in developing a machine learning algorithm designed to distinguish between ASD and ADHD among children, due to not only each respective condition's high prevalence among this age group but also the overlapping symptomology and behavior reported by the two conditions. The burden of time, cost, and effort in complex clinical diagnosis of behavioral disorders were noted as motivating factors for the use case of the model, emphasizing the potential of a model able to discriminate between ASD and ADHD for mobile and self serve use, minimizing time spent in clinical interactions (Duda et al., 2016). The initial publication focused on identifying a subset of questions from the Social Responsiveness Scale (SRS)—a longform rating scale of 65 questions designed to identify ASD among children—that distinguished between an ASD and ADHD diagnosis with high accuracy. Their later publication focused on improving the previous model through a larger sample size and different methods of data collection, namely survey based data collection as opposed to archived data from the original paper (2016).

In the initial development of the model, archival data of children with a diagnosis of ASD were selected from the Simons Simplex Collection version 15 (the source of over 80% of the

dataset), the Boston Autism Consortium, and the Autism Genetic Resource Exchange, as well as data of children with a diagnosis of ADHD (Duda et al., 2016). Because each of the sources of data are autism specific projects, the available data regarding ADHD cases consisted solely of siblings of children with ASD in these datasets, whose medical information and responses to the SRS were present in these projects as part of the extensive family history (2016). Because of this, the number of cases for ADHD was significantly limited in comparison to ASD, with 150 subjects of ADHD to 2775 subjects with autism (2016). Of these groups, Duda notes that male subjects made up 83.9% of the ASD set and 62% of the ADHD set.

A 10-fold cross validation was performed on six different machine learning algorithms[5], in which the 65 features of the SRS were ranked to maximize predictive power while reducing similarity across features. This identified 6 questions consistently ranked as most valuable across all trials, after which the dataset was undersampled to produce an ASD to ADHD ratio in the training and testing sets of 1.5:1 (2016). Duda et al., notes that this was done with the intention of mitigating bias that may persist across gender as well as that may result from an unbalanced representation of ADHD.

The study found that while tree-based algorithms did not provide optimal results, the other four algorithms each performed with an area under the curve (AUC) of about .93, indicating a high level of accuracy. Due to the limitations of the archival data, however, researchers were unable to validate the model, which provides the basis for the follow-up paper examining each algorithm's performance on a crowd-sourced dataset (Duda et al., 2017). Researchers collected samples for the validation dataset through online surveys advertised via

[5] The six algorithms were decision tree, random forest, support vector classification, logistic regression, categorical lasso, and linear discriminant analysis (Duda et al., 2016).

social media, parent support groups, recruitment flyers, and email lists of organizations in the Palo Alto area (2017). Survey respondents were asked to submit their demographic information (which appears to not have been included in the feature development of the model nor displayed in the paper itself) as well as answer 15 questions from the SRS that were ranked highest in the feature training from the initial study. From this, the data was subsampled to include participant responses who fit the criteria of having either only an ASD or only and ADHD diagnosis (2017).

The top four performing models (with the addition ElasticNet algorithm) were run through three unique sets of experiments: First, each model was trained with the original archival dataset (subsampled in the way described in the initial paper) and then tested on the survey dataset after parameter tuning. Next, each model was trained first with the survey dataset and then tested with a subsample of the archival dataset. Finally, the two datasets were mixed, with the training and the testing set both being subsamples of this combined set (2017).

The results of these trials revealed a high level of variation between the three groups of experiments, and a Kolmogorov-Smirnov test to examine variance in question responses across the two data samples revealed that while 4 of the 15 questions had a significantly different answer distribution in the ASD samples, every single one of the 15 questions noted a significant difference in answer distribution for the ADHD sample (2016). Duda et al., conducted a similar comparison analysis on the datasets and found now statistically significant variation in response across gender and age, but other dimensions such as race, SES, etc. were not evaluated.

After training and testing the model only on survey data, it was confirmed that each of the algorithms—despite some generalizing most easily to the survey set—performed their worst with the survey data as the training set, indicating a high variation of response in this sample as

well as a poor fit to the selected features. Despite this, Duda et al., note that the initial experiment of archival test set and survey training set produced an AUC of 0.82, which presents promising use for clinical settings (2017).

4.2 Case Two: DSM-V Screening Scale for Adult ADHD

Following the updates to the ADHD diagnostic criteria under DSM-V, Ustun et al. sought to compare responses of the Adult ADHD Self Reporting Screening Scale (ASRS)—developed by the WHO—against the presence of DSM-V symptomatology. From this they sought to develop a shortened list of questions in order to assist and ease the burden of ADHD diagnostic screening with high levels of accuracy. The purpose of the algorithm is to rapidly optimize and validate the chosen subset of questions so that a human may be able to properly assess a risk of ADHD—the algorithm itself is not intended for use in clinical and public health settings.

Researchers utilized the Risk Calibrated Supersparse Linear Integer Model (RiskSLIM)—which optimizes feature selection through the use of small integer coefficients and a surrogate loss problem (Ustun & Rudin, 2019)—to select an optimal number of questions as well as their score weights from the ASRS. This model was trained and tested upon a sample selected from two datasets: the National Comorbidity Survey Replication Survey (NCS-R) 2001-2003 and respondents of a 2004-2005 telephone survey of a managed health care plan (Ustun et al., 2017). The data for feature selection—responses to the ASRS—were recorded when performing the initial surveys, and participants were recontacted in 2016 for a follow up telephone interview conducted by trained and licensed PhD clinicians or social workers (2017). In this follow up interview, participants were assessed for DSM-V like ADHD conditions through the use of the Adult ADHD Clinical Diagnostic Scale (ACDS).

