Diagnostic errors in child mental health: Assessing treatment selection and its long-term consequences

Jill Furzer*

Most Recent Version

November 18, 2020

Abstract

The subjective nature of mental health diagnoses may cause missed diagnoses in some children and over-treatment in others, with differences by gender, race and socioeconomic status. To test for this, I estimate mental health risk in childhood using nationally representative health data. I calculate over-diagnosis by estimating risk-specific jumps in diagnosis by a child's school starting age. I find diagnosed young-for-grade children have a lower risk of mental illness, while many high-risk kids remain undiagnosed. Underdiagnosis appears to be driven by internalized mental health symptoms and pro-social behaviours leading to symptom masking, which is especially common in females. I investigate the impact of over- and under-treatment on longer-run outcomes—including educational attainment and the likelihood of social assistance receipt—using linked tax records. My results suggest that adverse treatment effects on education, social assistance receipt and income are concentrated in low-risk, over-treated individuals. However, mental health treatments improve these outcomes for high-risk children and accrue sizable improvements in the form of reduced rates of workplace injuries and suicide ideation. As missed or delayed diagnoses represent a larger misallocation of scarce support resources, these results suggest that targeted screening before age 11 would reduce inequities in diagnosing mental illness and downstream socioeconomic inequalities.

Keywords: mental health, child health, treatment selection, machine learning

JEL Codes: I10, I14, J01, J24

*Institute of Health Policy Management and Evaluation, University of Toronto, Canadian Centre for Health Economics email: jill.furzer@mail.utoronto.ca.

I sincerely thank Audrey Laporte, Elizabeth Dhuey, Shari Eli, Claire de Oliveira, Michael Baker, Mark Stabile, Lauren Jones, John McConnell, Daniel Millimet, Ludovica Gazzè, Boriana Miloucheva, Emmanuelle Arpin and participants at ASHEcon, APPAM, and the Monash University CHE seminar for helpful comments to improve this paper. The author acknowledges funding from the Canadian Research Data Centre Network (CRDCN) Emerging Scholars Grant. The analysis for this paper was conducted at the Toronto Regional Statistics Canada Research Data Centre, a part of the CRDCN. The services and activities provided by the CRDCN are made possible by the financial or in-kind support of the SSHRC, the CIHR, the CFI, Statistics Canada, and participating universities whose support is gratefully acknowledged. The views expressed in this paper do not necessarily represent the CRDCN's or that of its partners.

1 Introduction

In childhood, poor mental health is a large and growing issue, with one in five children diagnosed with a mental health disorder before age 18 (Ogundele, 2018). The prevalence of childhood anxiety and depression has notably risen in recent years, up to five percent in Canada (Statistics Canada, 2020a) and eight percent in the US (Bitsko et al., 2018). Suicide is also now the leading cause of death in those 10 to 14 years old (Statistics Canada, 2020b). Beyond direct health effects, disorders like anxiety, depression, and attention deficit hyperactivity disorder (ADHD) also lead to reduced educational attainment, likelihood of employment, and earnings in later life (for review see Prinz, Chernew, Cutler, & Frakt, 2018). Many interventions are available to mitigate these adverse trajectories, but their estimated treatment effects remain mixed. While some studies find improvements in risky behaviours and police contact for pharmacologically treated children (Chorniy & Kitashima, 2016; Dalsgaard, Nielsen, & Simonsen, 2014), and lowered educational outcomes with reduced intervention (Busch, Golberstein, & Meara, 2014), other studies find poorer educational outcomes (Currie, Stabile, & Jones, 2014; Kitashima & Chorniy, 2017). Understanding when and why treatment is effective is integral for informing health policy and addressing health-based inequalities in general.

Nevertheless, two main problems plague the estimation of mental health treatment effects. First, mental health is mostly unobserved, and secondly, treatment is sought non-randomly. In this paper, I argue that, while mental health interventions may prove beneficial for some patients, previously estimated treatment effects might include children who are in fact "low-risk", and who, as a consequence, experience minimal treatment benefits. At the same time, not everyone needing treatment may obtain appropriate medical services, resulting in forgone treatment benefits. This misallocation of mental health services aligns with the growing acknowledgment of over- and under-treatment in medicine, and the direct medical costs associated with either case (Kale & Korenstein, 2018; Chandra & Staiger, 2007; Abaluck, Agha, Kabrhel, Raja, & Venkatesh, 2016; Mullainathan & Obermeyer, 2019). Given the intersecting nature of mental health with human capital formation, mental health diagnostic decisions also have broader short- and long-term educational and socioeconomic consequences.

In this paper, I first examine under- and over-diagnosis of mental illness in childhood. Secondly, I estimate mental health treatment heterogeneity for children's health and socioeconomic outcomes while accounting for latent risk and treatment selection. In particular, I use a gradient boosted tree model to capture latent mental health severity based on mental health risk factors and current behaviour. With these estimates, I assess under- and over-treatment based on discrepancies in predicted mental health risk and realized diagnosis. To understand causal under or over-treatment, I estimate risk-specific diagnosis induced by a child's school starting age. Finally, I calculate short- and long-run mental health treatment effects by predicted mental

health risks using a rich nationally representative child survey linked to later life tax records. I exploit variation in individual risk and treatment status over time and within closely defined neighbourhoods to isolate treatment costs and heterogeneous benefits.

These combined strategies improve upon previous mental illness treatment estimates in several ways. Firstly, they capture a robust measure of mental illness risk that accounts for a significant source of omitted variable bias. Harnessing machine learning methods to estimate mental illness risk advances the use of self- or parent-reported symptoms to measure mental health severity. On their own, self- or parent-reported symptom measures may not fully capture the multi-factorial components of current symptom manifestation and being at-risk for mental illness. Use of the National Longitudinal Survey of Child and Youth (NLSCY) allows for the inclusion of mental health risk factors, including family environment, maternal psychopathy and hereditary endowments, in addition to current symptoms. Secondly, this paper uses discrepancies in predicted mental health risk and treatment choice to measure the precise effects of both risk and treatment by risk. Previous assessments have included underlying mental illness severity using self-reported symptoms instead of diagnostic status to curtail the treatment selection issue (i.e., Currie & Stabile, 2006; Cornaglia, Crivellaro, & McNally, 2015). Using variation in predicted individual risk and treatment status across time accounts for the impact of latent mental illness, the potential costs of treatment, and the benefits of treatment by risk profile. Thirdly, my use of tax records to measure long-term socioeconomic trajectories following mental health treatment provides a more fulsome picture of key later life outcomes over survey-based measures. Survey-based measures can suffer from misreporting and selection bias due to attrition. Finally, linkage of the NLSCY to tax records allows for the estimation of effects over a longer time horizon – upwards of 20 years – than has previously been assessed in the extant literature.

From my results, I first identify substantial treatment selection, including both over and under treatment. I assess marginal diagnoses across the predicted mental health risk distribution using variation in diagnostic likelihood by a child's school starting age. Starting school younger than average has been repeatedly shown to increase the identification of general disability and other mental health disorders (for review, see Dhuey, Figlio, Karbownik, & Roth, 2019). I find discontinuous jumps in diagnosis based on being young-for-grade come from across the risk distribution, especially the low-risk end of the distribution. These results suggest that being young-for-grade produces a faulty signal of risk, resulting in low-risk students receiving diagnoses before other higher-risk peers. Combining a data-driven risk prediction with a school starting age identification strategy provides a novel methodological application of machine learning causal inference. It sheds light on a common question around whether school starting age generates under- or over-diagnosis (Elder, 2010; Evans, Morrill, & Parente, 2010). Over-treatment of low-risk individuals is evident; however, a large proportion of high-risk kids are never diagnosed, and even more so for the oldest kids.

In estimating the costs and benefits of diagnosis and treatment, I identify negative effects across all human capital-based outcomes for low-risk children who received treatment. This result again suggests over-treatment is occurring and that it generates tangible harm. For those with the lowest mental health risk, receiving medical treatment in adolescence increases rates of dropout, expulsion and retention. Treatment also results in 0.72 fewer years post-secondary education and increases the likelihood of social assistance receipt in later life by 12 percentage points. In contrast, the probability of social assistance receipt decreases 19 percentage points for the highest risk adolescents under treatment. These same children obtain 0.81 more years of post-secondary education.

Compared to human capital outcomes, I find lower direct costs to treatment and more substantial treatment benefits. Treatment reduces injuries in childhood and adulthood, estimated via reduced workers compensation board (WCB) insurance claims. With $8 billion in annual WCB claims paid per year in Canada, I estimate a potential downstream yearly savings of $610 million if all adolescents in the top risk decile received medical intervention. Further findings showing reductions in suicidal ideation and adolescent drug use for those in the top risk deciles demonstrates mediating pathways leading to improved long-term socioeconomic outcomes with appropriate treatment. These results appear to deviate from previous results finding a negative gradient between mental health treatment and education, income, and employment. However, as opposed to blanket negative or null effects, I find adverse treatment effects concentrated in low-risk individuals. Conditional on being high-risk, treatment benefits are maximized when diagnosis occurs before the age of 11. Early diagnoses for low-risk children cut off human capital investment streams, while interventions that occur too late do not have enough time to accrue benefits.

Finally, I show evidence that sex differences in treatment outcomes found both in this paper and in previous studies (i.e., Busch et al., 2014; Fletcher, 2010) may derive from the under-estimation of mental illness risk for girls. Mirroring potential human biases, the diagnostic prediction model more heavily weights externalized symptom clusters (e.g., hyperactivity) that are more common in males. It also under-weights children with high levels of pro-social behaviours (e.g., cooperativeness), which are more common in females and can lead to symptom masking. Children with more internalized symptoms and pro-social behaviours then have a lower likelihood of diagnosis. Still, symptom differences explain only about half of the diagnostic gap between males and females. Due to this underweighting based on symptoms or unexplained bias, females who receive a diagnosis are then likely to be more at-risk and show comparably worse long-term outcomes.

These findings shed light on inequities in mental health care access that manifest in childhood. I show that mental health risk predictions help identify high-risk children and adolescents in terms of future mental health-related adverse events. I also show that mental health treatments prove highly beneficial for these high-risk individuals because they reduce adverse events and improve human capital accumulated over the

long-term. However, I identify that a risk misperception occurs in the diagnostic process, leading to sub-optimal rates of diagnoses, with lower-risk children and adolescents diagnosed before or instead of higher risk ones. Results showing school starting age acts as a faulty signal of risk highlight the role of school-based treatment referrals in this selection issue. These results have direct policy implications for targeted screening to ease barriers to diagnosis and treatment for mental illness and reduce socioeconomic inequities over the long-term.

2 Background

2.1 Child Mental Health Outcomes and Treatment

There is a relatively large literature showing that having mental health disorder in childhood, such as anxiety, depression or ADHD, reduces overall well-being and has lasting socioeconomic consequences. For example, mental health issues have been found to increase delinquency, the likelihood of grade repetition, and reduce math and reading scores (Currie & Stabile, 2006, 2009). Cornaglia et al. (2015) find a strong association between general mental illness in adolescence and the rate of dropout, especially for females. In the long term, mental illness is associated with reduced employment and earnings (Fletcher, 2013, 2014; Johar & Truong, 2014; Mousteri et al., 2019). Comparing across disorders, in one study, ADHD had the largest and most consistent effect on long-term outcomes like welfare receipt (Currie & Stabile, 2009).

For the treatment of mental health issues in children, clinical evidence supports the short-term use of antidepressants, selective serotonin reuptake inhibitors (SSRIs) (Kodish, Rockhill, & Varley, 2011; Walkup, 2017) and stimulants (Chan, Fogler, & Hammerness, 2016; Maia et al., 2017). However, clinical knowledge of the long-term effects of medication use in childhood is minimal, and where there is data it suggest that there are only small long-term benefits, and high rates of remission (Daley & Birchwood, 2010; Molina et al., 2009; Ginsburg et al., 2018). While psychotherapy shows positive effects, it is often not the first choice for patients, at least for adult patients (Cronin, Forsstrom, & Papageorge, 2020). At the population level, evidence of mental health treatments' effects remains mixed. Using a Medicaid sample of ADHD diagnosed children and adolescents, Chorniy and Kitashima (2016) find that stimulant use reduced sexually transmitted infections, pregnancies and drug use. Stimulants have also been shown to decrease hospital and police contacts in Denmark (Dalsgaard et al., 2014). Nevertheless, these stimulant medications have a limited impact on educational outcomes (Kitashima & Chorniy, 2017) and are associated with parental relationship deterioration (Currie et al., 2014). These patterns reverse in one study of antidepressant use in adolescents, where schooling outcomes declined when FDA warnings lead to a drop in usage (Busch et al., 2014). Results

were, again, especially strong for females.

The root cause of these mixed treatment effects may derive from a still limited understanding of the causes of mental illness, which can be based on genetic, biological, environmental and unknown factors (Ogundele, 2018). Some known factors include mother psychopathy, difficult family functioning, adverse events *in utero* or childhood, exposure to smoking, sex of the child, and family income (Cuffe et al., 2005; Froehlich et al., 2007; R. G. Frank & Meara, 2009). While many medical disorders have been recently linked to genetic risk factors, most mental health disorders have no identified genetic markers for vulnerability, nor is it clear the extent to which environment plays a role in activating these genes (Insel, 2014; Faraone & Larsson, 2019).

2.2 A Model of Treatment Selection

A largely understudied explanation of these mixed treatment effects may be that endogenous treatment selection generates these mixed results. Access to care, physician, family and country-specific preferences, and incorrect weighting of visible patient risk factors all influence diagnoses and treatment decisions. These decisions include the choice to pursue a diagnostic assessment for a given child and the physician's interpretation of any diagnostic results.

To understand treatment selection for child mental health, I build a simple cost-benefit model. I define a mental health risk measure α_i that runs from low to high risk on a scale of zero to one. The benefits and costs of mental health diagnosis and treatment were first outlined by R. Frank and McGuire (1999). Benefits $B(\alpha_i)$ include direct medical intervention effects on improved mood, memory or motivation. Reductions in accidental injuries, hospitalizations, suicidal ideation, and changes in risk-seeking behaviour are additional direct benefits. These health benefits, in addition to non-medical interventions like compensatory educational investments, may then accrue indirect human capital benefits. These combined benefits $B(\alpha_i)$ are assumed to be increasing in risk. The expected cost of identification has a fixed monetary and time cost of diagnosis c_D, and what I assume to be a fixed welfare cost in the form of stigma and adverse medication effects or complications c_L (Batstra & Frances, 2012; Owens & Jackson, 2017). Together these are defined as overall cost C.[1] Additional welfare costs include the opportunity cost associated with streamed academic pathways for the individual child and the opportunity cost of treatment for one child at the expense of another within the context of constrained school and medical resources.

A patient (or their family) opts for diagnosis and treatment ($D_i = 1$) if the benefit of diagnosis is greater than or equal to the cost. With cost fixed and benefits increasing in risk, there exists some risk threshold $\bar{\alpha}$ whereby benefits are greater than or equal to costs C when mental health risk $\alpha_i \geq \bar{\alpha}$ (Figure A1). Above

[1] In Canada, opting for assessment through the universal health care system accrues only the time costs spent awaiting assessment on a referral waitlist, which can be considerable. Families may opt for private assessments which they pay out of pocket or which can be covered by private insurance.

this threshold, there is a positive effect of identification and medical treatment on health. Individuals who are diagnosed with a mental health illness but who are less severe ($\alpha_i < \bar{\alpha}$) will have a negative treatment effect. Thus, the marginal child has a risk equal to a threshold $\bar{\alpha}$ where the benefits of identification equal the cost, again from the perspective of the family.

Translating this to human capital outcomes, I define two counterfactual human capital outcomes by diagnostic status as: $H_i^1 = \mu_0(\alpha_i) + X_i\beta + B(\alpha_i) - C + \epsilon_i^1$ and $H_i^0 = \mu_0(\alpha_i) + X_i\beta + \epsilon_i^0$. In the vein of Grossman (2000), initial endowment is measured through $\mu_0(\alpha)$, written here as dependent on underlying mental health traits (or non-cognitive skills). Child and family characteristics are measured through X_i. Previous literature tells us that human capital is decreasing in mental illness risk, so μ_0 decreases with risk. Unobservable human capital and other unobserved factors leading to a diagnosis or not are defined through ϵ_i^1 and ϵ_i^0. The treatment effect is then the difference between the two states: $H_i^1 - H_i^0 = B(\alpha_i) - C + (\epsilon_i^1 - \epsilon_i^0)$, with an expected benefit of:

$$E[H_i^1 - H_i^0|X_i] = E[B(\alpha_i) - C|X_i] + E[\epsilon_1 - \epsilon_0|X_i]$$

It is assumed that diagnosis occurs when the expected diagnostic effect, and resultant effect of treatment, is positive (see Panel B, Figure A1. The likelihood of seeking a diagnosis on part of the parent or diagnostic assignment by the physician is dependent on the *ex ante* expected benefit of diagnosis and treatment. Assuming that the benefits of diagnosis and treatment increase with risk under a fixed cost, the probability of diagnosis and treatment will also increase with risk. However, while in general the severity of mental illness health presentation correlates with service use, there are still likely high-risk patients with unmet medical needs (Wang et al., 2007) and the potential for over-diagnosis for others. Why might this misallocation occur?

Firstly, latent mental illness risk is at least partially unobservable. Any misperception in the risk of mental illness leads to misperception of diagnosis and treatment benefits. For child mental health, specifically, parents and sometimes teachers act as agents on behalf of the child to seek a diagnosis, further complicating risk perception. The views of a teacher or parent as to a given child's mental health may differ due to the mutability of behaviours over time or due to differences in a child's behaviour in the home as compared to the school environments (Mitchell & Shepard, 1966; Strickland, Hopkins, & Keenan, 2012). There may also be school funding incentives or budget constraints to consider that may differ by public or private school district type (Dhuey & Lipscomb, 2011, 2013). For physicians, the diagnostic and statistical manual of mental disorders (DSM), now in its fifth iteration, contains the most common criterion used to identify childhood disorders. These diagnosing criteria provide guidelines for assessing symptom severity and determining a diagnosis for all currently recognized mental health disorders. However, these guidelines are subjective and

evolving, leading to a wide variation in identification rates of these disorders dependent on country, region, time, and physician-specific preferences (Batstra & Frances, 2012; Kale & Korenstein, 2018).

Physicians may also apply different standards for mental health diagnosis by race or gender (Bertakis et al., 2001; Snowden, 2003; Bruchmüller et al., 2012; Quinn & Madhoo, 2014; Mullainathan & Obermeyer, 2017). Girls and non-Hispanic white children are more likely to receive depression diagnoses, conditional on an underlying depression score (Fletcher, 2008). Recent work has found misdiagnosis for ADHD based on school starting age (Evans et al., 2010; Elder, 2010; Schwandt & Wuppermann, 2016; Layton et al., 2018). This misdiagnosis has sex and income correlates (Furzer, Dhuey, & Laporte, 2019). In respect of differences in treatment, when more affluent families seek care, they are more likely to have a physician with a higher propensity to prescribe medication (Dalsgaard et al., 2014).

The probability of diagnosis and treatment also decreases with the cost. There are also well-documented discrepancies in access to mental health diagnostics by race, driven by racial segregation in service availability and discrimination (Bertakis et al., 2001; Snowden, 2003; Williams & Mohammed, 2009). Conversely, affluence may impact the likelihood of seeking a diagnosis based on perceived cost or stigma (Link & Phelan, 2010). Recent work has found that, at least in adult populations, low expected treatment benefits combined with the perceived cost of stigma are likely the driving forces behind hesitation to seek treatment for mental illness treatment hesitation, as opposed to monetary or time costs (Cronin et al., 2020). Finally, the distribution of children's cognitive and non-cognitive abilities by diagnostic status and the correlation in these ability distributions may also impact parental willingness to seek care. If parents feel that diagnosed children, on average, appear to fare worse (better) because of the diagnosis itself, they may delay seeking a (advocate for) diagnosis.

Given the potential for selection in seeking and obtaining a diagnosis as well as medical treatment, current measured effects of identification and pharmacological treatment on long-term outcomes may be biased. Assessing outcomes by diagnostic status alone assumes that all diagnoses and treatments received were appropriate and that no child who should have received diagnosis was missed. This is despite the rising concern of both over-diagnosis and over-treatment in numerous disorders (Kale & Korenstein, 2018; Chandra & Staiger, 2007; Abaluck et al., 2016; Mullainathan & Obermeyer, 2019). The estimated overall average treatment effect on direct health outcomes or long-term socioeconomic variables may be small or negative because this assumption does not hold. The small or negative average treatment effects in socioeconomic outcomes reported in the literature to date may thus result from heterogeneous treatment effects across the risk distribution and a higher density of low-risk diagnoses (i.e. of children being diagnosed with a condition who actually do not have that condition). To test this hypothesis, I rely on a machine learning prediction of health-based diagnostic risk. Repeated diagnostic risk predictions across childhood, and the variation in

both these and diagnostic and treatment status, are then used to estimate precise treatment effects by risk while accounting for observable and unobservable selective mechanisms that could bias these effects.

3 Data

To estimate latent mental health risk and measure the separable effect of diagnosis and treatment on long-term outcomes I rely on a nationally representative sample of 66,667 children and adolescents in the National Longitudinal Survey of Child and Youth (NLSCY) and administrative tax files linked to 13,076 children from the original 1994 NLSCY cycle, which I describe in detail below.

3.1 National Longitudinal Survey of Child and Youth

The NLSCY surveyed children across Canada, starting in 1994 with a cohort of children aged 0 to 11. This cohort responded to biannual follow-up surveys every two years until 2008. Each survey cycle added an additional grouping of children and youth to create parallel longitudinal and cross-sectional samples. I use the NLSCY primarily to form a diagnostic prediction as it contains an extensive collection of health related variables observable to a diagnosing physician. Importantly, the NLSCY includes parent and child health histories for all children, not just those diagnosed with a mental health condition. Through this complete response set, the NLSCY avoids the problem of the selective labelling in administrative files, whereby data contains only information on the portion of the population who sought a diagnostic assessment (Lakkaraju et al., 2017; Rose, 2018). In addition to child and family health histories, the NLSCY also includes the Children's Behavioural Survey (CBS), a DSM-validated scale given to parents of children under the age of 12 to screen for ADHD, anxiety, depression, aggression and pro-social behaviour. Children 12 and older provided self-reported CBS assessments and completed additional questions corresponding to a depression scale and a property offences scale (e.g., child reports having vandalized property). These will play a particularly important role in assessing diagnostic risk. A full list of the 109 variables (689 factorized) used for predicting health-based diagnostic risk is available in Appendix A.