The RiskSLIM models were estimated on an 8 question constraint with 4 distinct responses of severity for each question. They were optimized for a 10-fold cross validation against metrics of AUC and calibration accuracy (CAL). These models were trained upon a combined sample of both the NCS-R and managed plan samples, with the sample down weighted to account for the high prevalence of ADHD cases in the sample (2017). Evaluating the models in this manner revealed that the model with the best fit in both measures of AUC and CAL contained a 5 or 6 question set. This prompted the selection of the 6 question set presented below as the condensed ASRS screening set:

Table 1. Questions in the Optimal RiskSLIM *DSM-5* ASRS Screening Scale[a]

1. How often do you have difficulty concentrating on what people say to you, even when they are speaking to you directly? (*DSM-5* A1c)
2. How often do you leave your seat in meetings or other situations in which you are expected to remain seated? (*DSM-5* A2b)
3. How often do you have difficulty unwinding and relaxing when you have time to yourself? (*DSM-5* A2d)
4. When you're in a conversation, how often do you find yourself finishing the sentences of the people you are talking to before they can finish them themselves? (*DSM-5* A2g)
5. How often do you put things off until the last minute? (Non-*DSM*)
6. How often do you depend on others to keep your life in order and attend to details? (Non-*DSM*)

Figure 3: Optimal ASRS Question Set Selected by RiskSLIM model (Ustun et al., 2017)

These questions were first screened with the combined sample of NCS-R and managed care data. An evaluation of AUC values found that the highest levels of accuracy for this dataset existed at a thresholded cutoff score of 14—where those with a score of 14 or greater were assigned a positive outcome of diagnosis (2017). Seeking to corroborate these findings on a set of clinical data, 300 participants were collected from a 2014-2015 clinical sample through either the NYU Langone Adult ADHD Program—which was offering free evaluations to those who

had heard of the program through physician referral or media recruitment—or primary care offices near the school to be used as the validation data for the selected question set (2017). After undergoing the same interview process as participants from the NCS-R and managed health plan, the clinical samples were evaluated under metrics of accuracy, sensitivity, specificity, and positive predictive value (PPV). This analysis found that while a 13 score threshold marginally increased sensitivity—defined as the number of true ADHD cases that were captured by the model—it resulted in a major drop in PPV due to an overestimation of the prevalence of the condition (2017). Thus it was asserted that a 14 value threshold is necessary in clinical samples to prevent an increase in false diagnoses (2017).

Due to noted limitations of age restrictions and health plan types in the general populations sample, Ustun et al., recommend further validation of the question set through the use of expansive clinical samples including from primary care settings or of workplace identified cases (2017).

4.3 Model Cards

Discriminatory ASD and ADHD Classification

Model Details

- Comparison of ENet, Lasso, LDA, Ridge, and SVC performance upon various combinations of survey-retrieved and archived SRS data

Intended Use

- Distinguish between ASD diagnosis and ADHD diagnosis among children.
- Act as a guide to accelerate clinical observation process involved in ASD diagnosis
- Not intended to provide an actual DSM-V diagnosis

Factors

- Behavioral questions pulled from Social Responsiveness Scale—15 questions identified by previous study to be highly ranked by mRMR criterion in classification models of ASD and ADHD diagnosis

Metrics

- AUC evaluations of ROC of each learning algorithm over various data sets, identifying accuracy and variation of classifier performance.

Training Data

- Trial 1: 100 randomly chosen subsamples of the archival data
- Trial 2: All survey data
- Trial 3: subsampling of both archival and survey data (50 rounds of two-fold cross validation)

Evaluation Data

- Trial 1: All survey data
- Trial 2: 100 randomly chosen subsamples of the archival data
- Trial 3: subsampling of both archival and survey data (50 rounds of two-fold cross validation)

Ethical Considerations

- demographic data was recorded but not utilized in model evaluation
- Questions in survey limited only to those ranked highly on limited archival dataset
- Binary classification prevents outcome of co-occurring ADHD and ASD diagnosis

Caveats and Recommendations

- consider performance on patients with comorbid ASD/ADHD or no diagnosis
- Evaluate accuracy of classification across race and socioeconomic status

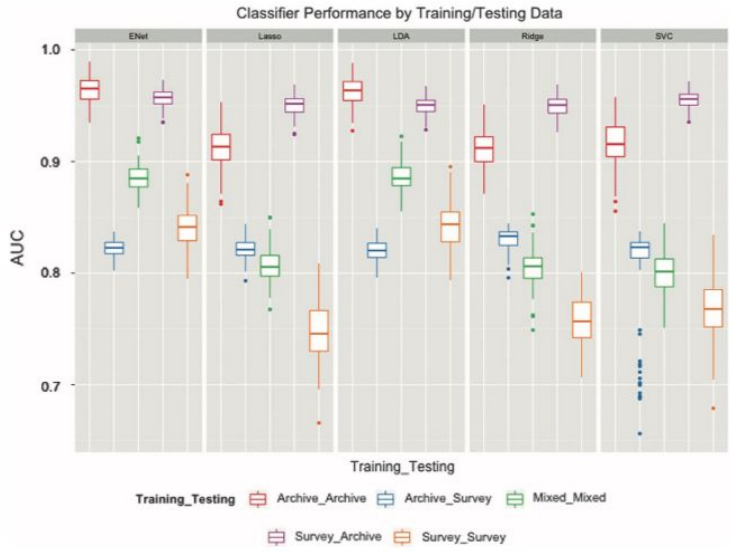


Figure 4. Performance of the five machine-learning models on five different training and testing set combinations across 100 trials.

Model Card- DSM-5 ASRS Screening Scale

Model Details

- Validating DSM-V to the Adult ADHD Self-Report Screening Scale (ASRS)
- Utilized Risk-Calibrated Supersparse Linear Integer Model (RiskSLIM) for scoring optimization

Intended Use

- Distinguish between ADHD cases under DSM-V criteria and non-cases
- Improve the operability of ASRS

Factors

- For Risk-SLIM model: 29-question ASRS
- Subset of 6 questions from the ASRS, with questions focused on inattentiveness, hyperactivity, and non-DSM-5 symptoms of executive dysfunction.

Metrics

- AUC and Mean Calibration Accuracy for Risk-SLIM Model across question set size
- Predicted Prevalence, Sensitivity, Specificity, AUC, and PPV evaluation of score thresholds ranging from 13+ to 17+ in home survey sample and physician referral sample

Training Data

- Household sample from National Comorbidity Survey Replication (NCS-R)
- Telephonic survey sample from population subscribed to managed health care plan

Evaluation Data

- NCS-R and managed care sample (weighted to account for oversampling)
- Clinical sample through NYU Langone Adult ADHD Program

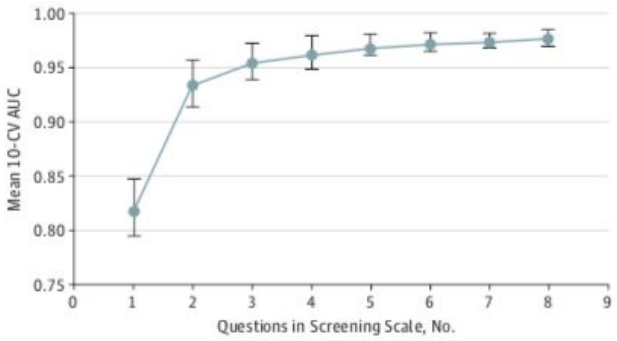
Ethical Considerations

- Managed care data is unlikely to include any data on uninsured individuals
- General population sample is relatively old, may not contain variation in ADHD present today

Caveats and Recommendations

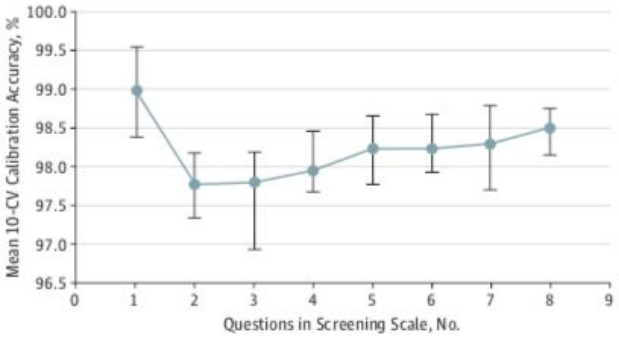
- Validate an a larger, more current community based sample to better represent changes to population
- Consider language of questions and ability to detect complex or unique presentations—particularly across gender

Figure 1. Ten-fold Cross-Validated (10-CV) Area Under the Curve (AUC) vs the Number of Questions in the Screening Scale



Pooled National Comorbidity Survey Replication and managed care samples (n = 337). The reported range represents the highest and lowest values of mean AUC across the 10 separate folds for 8 questions.

Figure 2. Ten-fold Cross-Validated (10-CV) Mean Calibration Accuracy vs the Number of Questions in the Screening Scale



Pooled National Comorbidity Survey Replication and managed care samples (n = 337). The reported range represents the highest and lowest values of mean calibration accuracy across the 10 separate folds for 8 questions.

CHAPTER 5: RESULTS & ANALYSIS

The intention of this analysis is to highlight various points in model development for ADHD diagnosis that bias may be introduced. Typically, at the final stage of model development, such biases are examined and confirmed through the use of fairness equations, which include calculating false positive and negative rates (FPR/FNR), equality of opportunity, or other similar measures across subpopulations. While both cases confirm the accuracy of the model through area under the curve (AUC) metrics and specificity/sensitivity calculations upon the test data as a whole, they lack a comprehensive examination of trends across these subgroups on the final classifications produced by the algorithm (Duda, 2016; Duda, 2017; Ustun, 2017). As a result many points in my analysis, rather than focusing on sources where actual bias has been encoded, highlight areas of the models' development that may pose a risk to vulnerable groups, and either note a need to evaluate these steps with an understanding of fairness and bias or conceptualize what others who are attempting to build a similar model may need to consider. Such areas have been codified into three themes: data collection, feature selection, and classification metrics.

5.1 Data Collection

Data Collection refers to the manner in which the data that a model is built upon—both training and testing—has been retrieved, and what this may indicate about the selected dataset itself. Because of a model's inclination to classify at the highest attainable level of accuracy, algorithms that learned the weight of their features through training upon a dataset in which a specific subpopulation is underrepresented are more able to achieve high levels of accuracy by prioritizing accurate classification on the majority group. This results in a lowered rate of

specificity and sensitivity to protected or minority subgroups, which—in the case of a tool such as ADHD diagnosis where a positive classification indicates an attainment of some benefit—not only broadly harms such a subgroup, but perpetuates rates of underrepresentation in future testing and development of the model. As such, the manner in which data collected—which can become skewed through clinician bias, survey bias, and other data attainment processes—remains a potential source of bias.

ASD/ADHD Case.

The archival data was pulled only from autism specific data sources, and researchers note that due to this all ADHD samples retrieved in the archived data—rather than representing a randomly selected population—consist solely of siblings of the individuals with autism retrieved from these sources (Duda et al., 2016). Due to emerging research indicating a genetic relationship both between ASD and ADHD (Stergiakouli et al., 2017), an implication to this limitation of the dataset is that individuals whose diagnosis of ADHD does not have a hereditary component may be underrepresented. Despite studies to indicate that genetics dominates much of the risk of developing ADHD with hereditary etiology nearing 75% (Faraone & Larsson, 2018), environmental factors such as toxins and maternal substance use remain closely linked influencers of ADHD that are now unlikely to be included. While there is no research indicating that the "causes" of ADHD and ASD results in varied symptomatology, the inability of the archived data to fully represent the affected community poses a risk to the true accuracy and efficacy of such a model.

Additionally, the majority of the samples in this archival set were sources from the Simons Simplex Collection (Duda et al., 2016). The Simons Simplex Collection (SSC) is a project within the foundation Autism research initiative to collect genomic data regarding ASD. Due to the wealth of clinical information needed in order for an ASD case to be included into such a project (not limited to full medical exams, family medical history, and aforementioned information surrounding genomic data), this dataset requires some form of extended clinician interaction and referral into such a project (SFARI, 2011). As a result, the SSC lends itself at a greater risk of clinician and referral bias—where cases of a condition that are both more severe appear at higher rates than what is representative of the condition in the general population. Additionally, as a rigorously managed database, there are several exclusionary criteria to the dataset (neonatal complications, family history of psychiatric disorders, down syndrome, etc.) that leaves a significant portion of ASD populations underrepresented (SFARI, 2011). Through this it is clear that the training such a discriminatory model on the archival dataset may leave the algorithm to deprioritize non-typical cases of both conditions in an effort to match these highly specific and localized representations of ASD and ADHD.

Duda et al., aims to mitigate the potential bias the archival data may introduce by validating their model against survey data designed to greater represent the general population of both conditions, but in doing so find that the accuracy of the algorithm decreases across all forms of linear modeling in this process. This confirms that the archival data likely diverges greatly from the actual data present in the survey set, and highlights a greater concern to all other aspects

of the model—particular feature selection—that was built upon the information obtained from the algorithms' performance on the archival dataset.

ASRS Case.