To indicate diagnosis, the NLSCY contains information on diagnosed ADHD, and a general indicator for emotional, psychological and nervous disorders (i.e., anxiety, depression, obsessive-compulsive disorder, conduct disorder). I construct a general indicator for any mental health disorder if a parent reports that their child was diagnosed with any of these medical conditions by a medical professional. I do not separate these disorders as mental health disorders often overlap in their symptoms and are often diagnosed concurrently. Current thinking on mental health also suggests that it is a continuum of traits over binary disorders (Caspi et al., 2014). I construct a treatment indicator based on reported use of broad categories of stimulants or

tranquillizers.[2] I set this indicator equal to one for every year after a diagnosis is first reported. Excluding any individuals without a diagnosis and treatment status, the pooled estimation sample of the NLSCY runs over eight cycles and 15 years and includes 209,880 total observations over 66,667 individual children and 57,131 households. There is an average of four cycles of information available per individual.

Descriptive analysis of key variables by diagnostic status for the full NLSCY data set are found in Table 1. Diagnosed children are less likely to be female, more likely to be non-Hispanic white, and come from families with lower income and higher maternal depression rates. Behavioural scores also differ significantly by diagnostic status.[3] As my constructed measure of mental health diagnosis is a composite diagnosis of ADHD and emotional, psychological or nervous disorder, Table 1 shows that within the diagnosed sample, 73 percent of kids have ADHD and 57 percent have anxiety and depression. Just over half of diagnosed children take stimulants, and 13 percent take tranquilizers.[4] Diagnosed children are also much more likely to have a learning disability, mental handicap, asthma and epilepsy, suggesting high levels of co-morbidity. For those not diagnosed at a given point in time, four percent will go on to receive a diagnosis in later years.

I also compile several short-term outcomes from the NLSCY to test direct and indirect effects of mental health treatment. Direct short-term outcomes include whether a child was injured or spent the night in a hospital in the past year. I capture health behaviours with indicators of adolescent-reported recreational use of marijuana and prescription pills. I additionally create an indicator regarding whether an adolescent reported ever having suicidal thoughts. To look at shorter-term indirect schooling outcomes, I create an indicator equal to one if a child drops out of school for more than one week, grade retention, and disciplinary action (i.e., suspensions or expulsions).

I divide the 209,880 observations of the NLSCY data set into a training and a test set for the gradient boosted tree model, which will be described in the next section. The training set is used to build and calibrate the model algorithm and the test set for results analysis. The test set combines all observations of each tax-linked individual child with a 10 percent random selection of non-linked children and adolescents. As the linked sample contains 13,076 children (a near 100 percent match for all children in both the first and second NLSCY cycles) with up to eight observations each, combining all observations of these children totals to 20 percent of the total sample of observations. Adding a 10 percent sample of all remaining individuals and all observations from those individuals generates a test set of 104,527 observations over 18,436 individual children. The remaining observations comprise the training set, with 105,356 observations over 48,231 children

[2]The NLSCY also reports if an individual takes anti-convulsants, which are sometimes used as a mood stabilizer to treat bi-polar disorder, schizophrenia, or borderline personality disorder. I do not include this in the main treatment variable given that this is considered off-label use and I cannot isolate what disorder they are being used to treat.

[3]One demographic to note is that the NLSCY is largely white. Using self-reported race and ethnicity variables, 95 percent of the sample identified as white. Diagnosed children are even more so, at 99 percent.

[4]Note that those with an ADHD diagnosis have a much higher treatment rate (82 percent) than those with emotional, nervous, or other psychological disorders (40 percent).

to approximate a 50 percent split of the data into test and training samples, consistent with common data splitting conventions in the machine learning literature.

Table 1: Descriptive Statistics of Key Covariates by Diagnosis Status

	Overall		Diagnosed		No Diagnosis		Diff
Variable	Mean	S.D.	Mean	S.D.	Mean	S.D.	
			Full NLSCY Sample				
Female	0.49	0.50	0.38	0.49	0.50	0.50	-0.12***
Non-Hispanic White	0.95	0.21	0.99	0.10	0.95	0.21	0.04***
Log Family Income	10.85	0.70	10.83	0.72	10.85	0.70	-0.03***
Family Size	4.11	1.22	3.83	1.28	4.12	1.21	-0.30***
Maternal Depression	4.22	5.16	5.95	6.71	4.14	5.07	1.81***
ADHD Scale	5.28	3.05	8.51	3.76	5.13	2.93	3.39***
Anxiety/Depression Scale	3.29	2.64	5.75	3.30	3.17	2.55	2.58***
Conduct Scale	3.71	3.03	4.30	3.31	3.68	3.01	0.61***
Ind. Aggression Scale	1.34	1.73	2.23	2.13	1.30	1.70	0.94***
Learning Disability	0.05	0.22	0.38	0.48	0.03	0.17	0.35***
Mental Handicap	0.01	0.07	0.06	0.24	0.00	0.05	0.06***
Asthma	0.18	0.38	0.37	0.48	0.17	0.37	0.20***
N Physician Visits	2.48	3.75	3.36	5.35	2.44	3.66	0.92***
Current Diagnosis	0.04	0.20	1.00	0.00	0.00	0.00	-
Future Diagnosis	0.08	0.27	1.00	0.00	0.04	0.20	-
% ADHD	0.06	0.24	0.73	0.45	-	-	-
% Anxiety/Depression	0.04	0.19	0.57	0.50	-	-	-
% On Stimulants	0.02	0.15	0.51	0.50	-	-	-
% On Tranquilizers	0.01	0.07	0.13	0.34	-	-	-
N	209,880		8,964		200,916		

3.2 T1FF Tax Linkage

Tax record linkages include 13,076 linked children surveyed in the first cycle and remaining in the NLSCY until cycle four, with tax records available from 2000 to 2015. The oldest children in the sample turned 18 in 2000. The tax records contain linkages for kids from ages 18 to 33 on labour force participation, unemployment (individual received unemployment insurance), and personal income. I create an indicator for ever claiming unemployment or social assistance based on all tax records from 2000 to 2015. The T1FF also provides information on worker's compensation board (WCB) injury claims and education tax credit claims. I proxy for higher education by calculating the number of years that education tax credits were claimed and similarly estimate the number of years in which a person received worker's compensation insurance payments. I normalize workers compensation claims to a maximum of 10 years, the average number of years of tax data available for each individual. Finally, I record income and labour force participation as of 2015. I use a measure of the log-average overall household income less personal income over the tax sample period

to account for whether a young adult is still living at home with their parents or with an income-earning partner.

These linked tax records provide a longer-term income and employment horizon than assessed in the previous literature while also limiting attrition and misreporting concerns common in self-reported income. Misreporting due to tax evasion or errors in filing or process may still result in measurement error (Kapteyn & Ypma, 2007). However, estimates using tax files are assumed to measure income more precisely than self-reported measures. Furthermore, use of long-term WCB claims allows for longer-term estimation of the known link between certain mental health disorders and the likelihood of injury well into adulthood (Nigg, 2013).

4 Empirical Strategy

4.1 Diagnostic Risk Prediction

Estimating heterogeneous treatment effects for mental illness requires a consistent estimation of latent mental health risk (α_{it}). Previous studies have gauged mental illness severity through use of behavioural symptom scales, using the DSM or general health surveys. These measures may not account for the multi-factorial etiology of mental health disorders stemming from hereditary, environmental, biological, and exogenous factors. They may also not account for the fact that a child may be at-risk, but not fully express symptoms at a given point in time. To address these complications, I incorporate a gradient boosted tree algorithm to create a high-dimensional non-parametric prediction of the health-related factors leading to diagnosis. The main dependent variable in this prediction is reported diagnosis of any mental health disorder in the NLSCY, measured at bi-annual time points over the survey period (Y_{it}). As noted above, I employ a large covariate set relating to child health, behaviour, parent health, adverse events and early health endowments to predict health-related diagnosis.

A gradient boosted tree model is a stage-wise additive model that builds decision trees by partitioning the data based on covariate value criterion (d) to create groupings (I_d) of individual observations ($I_d = \{i|q(X_i) = d\}$). The algorithm partitions the covariates to maximize the information gained from each split. After one tree build, the within-group predicted diagnosis is simply the average group diagnosis rate: $Y_i = \bar{Y}_{(i \in I_d)}$). The algorithm then estimates where this prediction went wrong, and iteratively builds trees based on a random sampling of faulty (high residual) predictions in previously built trees (J. H. Friedman, 2001). It continues to build until no new information can be gained, measured through a minimized mean squared error. The final diagnostic prediction is a summation of predicted diagnosis across all constructed trees for an individual i.

I employ a gradient boosted tree model firstly because its iterative learning approach is particularly useful for measuring relatively rare events like annual mental illness incidence (Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2012). While the lifetime prevalence of mental illness is around 20 percent, the annual incidence of child mental health disorders is relatively rare. Traditional prediction methods tend to over-predict non-events under this classification imbalance.[5] The "thresholding" approach to gradient boosted trees is also thought to well-approximate innate human decision making by assigning probabilities based on value thresholds over coefficient weights (Obermeyer & Emanuel, 2016; Athey, 2018). Finally, traditional regression methods are computationally limited in their ability to incorporate significant quantities of covariates and interactions. Given how little is still known about mental health and the potential for mis-identification in reported diagnosis, there is a high risk of measurement error. By incorporating a large covariate set in a non-linear framework, a gradient boosted tree model can help to minimize measurement error due to omitted variable bias.

To minimize algorithmic estimation error, the model is tuned to optimize in- and out-of-sample prediction. Several parameters are calibrated, including tree depth and the learning rate, or the proportion of faulty predictions used from previously built trees to inform the next tree build. This parameter determines how fast or slow the algorithm learns from its mistakes. The number of trees built is then inversely related to the learning rate: a fast-learning machine requires fewer trees to learn. With each tree build, the algorithm uses a random selection of observations and variables to incorporate stochasticity into the model to improve prediction accuracy (J. H. Friedman, 2002). These parameters are selected to maximize the out of sample goodness-of-fit in the hold-out test sample. Parameters are first tested in the training sample via cross-validation, where the training set is cut into k-folds and the model is fit on $k-1$ folds of the training data set. The training sample fit is then tested on the k^{th} fold. This process is repeated k times for each sample and each parameter - I use a five-fold cross-validation. Once the model is tuned to predict health-based diagnostic risk, the final out-of-sample prediction is performed on the test set. This out-of-sample estimate provides a prediction of whom the algorithm predicted should have been identified based on their underlying health needs. Comparing this predicted health-based diagnostic risk to actual diagnosis offers a first basis for who was overlooked or diagnosed unnecessarily.

Given the assumption of misdiagnosis, using current diagnostic status as a proxy to predict latent mental illness risk raises some concern regarding the potential for misidentification. If a diagnosis is biased by factors like race or gender, then any prediction using diagnosis as a dependent variable may propagate these biases (Rambachan, Kleinberg, Ludwig, & Mullainathan, 2020). However, given the subjective and distributional

[5]Indeed, using a linear probability model or a logit model to predict diagnosis renders zero or close to zero diagnosis rates. See Appendix B for more details.

nature of mental health, I argue that there is no better proxy for overall mental health risk outside of actual diagnosis. Suicide has been used as a predictive outcome in the machine learning literature on mental health (e.g., Gradus et al., 2020). Nevertheless, suicide only captures the tail end of the mental health distribution; it is not an outcome across all mental health symptom cluster types and is more common in males. If over-treatment is more of a concern, using treatment status may prove a better proxy. However, it is similarly not chosen given that females are less likely to be treated conditional on diagnosis, leading to a double misidentification. If under-diagnosis is more of a concern, future diagnosis status is an additional alternative dependent variable. One's future diagnostic status may reduce the rate of missed diagnoses by capturing delayed diagnoses or those with later ages of onset. However, there is also a higher propensity for over-treatment with time. I test the use of future diagnostic status and suicide ideation as alternative outcome variables. Suicide ideation and future diagnosis are available in the NLSCY for only older adolescents and thus is a truncated sample; I find this model to be less predictive of later life outcomes.