The data collection methods of this model also face some limitations in its data sample. In the training and testing of the RiskSLIM model two datasets—intended to represent the general adult population—were utilized, however both contained participants pulled only from between the years of 2001 and 2005. While data collection for this model does involve reinitiating contact and following up with a diagnostic interview designed to match current conditions, these two datasets exclude any person who recently became an adult between the years of 2005 and when the study was conducted in 2016. This may be a point of concern due to the malleability of ADHD symptoms—ADHD has been shown to change in presentation across various stages in a person's life, even through adulthood. While the ASRS data originally collected may provide a snapshot into early prevalence—questions within this screening scale specifically ask for the presence of symptoms in the previous six months (Ustun et al., 2017)—it remains possible that answers indicated then no longer coincide with the presentation and expressed severity of symptoms in the day to day of those interviewed later on in their life.

Of similar concern are the limitations of the (albeit large) managed care plan. Extensive details were not given to the nature of the plan, so it is not apparent whether the managed care sample included both those who were publicly and privately insured (Kessler et al., 2007). Even so, this plan excludes people who were without insurance at the time of the initial survey, a

group that has already been shown to face disproportionate barriers to diagnosis. As such, the validation of this model requires an examination of socioeconomic trends in its data and the subsequent impact on both feature selection and outcomes.

Overall the actual data collected itself appears sufficiently robust—each participant in the general population sample underwent a diagnosis-like interview process under DSM-V conditions—and the NYU Langone clinical sample consisted of individuals included in the study by way of both media recruitment and physician referral, which may mitigate concerns regarding clinician bias.

5.2 Feature Selection

Features refers to the attributes of information that are utilized within the model, where such a model learns patterns and relationships between the various combinations of these provided features in order to make an accurate prediction. Often, the features selected act as a proxy to the actual data model developers are hoping to observe, but what is interesting about ADHD diagnosis itself is that the underlying condition of the neurological disease and the manner in which it impacts the brain remains relatively unknown. As such, the current diagnosis system itself uses questions regarding symptoms to identify a case of ADHD—diagnosis of this neurological condition, even outside of the machine learning context, inherently utilizes these questions as "proxies" to the actual biomarkers that have not yet been attained! What comes with this is a tension regarding the applicability of such symptoms: due to the variance in ways that ADHD may present itself (from the axis of inattention vs hyperactivity to the internalization of

symptoms in women), what the proxies actually are, and what information do they convey, becomes crucial., This is particularly true for algorithms aimed at ADHD diagnosis, as they seek to reduce the number of features (in public health terms, rating scale questions) required to achieve high accuracy in classification upon this relatively heterogeneous condition. Within such an effort to generalize the question set exists the potential to skew results towards one specific presentation that fits the chosen features best, and in turn must be examined for bias.

ASD/ADHD Case.

The 15 questions were chosen from the Social Responsiveness Scale—a scale primarily used in diagnosing ASD—because they were heavily weighted to indicate a distinction between ASD and ADHD among the data the original iteration of the model was developed upon (Duda et al., 2016). Because the purpose of my own research was to primarily examine the diagnostic process of ADHD and potential sources of bias within it, difference in presentation of ASD across vulnerable subpopulations exists out of bounds of the scope of this paper. Despite this, the manner in which these questions were selected and its continued use may still be examined for potential avenues of bias.

As was previously identified, the archival dataset utilized in this model's development contains a very limited sample population of both ASD and ADHD—the ASD subset is likely to have undergone clinician and referral bias, and the ADHD population contains only individuals who are relatives of the ASD population. This is an incredibly limited scope, one which does not represent the prevalence and presentation of ASD and ADHD across the general population/ Because of the use of solely archival data in the initial development of the model, a more accurate descriptor of the 15-question set would not be distinguishers between ASD and ADHD,

but rather distinguishers between ASD and ADHD among family members. Duda et al. acknowledge these limitations, as this is the premise for the crowdsourced validation of the model as a follow up.

What is concerning about these questions is that they are used as the sole set of features for the crowdsourced validation of the model—in collecting the survey responses for the expanded dataset, participants were asked only the 15 identified questions as opposed to the filling out the entire Social Responsiveness Scale. Through this decision, model developers operated under the assumption that the original question set also presents an optimal fit for this much more diverse population, and compared various linear models upon this assumption. This was proven incorrect by the Kolmogorov-Smirnov test, where despite finding no significant difference in response across gender and age, each of the questions in the dataset produced significantly different response distributions between the ADHD samples from both datasets (2017). And yet, Duda et al. still advocate for the potential clinical use of the model, failing to consider that the differing distribution may indicate a pattern across a particular subgroup, one which may be falsely diagnosed at a disproportionately more frequent rate.

ASRS Case.

In this model, researchers found that the RiskSLIM algorithm maximized its accuracy at about the 5 or 6 question mark. As such, 6 questions were chosen for the resulting model: one for inattention, one for hyperactivity, two questions not related to the DSM-V, and two additional DSM-V questions (Ustun et al. 2017). These questions typically revolve around difficulty with tasks and activities, which presents a unique contrast to the emotion-infused lens by which researchers have noted ADHD presents itself in women. For example, ADDitude Magazine's

informal ADHD screening scale for women, which although not intended to be a replacement for clinical diagnosis has been reviewed and confirmed in its utility by medical professionals, focuses its questions primarily on emotional response to social functioning rather than physical actions. Some of the questions in this screening test include “Do you feel overwhelmed in stores, at the office, or at parties?” and “Do you shut down in the middle of the day, feeling assaulted? Do requests for ‘one more thing’ put you over the top emotionally?” (Solden, 2019).

These questions, while not conferring the abject and clinical manner by which DSM-V criteria may be written, utilize emotional language as this has been shown to fall in line with the presentation of ADHD that women experience most. The distinct difference between these questions and the questions selected by Ustun et al., indicates a potential source of bias in the model, one which prioritizes and favors symptoms that match a “typical”, externalized presentation of ADHD, while misdiagnosing those whose symptoms are not fully encompassed by this six question set (a group more likely to be dominated by women).