Consequently, I use current diagnosis as the baseline dependent variable to measure mental health risk. To minimize base rate error, I include no demographic, socioeconomic and location characteristics that might drive diagnosis outside of health-based risk factors. As a robustness test, I build a model incorporating demographic, socioeconomic, medical access and other factors that might lead to selection bias. Full modelling details including these models and comparisons to linear probability models and logit models and a model of suicide ideation risk can be found in Table A7 in Appendix B.

4.2 Evaluating Diagnostic Selection

Before estimating mental health treatment effects, it is of interest to understand drivers of treatment selection and misdiagnosis. As detailed in Section 2.2, risk misperception may be a critical driver of treatment selection. Understanding whether a risk misperception is occurring in mental illness diagnostic decisions requires a way to control for unobservable factors that may impact risk. School starting age offers a useful identification strategy for this context. Differences in student starting age based on eligibility dates for enrolling in primary school are known to cause a higher likelihood of being diagnosed with ADHD, learning disabilities and special education placement for the youngest children in a grade (e.g. Elder, 2010; Dhuey & Lipscomb, 2010). This jump in diagnostic likelihood occurs from an over-referral of the youngest students and under-referral of the oldest students based on their relative maturities (Furzer et al., 2019). As eligibility dates randomly sort students to be young or old for their grade based on an exogenous eligibility date, this higher diagnosis is independent of observable or unobservable factors outside of school entry age.

I extend this common identification strategy by estimating the school starting age discontinuity across the

predicted mental health risk distribution. If all marginal diagnoses induced by being young-for-grade come from the top of the risk distribution, there is likely some unobserved information driving diagnosis. In this case, a child's school starting age acts as a useful signal for diagnosing more high-risk children. On the other hand, if marginal diagnoses come from throughout the risk distribution or are concentrated in the low-end of the risk distribution, this would indicate over-treatment-children who are being diagnosed with a disorder they are unlikely to have. Under this scenario, school starting age causes a risk misperception, leading to diagnosis in young but low-risk children instead of their higher-risk peers.

I take a non-parametric approach to estimate the discontinuous jump in diagnosis conditional on risk for children born just before and just after their school eligibility cut-off date. I construct precise relative kindergarten enrollment ages using the longitudinal information available in the NLSCY. For most children, the year they began kindergarten is available in the NLSCY. I record their age on September 1st of their kindergarten starting year and their birth date relative to their school district cut-off date for that year and carry this value forward for all subsequent years. For those individuals with no direct kindergarten information, I impute age-at-entry and relative birth date using their birth date and province of residence, along with details on grade retention or skipping and if the family moved Provinces.[6] I use this to impute a precise age at school entry relative to the median age in their kindergarten class for their year and province. The average age at kindergarten entry was 5.14. The average relative birth date was 1.7 days before the cut-off.

I then compare rates of diagnosis for those whose birth dates fall just before and just after the school eligibility cut-off dates, within a 44-day window, by age-standardized predicted mental health risk ventile. This bandwidth was estimated using a mean-squared error optimal bandwidth selector (Calonico, Cattaneo, Farrell, & Titiunik, 2018). Age- and sex-standardized risk ventiles are calculated by dividing each age-sex cohort of children and adolescents into 20 evenly distributed bins based on their predicted risk. I then separately estimate the proportion of individuals diagnosed currently and cumulatively over time within each risk ventile for older and younger student groupings. Finally, I estimate the difference in these diagnosis rates for younger versus older children for each risk ventile. This approach yields a non-parametric discontinuity at each ventile of risk and provides a means for testing misdiagnosis, holding any unobservable information constant.

There is substantial work showing school starting age is random, with little evidence of birth timing to school entry cut-off dates (Dickert-Conlin & Elder, 2010). While seasonality in some disorders like asthma is of concern, there is limited evidence of seasonality for mental health disorders (Currie & Schwandt, 2013).

[6]Kids who entered the NLSCY after beginning kindergarten may pose some complications in that they started kindergarten in a different Province than they currently reside. Throughout the longitudinal data, only 1.2 percent of children report moving Provinces, so this number is likely small.

Furthermore, in Canada, eligibility dates differ across provinces, school district, and time, thus alleviating birth seasonality concerns. Holding one's child back from enrolment in the first year they are eligible is a concern in the school starting age literature (Evans et al., 2010). While parents cannot usually enrol their children in school early, they can hold them back if they feel they are not ready. To address this, I estimate marginal diagnoses based on both a compliant sample and a non-compliant sample. A compliant sample is where kids predicted to be young based on their birth date were indeed young-for-grade and predicted older kids were old-for-grade. Estimating on a compliant versus a non-compliant sample does not impact results overall (due to reporting rules from the CRDCN, these results cannot be publicly shared).

4.3 Short-term Treatment Effects

After obtaining a consistent estimate of health-based diagnostic risk, I use this to calculate the heterogeneous effect of mental health diagnosis and treatment on short and long-term outcomes. To model the effect of diagnosis and treatment on injuries, hospitalizations, and drug use, and schooling outcomes of retention, dropout and expulsion, I estimate the following equation:

$$Y_{it} = \beta_1 \alpha_{it} + \beta_2 D_{it} + \beta_3 \alpha_{it} * D_{it} + \omega X_{it} + \lambda_t + \pi_i + \epsilon_{it} \tag{1}$$

Consider first the outcome, Y_{it}, as a basket of adverse medium-term school outcomes of dropout, expulsion and grade retention. The matrix X_{it} contains time-varying family circumstances, including child age, number of family members, household income, maternal depression and smoking. I control for having a diagnosed learning disorder to isolate the mental illness diagnosis and treatment effect. Survey cycle year fixed effects are measured through λ_t. I include individual fixed effects through π_i to control for unobserved heterogeneity impacting upon the likelihood of diagnosis, treatment and adverse outcomes. I test both provincial and city-level fixed effects and cluster standard errors at the individual level. Similar models are run for all other direct outcomes. For binary variable outcomes, I test both linear and probit model specifications.

I measure the effect of mental health diagnosis and treatment through β_2. Alone, this coefficient contains information on the effect of mental illness severity, the effect of diagnosis, the effect of treatment and the potential for upward selection bias. To separate true severity from these factors, I include a predicted health-based diagnostic risk α_{it} and its interaction with diagnosis and treatment. The effect of mental health risk for non-diagnosed individuals is measured through β_1. The heterogeneous effect of diagnosis and treatment by risk is measured through $\beta_2 + \beta_3 \alpha_{it}$. I expect the marginal effect of diagnosis and treatment to be less than or equal to zero for any $\alpha_i \geq \bar{\alpha}$. The indifference point is $\bar{\alpha} = -\frac{\beta_2}{\beta_3}$. In application, I separate D_{it} into two indicators: one for diagnosis alone and one for diagnosis with pharmaceutical treatment, measured

in broad categories of stimulant medications and or tranquilizers. As the NLSCY does not contain reliable data on the use of non-pharmaceutical treatment, I assumed that some children with either diagnosis alone or diagnosis with pharmaceutical treatment also receive non-pharmaceutical therapy.

Simultaneity between Y_{it} and both D_{it} and α_{it} may remain a concern if adverse schooling outcomes also impact behaviour, increasing risk and likelihood of identification bias. The relationship between mental illness identification and adverse schooling outcomes is likely only to be one-directional, given that the identification process is not instantaneous. Therefore, while schooling outcomes are considered in the diagnostic process, previous schooling issues, not contemporaneous ones, impact identification. Doing poorly in school may nevertheless change current behaviour and affect contemporaneously calculated risk. To test for this, I regress predicted risk on a basket of negative schooling outcomes. I find no significant effect when controlling for individual fixed effects and other covariates and conclude simultaneity is not of concern.

4.4 Long-term Treatment Effects

To estimate the relationship between childhood mental health risk and diagnosis and treatment on long-term health and socioeconomic status, I estimate the equation:

$$Y_{i,y} = \beta_0 + \beta_1 \alpha_i + \beta_2 D_i + \beta_3 \alpha_i * D_i + \omega X_{i,y} + p_p + \lambda_y + \epsilon_{i,y} \tag{2}$$

For socioeconomic variables, the outcome variable $Y_{i,y}$ stands for years of post-secondary education, ever receiving social assistance and years with income below the low-income cut-off for individual i whose age 15 mental health risk and treatment was recorded in year y. Individual ages range from 20 to 33 in 2015. Predicted mental illness risk α_i, diagnosis and treatment status and all covariates are recorded at age 15. If measures at age 15 were not available, I used the next possible age at measurement, but restrict the lower age limit to age 10. I choose the period from ages 10 to 15, given that the rate of diagnosis peaks at this time. All covariates are analogous to equation (1), with the addition of endowment controls of mother's age at first child, birth weight, and sex of individual. Province fixed effects are measured through p_p, and fixed effects for the year of birth are measured through λ_y. Errors are clustered at the year-of-birth level. Health-related outcomes are similarly estimated to test the effect of adolescent mental health diagnosis and treatment, conditional on risk, on injuries via number of WCB claims and suicide ideation.

To control for additional observable factors impacting mental health identification and labour market outcomes, I additionally control for adolescent schooling outcomes and current day factors. Current day channels impacting labour market outcomes include average other reported home income over 2000 to 2015, parental status, and higher education. Combining these factors while also explicitly controlling for underlying

risk should absorb much of the omitted variable bias impacting both treatment selection and long-term outcomes by health-based risk. To account for any remaining unobserved heterogeneity, I use a neighbourhood fixed effects strategy. Neighbourhoods act as critical channels through which both mental health and labour market outcomes operate (Kling, Liebman, & Katz, 2007). Equation 2 is re-estimated as:

$$Y_{ij} = \beta_1 \alpha_{ij} + \beta_2 D_{ij} + \beta_3 \alpha_{i,j} * D_{ij} + \omega X_{i,j} + n_j + \lambda_y + \epsilon_{i,j} \tag{3}$$

Neighbourhood effects are recorded through n_j, and standard errors are clustered at the neighbourhood level. I base neighbourhoods on the first three digits of the child's household postal code. These digits corresponds to a close geographic region (forward sortation area) with an average of 8,000 individuals per forward sortation area. These geographic regions have families with similar characteristics. Recent work shows that one's teenage neighbourhood of residence matters most long term, making neighbourhood effects as measured at age 15 pertinent (Deutscher, 2018). One's neighbourhood is also likely a critical proxy for medical care access, physician diagnostic and prescription preferences (Chorniy & Kitashima, 2016) and unobservable family-level preferences for treatment. Use of these geographic neighbourhoods thus controls for neighbourhood-level heterogeneity in risk, identification and longer-term outcomes.

5 Results

5.1 Mental Health Diagnostic Risk

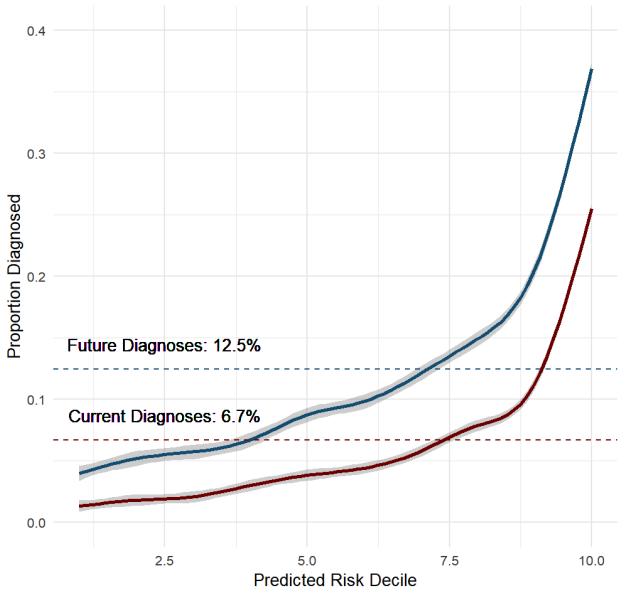
I begin my analysis with results from the machine-predicted health-based diagnostic risk measure, as evaluated on a hold-out test sample. Running on a scale of low (zero) to high (one) risk, the model predicts a majority of test sample individuals to be low-risk.[7] The most critical variables to predict health-based diagnostic risk, measured via frequency of use in the algorithm, included birth weight and child age (Figure A2). Behavioural composite scales, listed in order of importance were: ADHD, anxiety and depression, pro-social skills, and conduct/aggression. BMI and asthma diagnoses also contribute. Parents' ages, maternal depression, family functioning, and social support are also integral. Machine-driven predictions of variable importance in the diagnostic process broadly align with prevailing theories around the contribution of health endowments, family environment, hereditary health and adverse events to child health (Almond & Currie, 2011).

Separating predicted risk into age-standardized deciles, I plot the proportion of individuals diagnosed within each risk bin. Figure 1 shows a strong correlation between predicted health-based diagnostic risk and realized diagnosis, with 55 percent of all diagnoses coming from the top 20 percent of the distribution. In

[7] Around 84 percent of children in the sample have a risk score of less than 0.10 and three percent have a score greater than 0.25.

the top decile, 26 percent of children are diagnosed at a given point-in-time, and 37 percent over the course of the data sample. For children in the bottom half of the risk distribution, less than three percent are diagnosed within each decile. Subsequently, 25 percent of all future diagnoses come from this bottom half of the distribution. The correlation between predicted health-based diagnosis and actual diagnosis indicates a correct prediction of mental health risk.[8]

Figure 1: Proportion Diagnosis by Predicted Risk Decile



Note: Figure shows the proportion of children and adolescents diagnosed at a given point in time (red) and cumulatively in the future (blue) by predicted mental health risk decile. Dotted lines indicate the average overall proportion diagnosed currently (bottom) and subsequently (top) in a hold out test sample.

Nevertheless, if mental health diagnosis and treatment decreases risk via symptom containment, those already diagnosed might be predicted to be low-risk. I measure average risk by time-to- and since-diagnosis to investigate this possibility. Appendix Figure A3 shows predicted risk to be highly correlated with time-since-diagnosis, with no downward discontinuity in average mental health risk after diagnosis. This finding mitigates concerns that diagnosed and treated low-risk individuals are individuals benefiting from treatment. For subsequent diagnoses of currently low-risk kids, I note that risk predictions are point-in-time estimates and may not capture episodic mental health issues or onset of illness occurring at later ages that would not be captured in hereditary risk. While I do include exogenous factors, like experiencing an adverse childhood event in each point-in-time estimate, I do not control for exogenous risk transitions over time. Still, predicted risk is stable as it is based on many overall risk factors, not just current symptom manifestations. Children and adolescents tend to stay within the same risk decile or a neighbouring decile.

I further underscore the risk prediction's utility by assessing adverse events in non-diagnosed individuals

[8]Modeling metrics are available in Appendix B.

by risk. Appendix Figure A4 shows rates of suicide ideation (Panel A) and workplace injuries (Panel B) by risk decile in individuals never receiving a diagnosis. Both rates of suicide ideation and injury rise with predicted risk, showing that the algorithm is identifying individuals with a high risk of mental health-related adverse outcomes but not receiving appropriate care. To conservatively estimate this unmet care need, I calculate the number of undiagnosed individuals in the top 20 percent of the risk distribution who experience an adverse event. In the longitudinal sample, 19 percent of high-risk but never diagnosed individuals experience suicide ideation. Diagnosing these adolescents but leaving all other diagnoses unchanged would increase the overall mental health diagnosis rate by three percentage points to 15 percent.[9]

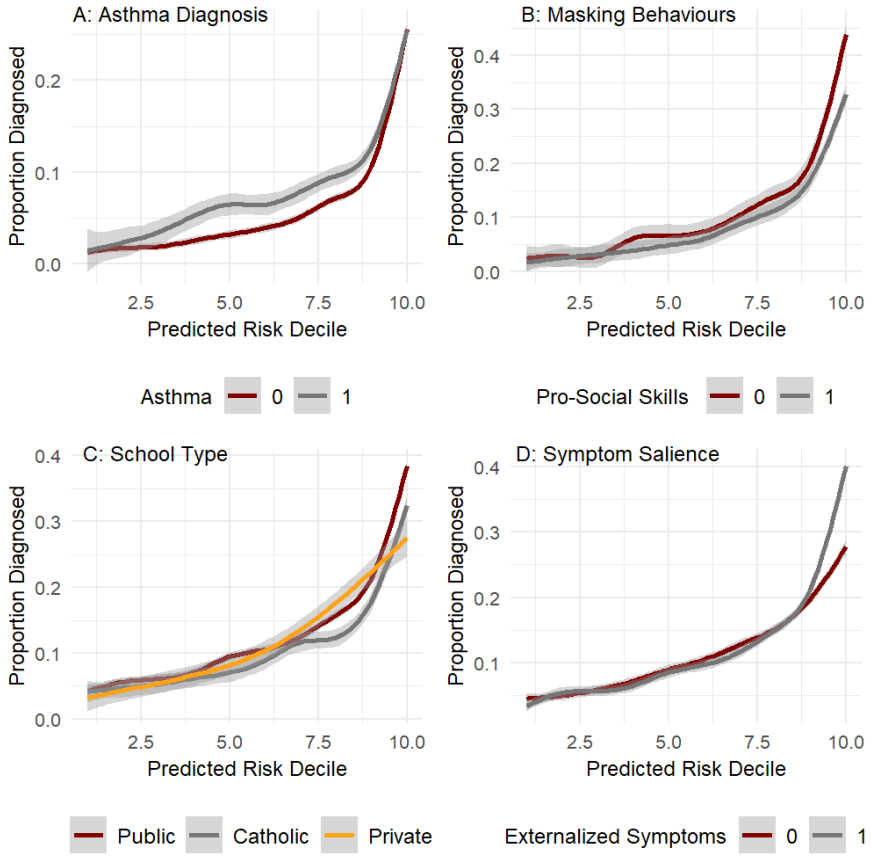
5.1.1 Drivers of Misdiagnosis

What might drive low-risk diagnosis or missed high-risk diagnoses? I first evaluate differences in demographic, medical and location characteristics for those with discordant predicted risk and diagnostic status (Appendix Table A2). Prediction outcomes fall into four categories: those predicted to be at high risk and diagnosed (True Positive) or not diagnosed (Type I error or False Positive), and low risk diagnosed (Type II error or False Negative) or not diagnosed (True Negative). Type II errors are associated with prior asthma diagnoses, Quebec residency, and living in an urban metropolitan area. I specifically show results from Quebec as this was the only Province to have a universal pharmaceutical drug coverage program at the time, which lead to increased mental health treatment (Currie et al., 2014). There is also evidence of higher Type I errors in non-white children, though given that the overwhelming majority of children in the test set are white-identifying this result is not precisely estimated. Despite excluding learning disability from the equation, those with a diagnosis and predicted either high or low risk have significantly higher rates of learning disorders than the average child.

Expanding this, I use a linear probability model to assess what factors impact Type I and Type II error. I incorporate both individual and time fixed-effects to control for unobserved selection and control for the predicted risk, the 20 variables with the highest algorithmic gain, and demographic and medical access factors left out of the initial risk prediction. Pro-social behaviours and being female (a 1.9pp increase) were the only variables to predict a Type I error conditional on predicted risk and characteristics driving risk. Medical contacts and having a learning disorder increased Type II errors. Maternal health conditions and being in a public school further exacerbated Type II errors, while being in a private school decreased Type II errors (all results available upon request).

[9] About 12 percent of the total non-diagnosed population experiences suicidal ideation. If everyone experiencing suicide ideation were to receive diagnosis and treatment, this would increase the diagnostic rate by eight percentage points. Given diagnostic costs associated with low-risk diagnosis, as I will detail in Section 5.2, this is likely an over-estimate. As 14 percent of individuals in the top 20% of risk experience a later life workplace injury, estimating under-treatment by this adverse outcome would result in a two percentage point increase in diagnoses.

Figure 2: Diagnosis Rate by Predicted Risk and Diagnostic Mechanisms



Note: Figure shows the proportion of children and adolescents who are diagnosed by predicted mental illness risk decile and by prior asthma diagnosis (Panel A), whether they have higher than median pro-social skills (Panel B), the school type they attended (Panel C) and whether their behaviours are more externalizing or internalizing (Panel D). Panel A uses current diagnostic status. All other panels use cumulative diagnoses over time.

To understand some of these mechanisms further, I re-estimate the proportion of individuals diagnosed within each predicted mental illness health risk decile, and divide by prior asthma diagnosis (Panel A), pro-social behaviours (Panel B), school type (Panel C), and externalized symptom dominance (Panel D) in Figure 2. I estimate the current mental health diagnosis rate by diagnosed asthma status, knowing the high co-morbidity between these disorders reported here and in previous studies (Chorniy, Currie, & Sonchak, 2018). The differences in diagnosis rates by prior asthma diagnosis are greatest in the middle of the distribution, suggesting prior asthma diagnosis and more access to physicians drives the likelihood of diagnosis of a mental illness for lower-risk children. I next generate a value for low or high pro-social skills by estimating the median level of pro-social skills on the CBS scale. High-risk children with high levels of pro-social behaviours maintain consistently lower diagnostic rates. This result suggests the presence of masking behaviour, whereby individuals can hide their mental health symptoms through compensatory behaviour; masking is especially common in girls with mental health issues. Plotting diagnosis by risk for

those in public, private or Catholic school districts, I find that Catholic school districts have lower diagnosis rates across the risk distribution. Public schools notably have a higher proportion of high-risk diagnoses compared to either private or catholic schools. These results underline how differing funding incentives and budget constraints also likely lead to different diagnosis rates by school type. While these incentives drive more diagnoses in public schools across the risk distribution, it also decreases the rate of under-diagnosis of the most at-risk. Finally, estimating current diagnosis by whether a child exhibits more externalized or internalized symptoms, I find that internalized symptoms decrease diagnosis for high-risk kids only. The fact that more internalized symptoms are less likely to be diagnosed highlights the influence of symptom salience in the diagnostic process.