While this concern is in part mitigated by a further examination of the original screening scale under DSM-IV, which reveals that the six-question set utilized at that stage accounted for dichotomous responses across various demographic attributes including race, gender, and socioeconomic status, it is not explicitly made clear if that same evaluation is performed on the question set selected by RiskSLIM (Kessler et al., 2007, Ustun et al., 2017). Even if it were, a point of tension remains that the landscape of ADHD diagnosis for women has changed drastically since 2003; the increase in the growth rate of female ADHD diagnoses over the past ten years has been attributed not just to an increase in acceptance of a woman’s ability to have the condition, but also a fundamental change in the way symptoms are understood to manifest in

women and girls. While the use of DSM-V like diagnostic interviews may take proper steps in accounting for this unique set of presentations among women—particularly among a dataset that is noted to be 50.5% women (Ustun et al., 2017)—this study could benefit from collecting participants from a more current sample in order to ensure broad representation of ADHD as it exists today.

5.3 Classification Metrics

A classification metric in this context refers to the actual assignment of an outcome based upon values passed through to the features, both in a manner of how the outcome is selected and what possibilities an outcome may encompass. Both of the identified cases utilize a preexisting rating scale as features—where a particular response to a question on the rating scale is associated with a particular score. They both also produce a binary classifier, where the outcome always is one of two things. As such, the models utilize some form of cutoff or value analysis to establish when a score indicates a certain outcome or the other. For the purposes of the ASRS screening model, the implications of the chosen threshold will be analyzed, whereas for the ASD/ADHD discriminatory tool the two values of the classifier itself will be examined.

ASD/ADHD Case.

Under this model, only the following two results may be produced: an indication of ASD risk, or an indication of ADHD risk. What is missing from this discrete plotting of outcomes are classifications of no presence of ADHD/ASD at all, or a comorbidity of both in the individual (Duda et al., 2016). Although developers of this model acknowledge that the co-occurence of both ASD and ADHD is now possible through the changes introduced by DSM-V, both the data cleaning process conducted on the crowdsourced survey data and the inability of the model itself

to assign a risk of comorbidity limits the actual scope of diagnostic complexity that the model is able to achieve.

Above all else, this poses a threat to the efficacy of the model, as it has not been tested and validated on data that matches what it may see in a real world setting—as of yet it has not been proven high levels of accuracy on adversarial data in which co-occurrences of disease cannot be scrubbed out. Beyond concerns of accuracy, the limitation of a discrete, binary classification may disparately hurt individuals seeking a diagnosis in the following ways:

a) *Clinical Care Timeline.* While a reduction in time and physician visits is crucial to improving the quality of care that those with these diagnosis may receive, it is important to acknowledge that a diagnosis of one condition—despite its similarity in symptoms—remains distinct from another because the treatment, medications, educational response, disability services, and more following such a diagnosis differ. For example, Leitner notes that an ADHD diagnosis among children typically results in training for parents focused on behavior management, whereas with ASD diagnosis the treatment is focused around parental education surrounding skills to promote normative social functioning (2014). For those with ASD/ADHD comorbidity—where educational achievement and severity of symptoms are worse than those with either of the conditions alone (Leitner, 2014)—attainment of a treatment which accounts for both disorders is especially important. Although the effective outcome of this is unclear, it is important to consider that an individual who may require dual forms of treatment and care due to a comorbid diagnosis might, under the currently established model, be sent towards a

continuum of care based on their singular diagnosis which prevents the integration of a secondary diagnosis, inhibiting the level of overall benefit they may receive.

b) *Magnification of differing disparities.* While ASD and ADHD have indicated similar symptomatology, prevalence rates and disparities for the two are vastly different. For example, a lower family income is significantly associated with higher rates of ADHD diagnosis (Pastor & Hawkins, 2015) among children but lower rates of ASD diagnosis (Durkin, 2017). Because the algorithm is designed to compare severity of ADHD against severity of ASD in order to produce its classification, any bias encoded into the questions regarding symptom severity and its relationship with socioeconomic status may be magnified to consistently classify low income individuals with ADHD even where co-occurrence exists and symptoms of ASD may be present.

ASRS Case.

While the question set produced by the model allows for a range of scored results, the intention of its use as a binary classifier is met by the inclusion of a threshold. Ustin & Rudin highlight this as one of the strengths of the risk scoring model—that the cutoff score can be adjusted based on the specific needs of the use case, and the algorithm itself does not impede on the decision makers ability to do so (2019).

Although tests run on both the general population set and the clinical sample set resulted in the same optimal threshold score, the notion that training data, rather than an external review board, may be the ultimate decider of the cutoff presents an interesting point of analysis. On one hand, it completes the task of establishing a "non-normative" level of severity based upon already existing examples of such, reducing costly time and analysis. On the other, it introduces

the possibility of a skew in said training data, overfitting the model to a thresholded score that is either too strict or too loose. In the context of ADHD, this can manifest in the over—or rather, incomplete—inclusion of underrepresented and sensitive subgroups. For example, Miller et al. noted that in a study conducted where teachers assessed all boys in their classroom for ADHD regardless of race, it was found that African American boys were more likely to have higher levels of symptom severity (2009). A similar phenomenon is noted by Mowlem et al. for girls—where diagnosis among girls was associated with a slightly increased severity of symptoms (2019).

While the underlying causes for both of these trends may differ, what the data indicates is clear: it is not so much that individuals of these subgroups are less likely to experience ADHD along with lowered symptom severity, but rather that those who do are largely missing from diagnosis and prevalence data. Thus, when considering how this may impact aspects of model development such as threshold selection a point of concern arises: *even* in cases where marginalized groups are included, if the scope of the data upon this group is incomplete then the model is at risk of skewing upwards to a stricter cutoff, feeding back into the cycle where presentations of the condition with lowered severity are further unable to access diagnosis.

CHAPTER 6: DISCUSSION & CONSIDERATIONS

While the two case analyses provided varying degrees of robustness regarding the setup and utilization of each model, there remained a distinct lack of confirmation to its ability to avoid perpetuating bias. Despite both cases indicating moments at which gender bias and oversampling was accounted for (in calculating a weighted prevalence, examining distribution of question responses across age, etc.) they fundamentally lacked a presentation of the results across subgroups as well as a comparison of trends as it relates to known disparities in the problem. This confined my analyses to remain purely speculative and even overly critical, as it emphasized that frameworks of fairness and equity still exist separate from the development of a model rather than an integrated step of the process.

Despite this, the research that was conducted both on the background of ADHD disparities in Chapter 2 and on the case studies produced a series of questions that—while may not be easily answered—I find necessary to consider as the machine learning community further seeks to integrate algorithms into the ADHD diagnostic and clinical process.

6.1 Biological vs. Non-Equitable Bias

Research has indicated that while a significant portion of the ADHD diagnosis disparity across gender is due to a lack of understanding and engagement with the different presentations of the condition in girls as compared to boys, the findings produced by Ramtekkar show that even when diagnosis itself is set aside and symptom assessment alone is examined, girls are still less likely to present ADHD-like symptoms than boys (2010). As such, it has been posited by many clinicians in the field that from a biological perspective, ADHD—at least by the standards

of the current definition—appears more frequently in men than in women, particularly among children.

This provides a host of complexity to existing methods of ensuring fairness—in these definitions, the expectation of some form of "equality"—whether that be through probability of classification, false positive rate, or the like—asserts that in the process of producing ethically sound models, protected subclasses must achieve a level of sameness. In the context of ADHD diagnosis across gender, however—and likely across many other protected class of which the research remains limited—seeking out sameness would actually introduce a new form of bias into the classifier, one which does not reflect the innate differences among populations affected by this condition.

While uniquely amplified through ADHD, this constraint to fairness arises in many other machine learning applications in the public health field. Solving this issue would hold great value to the Machine Learning community, as it is also linked to efforts of proper proxy selection and retrieval; identifying the differences promoting actual bias versus disparity driven bias may lead to more specific feature use, reduced variance in model classification, and higher accuracy. To do so, however, requires greater understanding of the SDOH underlying the experiences of protected classes and the methods by which they may manifest in available clinical data.

6.2 How current should the data be?

A problem relatively unique to ADHD—among other psychiatric disorders or illnesses categorized under the DSM—is the frequency by which diagnostic criteria is updated. ADHD in particular has seen a remarkable change in the past two decades as marked by its change in terminology, symptomatology, and affected populations. As such, it must be acknowledged that

after such a change—particularly one as drastic as was presented between DSM-IV and V—there will likely be remarkable differences in diagnosis rates and populations affected. While observations of the physicians in the field have revealed that part of the proposed DSM-V changes were influenced by a lack of adherence to the original criteria (Epstein & Loren, 2013), the symptom criteria still acts as a gatekeeper for proceeding steps of diagnosis by way of parent and teacher evaluations, scoring systems, and clinical research. As such, it must be acknowledged that after such a change—particularly one as drastic as was presented between DSM-IV and V—there will likely be notable differences in diagnosis rates and populations affected.

The populations able to be diagnosed with ADHD, in particular, has expanded greatly since both DSM-III and IV through the reduction of exclusionary conditions and required symptoms, the shifting of age limits, and the categorization into two forms of presentations. As such, while it is unlikely that older data from DSM-III and DSM-IV would include individuals who would be wrongfully diagnosed under DSM-V, it is probable that many who *would* have been assigned the condition were not properly assessed. This is not only irretrievable data (It is particularly expensive, if not nearly impossible, to distinguish between an individual without ADHD and an individual wrongfully diagnosed to not have ADHD barring the assistance of the diagnosing physician of the time), but may lead to an algorithm learning a much stricter classifier than is reflected in DSM-IV.

Researchers seeking to develop an algorithm to accurately diagnose ADHD under the most up to date criteria must first consider under which diagnostic criteria the outcomes of the collected data were assigned, and how this may influence the purpose and intended outcome of

their model. This consideration is made somewhat simpler through the use of clinical data, where age and date of diagnosis may be encoded, but is not a common question asked in data collection conducted by online and telephonic surveys—such as the type conducted by Duda et al., (2017).

6.3 Model utilization under a diagnostic transition.

What is the expected use of such an algorithm during the transition from one set of criteria to another? Are there ethical implications in continuing use of the model on data reflecting the outcomes of one set of criteria until significant data has been established for the new criteria? Additionally, would the use of such models as an ubiquitous screening tool for ADHD diagnosis—particularly at a point in time where their accuracy has decreased due to criteria changes not yet reflected in the model—impact the overall change of incidence, prevalence, and symptomatology produced by the new criteria? Although these questions are proposed as somewhat vague and distant hypothetical, these perspectives are incredibly important to consider in the development of learning algorithms for mental and public health uses—especially in considering the effective use of the model in hospitals and physician practices.

6.4 Accounting for the intersection of subgroups.

A significant limitation in the analysis was my treatment of the identified subgroups as discrete categories rather than overlapping identities and characteristics an individual may hold at once. For example, diagnosis rates across gender reveal varied trends in age of diagnosis when comparing female adolescents to their male counterparts (Danielsen et al., 2018), but potential sources of bias were not examined in the context of this relationship during analysis. While this highlights the need for further examination of the impact of intersecting identity within the two

cases, it is interesting to note that emerging literature in machine learning has already begun to identify and assess this very problem. The literature identifies both the value that an intersectional perpsective provides in discrimination analysis (Buolamwini & Gebru, 2018) and proposes methods to mitigate algorithmic cost of trying to solve such a problem in model development (Cabrera et al., 2019; Kearns et al., 2019). Through this, the literature highlights a core tension that exists at the heart of both machine learning and public health: how do we weigh the unique identity of each data point against the desire to establish patterns and distinct classifications? How do we find the balance between the benefit of generalizability across a larger population and the desire to maintain accuracy upon the individual? This becomes a particularly necessary question to consider in the context of ADHD diagnosis, which continues to be a highly subjective and individualized process even in the face of efforts towards standardization

CHAPTER 7: CONCLUSION & RECOMMENDATIONS

The purpose of this thesis has been to examine the integration of machine learning algorithms into key steps of the ADHD diagnosis process, particularly through an understanding of the social determinants underpinning this topic. In doing so, I have examined not only the public health context of ADHD and the complexities of diagnosis, but also how this context creates a need for a thorough, more integrated approach to bias mitigation. To conclude my analysis, I propose a series of recommendations for those seeking to further integrate machine learning into the public health field and vice versa. While my recommendations speak specifically to the needs present in ADHD diagnosis, they can be extrapolated to other use cases of machine learning in health.

7.1 For Machine Learning Sector:

Seek out sources of data for populations underrepresented in ADHD diagnosis.