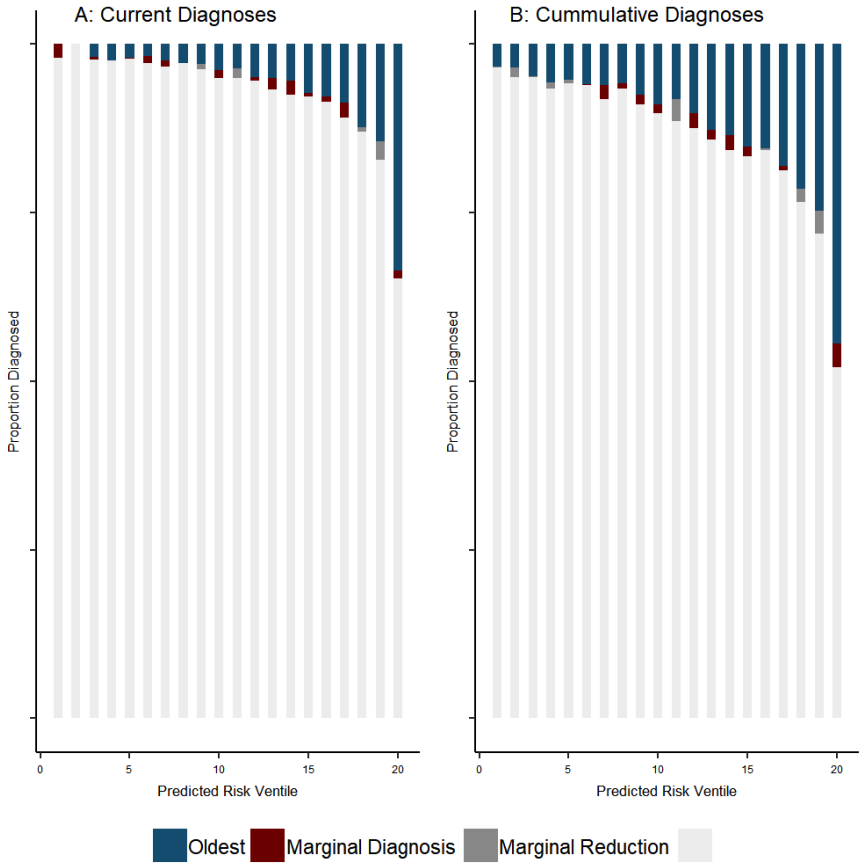
5.1.2 Risk Misperception by School Starting Age

Comparing discrepancies in predicted health-based diagnostic risk versus diagnostic status across various mechanisms shows that behaviour masking, salience, and medical access contribute to potential misdiagnosis. However, there may be factors observable to a physician, but unobservable to the risk algorithm, leading to risk mismeasurement. I use a child's primary school starting age to as a driver of diagnoses at-the-margin and control for potential unobserved heterogeneity. From Figure 3, a marginally diagnosed young-for-grade child comes from across the risk distribution. To be conservative, consider a "low-risk" diagnosis to be anything in the bottom half of the risk distribution. Being young-for-grade increases the likelihood of a low-risk diagnosis by 24 percent in early years. In the relatively younger student cohort, 20 percent of diagnoses come from the bottom half of the risk distribution. In comparison, 16 percent of diagnoses come from the bottom half of the distribution for the oldest children. This gap closes somewhat over time, with a 15 percent difference in low-risk future diagnoses by school starting age; 25 percent of diagnoses come from the bottom half of the youngest-for-grade distribution, but only 22 percent of diagnoses in the older peer sample do.

Concurrently, a small portion of marginal diagnoses come from the top ventile of risk. Although there is no apparent difference in the proportion younger (35 percent) versus older (34 percent) students diagnosed in the top ventile, this widens with time. However, the diagnostic difference between younger and older students in a grade rises to four percentage points (or 9 percent) when assessing cumulative future diagnoses. This difference indicates that school starting age does signal high-risk children at least some of the time. Yet, many low-risk children are still being diagnosed over high-risk ones, even when controlling for unobservable selection. These results suggest that a misprediction of mental health risk by either physicians or parents, or both, is likely occurring, leading to misdiagnosis.

I note that youngest and oldest children per grade are equally likely to be in the top risk decile and have similar average risk scores in the top decile. The youngest children in the top decile have an average risk

Figure 3: Marginal Diagnosis by School Starting Age and Predicted Risk



Note: Figure shows the proportion of children and adolescents diagnosed currently (Panel A) and cumulatively (Panel B) by ventile of risk. Figures are based on a test-set sample of children born 65 (44) days before or after their school enrollment cut-off date – calculated using an MSE-optimal bandwidth estimator. I estimate the proportion of oldest kids in a grade diagnosed by risk ventile (blue). Marginal diagnoses is the proportion of youngest children diagnosed by ventile less the proportion diagnosed from the oldest sample. Positive marginal diagnosis (red) and negative marginal diagnoses (grey) are both present. Each figure is truncated by 0.25 for visualization purposes. Testing marginal diagnoses on a compliant sample, excluding any late or early-enrolled children does not change the results. I test for balance of covariates across school starting age and find no significant difference across income, mental handicap, sex, household income, asthma diagnosis and maternal depression. When looking at diagnosis by risk for ADHD diagnoses only or anxiety and depression only, I find that both are exhibit a similar diagnosis pattern.

score of 0.39, while the oldest have an average risk score of 0.38. For the youngest, the relative immaturity of younger students is generating more diagnoses, but this is unrelated to underlying risk as most of the marginal diagnoses come from low-risk children. Future diagnoses exacerbate this pattern.

5.2 Short-run Mental Health Diagnosis and Treatment Effects

Results estimating the impact of mental health treatment on short-term health and education outcomes, conditional on predicted risk, are presented in Table 2. For comparison purposes, I present four specifications for each outcome over two panels. All models include individual fixed effects. Models in Panel A include baseline effects of diagnosis without pharmaceutical treatment and the cumulative effect of diagnosis and

treatment, with and without predicted risk. Panel B incorporates heterogeneous effects by diagnosis with and without medication treatment. In alignment with previous literature, those diagnosed with a mental health disorder show a higher risk of injury, hospitalization, drug use, and increased risk of dropout, expulsion and retention. Those on treatment see a higher likelihood of adverse outcomes than those with diagnosis alone. For instance, diagnosed individuals on treatment see a 30-percentage point increase in injuries and a 6-percentage point increase in adverse schooling outcomes. When incorporating risk, I note a strong correlation between predicted risk and these short-term outcomes, reaffirming that machine learning guided prediction is accurately identifying individuals at high-risk of adverse outcomes. Confirming that underlying mental health risk is a major source of omitted variable bias, incorporating risk serves to decrease diagnosis and treatment effects across all outcomes.

When further allowing for heterogeneous treatment effects by risk, the likelihood of adverse outcomes increases for low-risk diagnosed individuals (Panel B). The negative interaction for diagnosis and treatment by risk reveals an attenuating effect of diagnosis and treatment on adverse outcomes for high risk individuals. To control for remaining selection on observable characteristics, I present a second fully specified model that incorporates a suite of potentially confounding variables-health endowment, family environment and parental health. Incorporating these features reduces the estimated effects of diagnosis and treatment by risk for high and low-risk children. Treatment benefits in the form of reductions in injuries, prescription pill usage, and adverse schooling outcomes for high-risk children remain. While abating effects of treatment on marijuana usage are detected, this is no longer precisely estimated.[10] Outside of marijuana use, increased risk of adverse outcomes remain for diagnosis and treatment across all outcomes.

Schooling outcomes appear to be associated with a higher costs of low-risk diagnosis (Panel A, Figure 4). A low-risk child with a predicted risk of 0 would have a 2.4-percentage point increase in adverse schooling outcomes if they were identified with a mental health disorder and a 5.4-point increase if they also were subsequently on treatment. With a risk score of 1, the highest risk child would have a 6.7-point increase in adverse schooling outcomes compared to their lowest-risk peer. This effect almost disappears with treatment for the highest risk child. Combining the cost of treatment for low-risk children (a 5.4pp increase) with the heterogeneous benefit of treatment by risk (10.1pp decline) generates a 4.7-point reduction in adverse schooling outcomes for the highest-risk child. From these effects I estimate that positive treatment effects are found for anyone above a risk threshold of 0.53. There is a comparably lower threshold point for which treatment benefits accrue for injuries, as treatment appears to confer smaller fixed costs for direct health outcomes. For those above a predicted risk of 0.24, being on treatment reduces the likelihood of injury. This

[10]Mental health risk is also not related to marijuana usage when controlling for other confounding variables. As nearly 50 percent of kids in the sample report trying marijuana, it is logical that use may not be directly tied to underlying mental health.

Table 2: Short-Term Mental Health Diagnosis and Treatment Effects

	Direct Health Outcomes								Schooling Outcomes	
	Injuries		Hospitalization		Marijuana		Pres. Pills		Welfare	
Panel A: Baseline Treatment Effects										
Diagnosis	0.274***	0.209***	0.079***	0.059***	0.216***	0.180***	0.014+	0.011	0.036***	0.030***
	(0.012)	(0.012)	(0.008)	(0.008)	(0.029)	(0.029)	(0.008)	(0.008)	(0.009)	(0.009)
D*Treatment	0.296***	0.182***	0.117***	0.083***	0.469***	0.405***	0.024***	0.018**	0.056***	0.044***
	(0.013)	(0.014)	(0.010)	(0.010)	(0.052)	(0.052)	(0.006)	(0.006)	(0.012)	(0.012)
Risk		0.649***		0.197***		0.591***		0.037**		0.065***
		(0.022)		(0.013)		(0.056)		(0.013)		(0.016)
IndividualFE	x	x	x	x	x	x	x	x	x	x
N	105,000		105,000		34,000		81,000		105,000	
Panel B: Heterogeneous Treatment Effects										
Diagnosis	0.275***	0.044*	0.075***	0.031*	0.224***	-0.036	0.012	0.001	0.036***	0.024*
	(0.013)	(0.017)	(0.009)	(0.012)	(0.033)	(0.036)	(0.009)	(0.012)	(0.010)	(0.012)
D*Treatment	0.296***	0.030+	0.105***	0.027*	0.467***	0.051	0.027***	0.023**	0.064***	0.054*
	(0.016)	(0.017)	(0.012)	(0.012)	(0.060)	(0.050)	(0.007)	(0.008)	(0.013)	(0.014)
Risk	0.901***	0.103***	0.249***	0.061***	0.715***	0.034	0.052***	0.054**	0.102***	0.067**
	(0.028)	(0.029)	(0.015)	(0.018)	(0.065)	(0.066)	(0.015)	(0.018)	(0.019)	(0.022)
D*Risk	-0.609***	0.002	-0.138***	-0.015	-0.375**	-0.048	-0.019	-0.022	-0.070	-0.065
	(0.059)	(0.069)	(0.034)	(0.045)	(0.117)	(0.137)	(0.043)	(0.048)	(0.046)	(0.062)
D*T*Risk	-0.692***	-0.123**	-0.136***	-0.002	-0.359**	-0.180	-0.054*	-0.058+	-0.115**	-0.101*
	(0.047)	(0.047)	(0.030)	(0.033)	(0.139)	(0.134)	(0.027)	(0.031)	(0.043)	(0.047)
IndividualFE	x	x	x	x	x	x	x	x	x	x
Covs		x		x		x		x		x
N	105,000	72,000	105,000	72,000	34,000	20,000	81,000	58,000	105,000	72,000

Note: Table presents results from individual fixed effects models of diagnosis and treatment on health and schooling outcomes in childhood. Panel A shows baseline models with no additional controls. Panel B shows two models for each outcome: baseline models controlling only for the specified variables and fully specified models controlling additionally for learning disabilities, child age, log-household income, number of siblings, blended family status, attendance at public, private or catholic school, parent smoking, maternal depression, and year and city fixed effects. School welfare measures if a child was retained in a grade, dropped out for more than a week, or was expelled. All N are rounded to the nearest 1000. Standard errors in parentheses are clustered at the individual level: $***p < 0.001$, $**p < 0.01$, $*p < 0.05$,$+p < 0.10$

predicted risk level is still well within the highest risk categories. For reference, 0.25 is the average risk score of a child in the top decile of risk.[11]

5.3 Long-run Mental Health Diagnosis and Treatment Effects

I next present estimated effects of diagnosis and treatment on longer term health and socioeconomic outcomes (Table 3). I present two models for each variable of interest. In the first model for each variable, I explicitly measure treatment heterogeneity by incorporating risk and allowing for differing diagnosis and treatment effects by risk. I also control for a suite of observable health endowment, family environment, adolescent schooling and health, and exogenous adulthood factors that might lead to upwards bias in the estimated coefficients. In the second column of each outcome, I include neighbourhood fixed effects to control for remaining unobserved heterogeneity. When comparing results with and without neighbourhood fixed effects, the estimated coefficients remain largely consistent when incorporating neighbourhood fixed effects. This consistency suggests that the models correctly capture observable and unobservable heterogeneity to produce an unbiased treatment effect estimate. I focus on the fully specified neighbourhood fixed effects models to detail the results below.