While the utilizations of fairness equations may help identify and mitigate forms of bias that as a research community we know to be harmful, there is much information regarding disparities in ADHD (such as: at which point does the disparity reflect biological differences in gender? What biases exist that have not yet been researched?) that has yet to be studied and quantified. As such, a fixed effort to include data points from marginalized identities acts as a protective factor of sorts—ideally this would mitigate the impact of missing data on unfairness and allow for further examination of other sources of bias such as skew and feature selection. The onus should be placed on the developers to preemptively include marginalized groups as the accuracy, impact, and overall use of their model may be severely reduced without such measures. Some methods of accessing these populations include reaching out to prominent racial

religious figures, teachers in low income classrooms, hospitals and clinics that serve uninsured populations, girl scouts, etc.

Treat bias identification as a proactive rather than reactive measure.

Algorithms for public use or social good results in the impact to a human life each time it is tasked to make a decision (this is particularly true for health based learning algorithms such as the ADHD diagnostic tools discussed earlier). A reactive approach to fairness, where models are not routinely evaluated for harmful bias until that bias has *already* been enacted, necessarily harms individuals' lives and well-being, even when caught early on in model use. Therefore it is crucial that the evaluation of fairness be integrated into the development process itself, proactively searching for instances where the chosen fairness metric "fails" upon an identified subpopulation. This expectation must be enforced through community based accountability, where journals, symposiums, and leaders in the machine learning field mark fairness evaluation as an indicator of, and eventually a requirement for, a well-researched presentation or paper.

7.2 For Public Health Sector:

Reconsider the use case of the developed machine learning tool as a method of mitigating bias.

In some regards, biases that a machine learning algorithm may accidentally encode can act as a valuable resource, as their existence points to the influence of SDOH on the intended outcome in a way that has yet to be identified. In a field like public health the identification of such a bias is not necessarily viewed negatively—and in fact is typically the first step in implementing an equity based framework to reduce such a disparity. This highlights the role that the context of how a bias is utilized can play into subsequent outcomes and even demographic trends for future uses of the model. Chouldechava introduces this idea in the examination of the

recidivism tool described in chapter 2 (2016). There, she notes that the negative impact of the identified bias was magnified by the fact that the model's intended use was to assign parole sentences to former inmates (2016). In contrast, if the model were to be used to introduce unincarcerated folk to recidivism reduction programs, this bias may now be seen as a benefitting factor. In this way, it becomes clear that context defines much of the way in which biases are discussed, and at times changing the context can mitigate the harm of a problem, particularly in the use of black-box algorithms where adjusting the algorithm itself is not within a public health official's reach.

If we theorize this same mindset to a tool that is designed to increase the sensitivity of the ADHD diagnostic process, we can imagine that such a tool may produce a disparate impact in who is able to proceed to further levels of care when it is used as a screening tool *after* a patient has expressed desire for referral or consultation—this has the potential of introducing biases that occur before step one in the diagnosis process later on in the clinical timeline, further filtering out already underrepresented groups. However, if that same model is used as a general screener given to all patients (for example, part of a pediatrician's initial assessment or yearly exam), it has a greater possibility of catching cases where individuals lack insight into the condition, actually increasing the number of cases that are caught.

Advocate for the inclusion of public health voices in the model development process.

It can be both a daunting and ineffective task to utilize an algorithm where aspects of its classification process are obscured—whether through the process of model development existing in a silo or the language surrounding an algorithm not providing enough transparency into its components. For algorithms specifically designed to be utilized in a healthcare setting, the

inclusion of public health officials, physicians, and social workers introduces key voices who are able to approach the problem through an understanding of SDOH and prompt a proactive anticipation of existing biases, an understanding of the scope of the data, and a more direct application to the intended use case.

As the fields of both machine learning and fairness continue to evolve, their usability in the healthcare context magnifies in potential., However, fully realizing the role algorithms may play in health—particularly mental health diagnosis—requires a further look at disparate impacts through the lens of SDOH and equity frameworks.

WORKS CITED

ADHD Editorial Board. (2019, November 1). *Diagnosing ADHD: How to Evaluate a Child for ADHD ADD*. https://www.additudemag.com/diagnosing-adhd/

American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition. Arlington, VA: American Psychiatric Association, 2013.

Beutel, A., Chen, J., Zhao, Z., & Chi, E. H. (2017). Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. *ArXiv:1707.00075 [Cs]*. http://arxiv.org/abs/1707.00075

Buolamwini, J., & Gebru, T. (n.d.). *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. 15.

Cabrera, A. A., Kahng, M., Fred, H., Jamie, M., & Duen Horng, C. (2019). Discovery of Intersectional Bias in Machine Learning Using Automatic Subgroup Generation. *Debugging Machine Learning Models Workshop (Debug ML) at ICLR*.

Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *ArXiv:1610.07524 [Cs, Stat]*. http://arxiv.org/abs/1610.07524

Chouldechova, A., & Roth, A. (2018). The Frontiers of Fairness in Machine Learning. *ArXiv:1810.08810 [Cs, Stat]*. http://arxiv.org/abs/1810.08810

Coker, T. R., Elliott, M. N., Toomey, S. L., Schwebel, D. C., Cuccaro, P., Emery, S. T., Davies, S. L., Visser, S. N., & Schuster, M. A. (2016). Racial and Ethnic Disparities in

ADHD Diagnosis and Treatment. *Pediatrics*, *138*(3). https://doi.org/10.1542/peds.2016-0407

Danielson, M. L., Bitsko, R. H., Ghandour, R. M., Holbrook, J. R., Kogan, M. D., & Blumberg, S. J. (2018). Prevalence of Parent-Reported ADHD Diagnosis and Associated Treatment Among U.S. Children and Adolescents, 2016. *Journal of Clinical Child & Adolescent Psychology*, *47*(2), 199–212. https://doi.org/10.1080/15374416.2017.1417860

Duda, M., Haber, N., Daniels, J., & Wall, D. P. (2017). Crowdsourced validation of a machine-learning classification system for autism and ADHD. *Translational Psychiatry*, *7*(5), e1133–e1133. https://doi.org/10.1038/tp.2017.86

Duda, M., Ma, R., Haber, N., & Wall, D. P. (2016). Use of machine learning for behavioral distinction of autism and ADHD. *Translational Psychiatry*, *6*(2), e732. https://doi.org/10.1038/tp.2015.221

Durkin, M. S., Maenner, M. J., Baio, J., Christensen, D., Daniels, J., Fitzgerald, R., Imm, P., Lee, L.-C., Schieve, L. A., Van Naarden Braun, K., Wingate, M. S., & Yeargin-Allsopp, M. (2017). Autism Spectrum Disorder Among US Children (2002–2010): Socioeconomic, Racial, and Ethnic Disparities. *American Journal of Public Health*, *107*(11), 1818–1826. https://doi.org/10.2105/AJPH.2017.304032