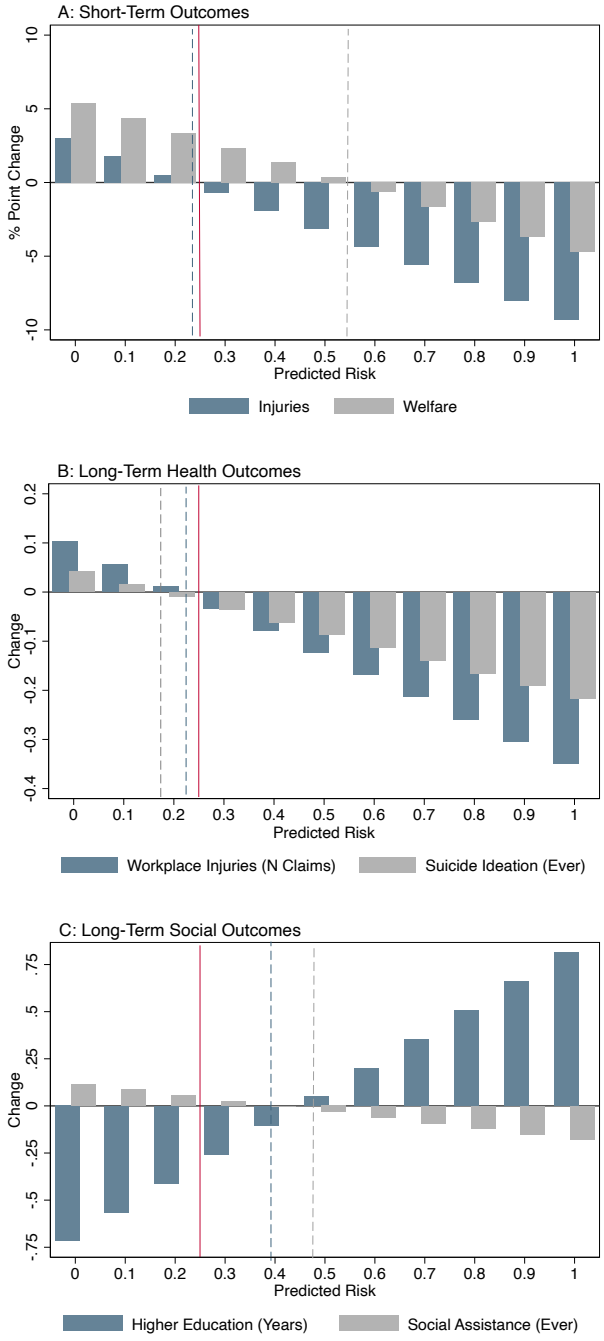
Table 3 first shows that suicide ideation and later life injuries (measured through WCB claims) both rise with underlying predicted health-based diagnostic risk. The highest risk individuals have as much as a 46-point increase in suicide ideation, or 1.36 standard deviations. However, diagnosed low-risk adolescents have a 4.2- to 5.4-point increase in suicide ideation depending on treatment status. This aligns with the small increased risk of suicidality with anti-depressant use in individuals under the age of 18 (R. A. Friedman, 2014). Nevertheless, high-risk individuals see reductions in suicide ideation by 26-percentage points with diagnosis and treatment (Panel B, Figure 4). At the same time, injuries increase nearly half of a standard deviation for the highest risk individual. Yet, high risk individuals who are diagnosed and on treatment see reductions in injury claims by 0.35 percentage points (0.76 standard deviations). With 8 billion in annual WCB claims paid per year in Canada. Using a back-of-the-envelop calculation, I estimate a potential annual downstream savings of approximately 612 to 879 million if the diagnosis rate for those in the top decile of risk increased from 40 percent to 100 percent.[12]

For socioeconomic outcomes, underlying mental health risk remains a robust predictor of adverse outcomes in later life, including fewer years of higher education, more years in low-income, and a higher likelihood

[11] I present results from a linear probability model for better interpretation but similar results using a probit specification. I use a random-effects model and individual, age, year and both city size and province fixed effects for the probit specification (results available upon request).

[12] Numbers based on three million adults assumed to be in the top decile of mental health risk, with 60 percent non-diagnosed (1.8M). 14 percent of non-diagnosed individuals in the top decile of risk have a claim, or 252,000 claims. Diagnosis alone reduces claims for this group by 61,230, or 87,950 with pharmaceutical treatment at $10,000 per claim on average.

Figure 4: Heterogeneous Effects of Diagnosis and Pharmaceutical Treatment by Risk



Note: Figure shows the calculated treatment effects of diagnosis and pharmaceutical treatment by predicted mental health risk for short term health and schooling welfare - dropout, expulsion and retention (Panel A). Longer-term health outcomes (Panel B) include the change in likelihood of suicide ideation (grey) and number of workplace injury claims (blue). Longer-term social outcomes (Panel C) include likelihood of social assistance receipt (grey) and years of post-secondary attendance (blue). For long-term outcomes, predicted mental illness risk and treatment status is measured at age 15. Red lines indicate the average risk score for an individual in the 10^{th} decile of risk (0.25). Dotted lines represent the threshold risk level by which positive treatment benefits begin for a given outcome.

of requiring income support. Mirroring short-term outcomes, there is a more robust penalty for low-risk diagnoses in socioeconomic outcomes than health outcomes. Low-risk and diagnosed individuals have a 0.6 to 0.7 fewer years of higher education and a 6.5- to 11.5-percentage point increase in the likelihood of requiring social assistance (Panel C, Figure 4). Diagnosis and treatment for high-risk individuals increases higher education training by 0.78 years and 0.25 fewer years in low-income. The likelihood of requiring social assistance also declines by 18 percentage points.

5.3.1 Mental Health Treatment and Income

I next model the effects of adolescent mental health diagnosis and treatment on long-term income (Table 4). At baseline, I find a significant negative effect of having a diagnosed mental health disorder in adolescence on later life income. Unconditionally, those with a diagnosed mental health disorder experience a 23 to 30 percent income reduction, dependent on treatment status (Column 1).[13] These models include fixed effects for year, age and the current province of the individual. When controlling for early life health endowments, the negative effect of treatment falls by about 10 percentage points (Column 2). These results align well with the extant literature predicting an approximately 20 percent reduction in self-reported income due to mental health incidence (Fletcher, 2013, 2014; Johar & Truong, 2014). When further controlling for risk (Column 3), educational outcomes (Column 4), and adulthood controls (Column 5), the negative gradient between diagnosis and treatment remains consistently negative, as does the relationship between predicted early-life mental health risk and income. However, once conditioning on both labour force participation and employment, the highest risk individuals receiving pharmaceutical treatment have a 29 percent increase in income (Column 5). Incorporating neighbourhood fixed effects increases this income effect to 64 percent (Column 6). These results demonstrate some potential for positive treatment benefits on income. However, mental illness's cascading impact on health generally, education, and labour force attachment make accruing these benefits exceedingly difficult.

5.4 Sex Differences and Symptom Clusters

Given the reported evidence of higher mental illness impact on short and long term outcomes for females (i.e., Fletcher, 2010; Busch et al., 2014), I assess sex differences in treatment effects. When subset by sex, females' predicted mental health risk is more predictive of all short- and long-term outcomes, except long-term injuries and marijuana use (Appendix Table A4). These results also appear to show a higher cost to low-risk diagnoses and treatment in the short-term for females but higher costs long-term for males in terms of injuries and reduced levels of educational attainment. Females experience a higher rate of suicide ideation

[13] Percentage change in income calculated via $e^{\beta_2} - 1$

Table 3: Long-Term Mental Health Diagnosis and Treatment Effects

	Health Outcomes				Socioeconomic Outcomes					
	Injuries (N WCB Claims)		Suicide Ideation Ever		Years Higher Education		Years Low-Income		Social Assistance Ever	
Mean	0.175		0.136		3.105		1.761		0.481	
(Std. Dev.)	(0.59)		(0.34)		(2.83)		(2.64)		(0.50)	
Diagnosis	0.099*	0.093*	0.051*	0.054*	-0.609***	-0.622***	0.461**	0.362*	0.065**	0.065*
	(0.046)	(0.042)	(0.022)	(0.024)	(0.125)	(0.140)	(0.154)	(0.162)	(0.025)	(0.028)
D*Treatment	0.106**	0.102**	0.051**	0.042*	-0.740***	-0.717***	0.425**	0.485***	0.111***	0.115***
	(0.035)	(0.035)	(0.020)	(0.021)	(0.106)	(0.109)	(0.141)	(0.143)	(0.025)	(0.026)
D*Risk	-0.318+	-0.336+	-0.236+	-0.271+	2.338***	2.297***	-1.207	-0.712	-0.201+	-0.158
	(0.186)	(0.182)	(0.128)	(0.141)	(0.575)	(0.633)	(0.809)	(0.850)	(0.118)	(0.135)
D*T*Risk	-0.449***	-0.451***	-0.291***	-0.259**	1.638***	1.531***	-0.587	-0.753	-0.284***	-0.295**
	(0.122)	(0.129)	(0.085)	(0.089)	(0.389)	(0.419)	(0.512)	(0.538)	(0.081)	(0.092)
Risk	0.257**	0.258**	0.453***	0.463***	-1.986***	-1.949***	1.925***	1.911***	0.356***	0.327***
	(0.095)	(0.098)	(0.056)	(0.059)	(0.284)	(0.293)	(0.340)	(0.353)	(0.054)	(0.060)
Endowment Covs	x	x	x	x	x	x	x	x	x	x
Adolescent Covs	x	x	x	x	x	x	x	x	x	x
Current Day Covs	x	x	x	x	x	x	x	x	x	x
NBhood FE		x		x		x		x		x
Adj. R^2	0.083	0.059	0.070	0.071	0.328	0.299	0.317	0.306	0.167	0.155
N	12,000	11,000	10,000	10,000	12,000	11,000	12,000	11,000	12,000	11,000