Epstein, J. N., & Loren, R. E. A. (2013). Changes in the Definition of ADHD in DSM-5: Subtle but Important. *Neuropsychiatry*, *3*(5), 455–458. https://doi.org/10.2217/npy.13.59

Evans, A. (2019). *Perceptions of ADHD Among African American Parents and Caregivers of Boys 5-14 Years Old.*[Columbia University]. https://academiccommons.columbia.edu/doi/10.7916/d8-63ky-3e87

Faraone, S. V., & Larsson, H. (2019). Genetics of attention deficit hyperactivity disorder. *Molecular Psychiatry*, *24*(4), 562–575. https://doi.org/10.1038/s41380-018-0070-0

Fairman, K. A., Peckham, A. M., & Sclar, D. A. (2020). Diagnosis and Treatment of ADHD in the United States: Update by Gender and Race. *Journal of Attention Disorders*, *24*(1), 10–19. https://doi.org/10.1177/1087054716688534

Ginsberg, Y., Quintero, J., Anand, E., Casillas, M., & Upadhyaya, H. P. (2014). Underdiagnosis of Attention-Deficit/Hyperactivity Disorder in Adult Patients: A Review of the Literature. *The Primary Care Companion for CNS Disorders*, *16*(3). https://doi.org/10.4088/PCC.13r01600

Gramlich, J. (2019, April 30). The gap between the number of blacks and whites in prison is shrinking. *Pew Research Center*. https://www.pewresearch.org/fact-tank/2019/04/30/shrinking-gap-between-number-of-blacks-and-whites-in-prison/

Gualtieri, C. T., & Johnson, L. G. (2005). ADHD: Is Objective Diagnosis Possible? *Psychiatry (Edgmont)*, *2*(11), 44–53.

He, L. (2019, October 2). *Busting Bias in Machine Learning for Cancer Research*.

Kaiser Family Foundation. Aug 28, P., & 2018. (2018, August 29). Uninsured Rate Among the Nonelderly Population, 1972-2018. *The Henry J. Kaiser Family Foundation*.

https://www.kff.org/uninsured/slide/uninsured-rate-among-the-nonelderly-population-1972-2018/

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2019). An Empirical Study of Rich Subgroup Fairness for Machine Learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, 100–109. https://doi.org/10.1145/3287560.3287592

Leitner, Y. (2014). The Co-Occurrence of Autism and Attention Deficit Hyperactivity Disorder in Children: " What Do We Know? *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00268

Littman, E. (2012, December). The Secret Life of Girls with ADHD. *Attention Magazine*, 3.

Marshall, A. T., Betts, S., Kan, E. C., McConnell, R., Lanphear, B. P., & Sowell, E. R. (2020). Association of lead-exposure risk and family income with childhood brain outcomes. *Nature Medicine*, *26*(1), 91–97. https://doi.org/10.1038/s41591-019-0713-y

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. *ArXiv:1908.09635 [Cs]*. http://arxiv.org/abs/1908.09635

Miller, T. W., Nigg, J. T., & Miller, R. L. (2009). Attention deficit hyperactivity disorder in African American children: What can be concluded from the past ten years? *Clinical Psychology Review*, *29*(1), 77–86. https://doi.org/10.1016/j.cpr.2008.10.001

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. *Proceedings of the*

Conference on Fairness, Accountability, and Transparency - FAT '19*, 220–229. https://doi.org/10.1145/3287560.3287596

Morgan, P. L., Staff, J., Hillemeier, M. M., Farkas, G., & Maczuga, S. (2013). Racial and Ethnic Disparities in ADHD Diagnosis From Kindergarten to Eighth Grade. *PEDIATRICS*, *132*(1), 85–93. https://doi.org/10.1542/peds.2012-2390

Mowlem, F. D., Rosenqvist, M. A., Martin, J., Lichtenstein, P., Asherson, P., & Larsson, H. (2019). Sex differences in predicting ADHD clinical diagnosis and pharmacological treatment. *European Child & Adolescent Psychiatry*, *28*(4), 481–489. https://doi.org/10.1007/s00787-018-1211-3

Pastor, P. N., & Hawkins, L. D. (2015). *Association Between Diagnosed ADHD and Selected Characteristics Among Children Aged 4–17 Years: United States, 2011–2013*. *201*, 8.

Ramtekkar, U. P., Reiersen, A. M., Todorov, A. A., & Todd, R. D. (2010). Sex and age differences in Attention-Deficit/Hyperactivity Disorder symptoms and diagnoses: Implications for DSM-V and ICD-11. *Journal of the American Academy of Child and Adolescent Psychiatry*, *49*(3), 217-28.e1-3.

SFARI | The Simons Simplex Collection. (2011, September 16). SFARI. https://www.sfari.org/funded-project/the-simons-simplex-collection/

Solden, Sari. (2019, August 28). *ADHD in Women: Symptom Checklist and Self-Test for Adults*. ADDitude. https://www.additudemag.com/adhd-symptoms-in-women/

Ustun, B., Adler, L. A., Rudin, C., Faraone, S. V., Spencer, T. J., Berglund, P., Gruber, M. J., & Kessler, R. C. (2017). The World Health Organization Adult

Attention-Deficit/Hyperactivity Disorder Self-Report Screening Scale for DSM-5. *JAMA Psychiatry*, *74*(5), 520–526. https://doi.org/10.1001/jamapsychiatry.2017.0298

Ustun, B., & Rudin, C. (2019). Learning Optimized Risk Scores. *ArXiv:1610.00168 [Math, Stat]*. http://arxiv.org/abs/1610.00168

World Health Organization. (2010). *A conceptual framework for action on the social determinants of health: Debates, policy & practice, case studies*.http://apps.who.int/iris/bitstream/10665/44489/1/9789241500852_eng.pdf

Xu, G., Strathearn, L., Liu, B., Yang, B., & Bao, W. (2018). Twenty-Year Trends in Diagnosed Attention-Deficit/Hyperactivity Disorder Among US Children and Adolescents, 1997-2016. *JAMA Network Open*, *1*(4), e181471–e181471. https://doi.org/10.1001/jamanetworkopen.2018.1471