Note: Table presents the heterogeneous costs and benefits of diagnosis and treatment by risk at age 15 on long term outcomes. Health measures include number of years of WCB injury claims and whether the individual ever reported suicidal ideation. Socioeconomic outcomes include number of years claiming a higher education tax credit and being below the low-income cut-off and ever receiving social assistance. All models include endowment controls of birth weight and maternal age at first birth. Age 15 controls include log household income, a general health score, education indicators for drop-out, expulsion and retention, as well as math test scores. Adulthood controls include presence of a child in the home and log average other household income from 2000 to 2015. Neighbourhood fixed effects are based on residence at age 15. All N are rounded to the nearest 1000 in accordance with rules for this tax data set. Standard errors are in parentheses and clustered at the neighbourhood level: $^{+}$ $p < 0.10$,* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Effect of Mental Health Diagnosis and Treatment on Longer-Term Income

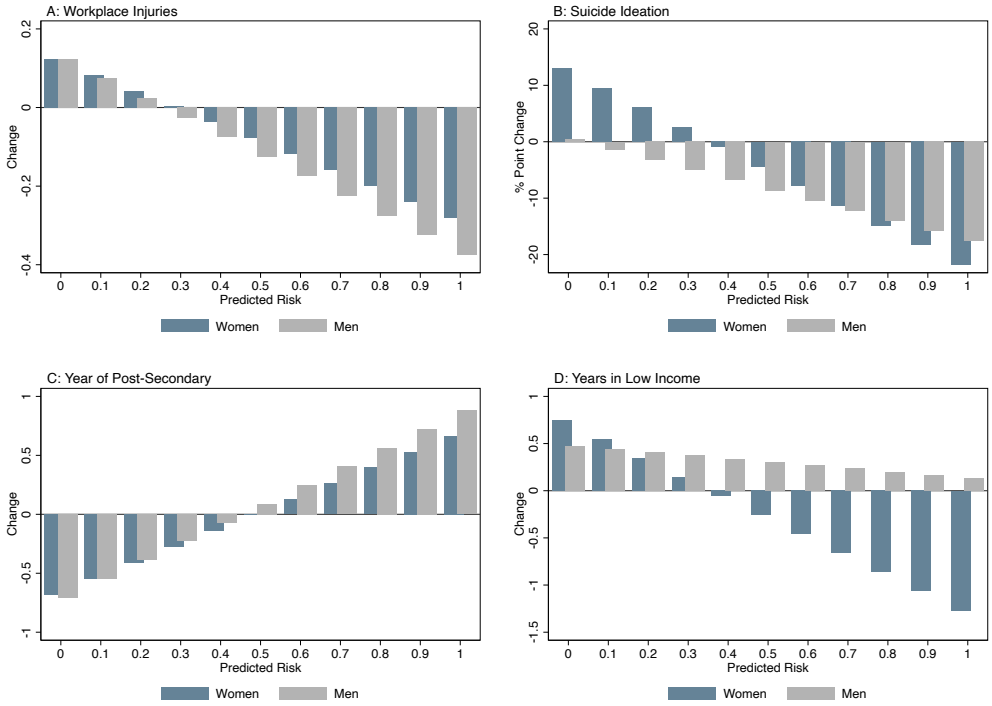
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Full Sample							Males	Females
Diagnosis	-0.266***	-0.254**	-0.287**	-0.210*	-0.141	-0.126^{+}	-0.106	-0.107	0.017
	(0.065)	(0.092)	(0.105)	(0.100)	(0.093)	(0.073)	(0.082)	(0.093)	(0.147)
D*Treatment	-0.352***	-0.223***	-0.153*	-0.111^{+}	-0.040	-0.172**	-0.211**	-0.143	-0.277**
	(0.045)	(0.062)	(0.062)	(0.065)	(0.059)	(0.065)	(0.068)	(0.104)	(0.096)
Risk			-0.448*	-0.324^{+}	-0.329*	-0.391**	-0.429**	-0.505*	-0.342
			(0.190)	(0.195)	(0.150)	(0.138)	(0.153)	(0.224)	(0.246)
D*Risk			0.489	0.224	0.048	0.232	-0.118	-0.144	-1.404
			(0.432)	(0.462)	(0.433)	(0.318)	(0.342)	(0.358)	(0.899)
D*T*Risk			-0.035	-0.124	-0.205	0.370	0.602*	0.540	0.796^{+}
			(0.317)	(0.343)	(0.307)	(0.239)	(0.246)	(0.353)	(0.418)
Endmt		x	x	x	x	x	x	x	x
EducCovs				x	x	x	x	x	x
AdultCovs					x	x	x	x	x
LFP&UMP						x	x	x	x
NeighFE							x	x	x
N	12,000	7,000	7,000	7,000	7,000	5,000	5,000	2,000	3,000
Adj.R^2	0.109	0.135	0.136	0.137	0.113	0.335	0.340	0.334	0.320

Note: Table presents estimates of the effect of mental health in adolescence on later life income. Baseline models include age, year and province fixed effects. Health endowments include birth weight, maternal age at first birth and family income. Neighbourhood fixed effects are based on residence at age 15. Education controls include education indicators for drop-out, expulsion and retention, as well as math test scores. Current day controls include presence of a child in the home, other household income, and whether individual obtained any form of higher education. Labour force participation and employment condition on current labour force attachment. All N are rounded to the nearest 1000. Standard errors in parentheses and clustered at the neighbourhood level: $^{+}$ $p < 0.10$,* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

for low-risk individuals who receive treatment. For treatment benefits in the predicted high-risk group, males see greater benefits in the form of injury prevention in both the short and long-term (Panel A, Figure 5). Both males and females experience reductions in suicide ideation with treatment – a 18 to 35-point decline depending on treatment status (Figure 5). Predicted high-risk males receive more treatment benefits in the form of increased post-secondary education (Panel C), yet females accrue fewer years in low-income.

Note from Table 4, that income effects are only realized after conditioning on one's labour force attachment. Female mental health particularly impacts labour force participation. The highest risk females who received treatment have a 0.29 (0.16) point reduction in labour force participation, while males experience no real impact. At the same time, male employment is more affected by mental health, with the highest risk males on treatment experiencing a 0.36 (0.12) decline in employment (-0.18 for females, but not precisely estimated). These results disentangle overall treatment effects in the extant literature to show strong differential treatment effects by risk. Previous work finds an overall 5.5 percentage point reduction in survey-reported employment for males with depression and a 8.8 percentage point reduction for females (Fletcher, 2013). There is a similar 10 percentage point reduction from ADHD for both males and females (Fletcher, 2014). Using the average risk score value for a female in the top decile (0.25) of risk shows an 8.9 percentage point reduction in labour force participation across this general mental health measure. Similarly, the average

Figure 5: Heterogeneous Effects of Diagnosis and Pharmaceutical Treatment by Sex



Note: Figure plots estimated adolescent mental health treatment effects on long-term workplace injury claims (Panel A), suicide ideation (Panel B), year of post-secondary education (Panel C), and years with income below the low-income cut-off (Panel D). Each plot includes separately estimated models for females (blue) and males (grey). All models control for health endowments, family environment, early schooling, and neighbourhood fixed effects.

high-risk male would see a 3.7 percentage point reduction in employment. These results point out that labour market outcomes are highly risk-dependent.

One critical reason for these sex differences may be a by-product of the risk algorithm. I use a blinded the algorithm incorporating only health risk and symptom factors and removing demographic and socioeconomic factors that may drive base rate error in the model. Despite this blinding, a sex difference in risk remains. The average decile rank for females is 5.25 versus 5.75 for males. A lower predicted risk for females likely derives from the algorithm weighting externalized disorders like ADHD more heavily than internalized disorders like anxiety and depression, and down-weighting risk with pro-social behaviours (Figure A2). These two factors, which are more common in females, may result in a female's true risk being deflated, leading to a higher estimated association. Moreover, if under-weighting impacts the diagnostic decision, diagnosis for any individual with more internalized symptoms might be missed or delayed because of symptom under-weighting. Appendix Figure A5 shows that substantially fewer females are diagnosed, even for the highest risk categories. Yet, in early years, females who are diagnosed are more likely to be in the top 20 percent of risk (Appendix Table A1). This discrepancy evens out over time, with a similar proportion of high risk males and females diagnosed cumulatively over time. These results present an alternative explanation to

previous results finding worse outcomes for females with mental health issues. If female mental health risk is underestimated by parents, teachers or physicians acting as agents on behalf of a child, only the most at-risk females are receiving diagnoses early. If only the most at-risk receive diagnoses and have worse outcomes than the diagnosed but, on average, less at-risk males, then it would appear that mental health issues make females experience more severe mental health issues than males.

Of course, sex-based differences in outcomes by risk and treatment may also be induced by sex differences in diagnostic timing and the symptom clusters themselves. For reference, females make up only 29 percent of the ADHD diagnoses, but 46 percent of the anxiety and depression diagnoses. Mental health risk associations between anxiety, depression and ADHD with short and long-term outcomes appear quite similar Table A6. For low-risk diagnoses, both disorders are associated with diagnostic and treatment costs on social assistance, higher education attainment, and poverty. Anxiety and depression diagnoses appear associated with a higher incidence of suicide ideation in low-risk individuals. Receiving a diagnosis and treatment for ADHD appears to have minimal impact on high-risk individuals' health and socioeconomic outcomes. However, treatment for anxiety and depression appears to reduce suicide ideation, injuries, receipt of social assistance, the likelihood of being in poverty, while increasing years of higher education. These effects might help to explain some sex differences.

As for diagnostic timing, the average age of diagnosis for males is 8.88 versus 10.7 for females (9.57 overall). This timing is mainly due to the earlier age of diagnosis for ADHD, which is more common in males. The average age of diagnosis for ADHD at 8.50 versus 10.65 for anxiety and depression. Sex differences in diagnostic timing are, however, present within symptom clusters. Males with ADHD receive diagnoses at an average age of 8.44. Females are usually aged 8.66. Simultaneously, the average age of diagnosis for anxiety and depression is 9.47 for males and 11.93 for females. As earlier diagnosis appears more optimal for high-risk individuals (see Appendix Table A5), females may be more at risk of having diagnosis occur too late to accrue meaningful benefits. At the same time, males may be at risk of being diagnosed too early, resulting in a diminished human capital channel.

Of course, missed or delayed diagnoses may derive from bias in decision making. I train two additional algorithms to predict health-based diagnostic risk for males and females separately on sex separated training sets to investigate this possibility. I use these male and female risk models to then predict diagnosis on the full test set. I then calculate the diagnostic differences when an individual is assessed "as if" male or female. Firstly, the male algorithm always predicts a higher risk level, with 5.8 percent of males predicted as high-risk and 3.8 percent of females (Appendix Table A7). The two percent gap in risk prediction likely reflects symptom differences. The female risk algorithm maintains a sex gap in those predicted high-risk: 2.6 percent of males are high-risk and 1.5 percent of females from the female-only trained algorithm. These results

suggest that sex-based symptom differences can account for some differences in diagnosis. Nevertheless, there is an eight percent gap in diagnosed males versus females in the top decile of risk. Calculating this difference over the male trained algorithm (17 percent) and the female trained algorithm (13 percent) and taking their difference reveals a four percent gap that cannot be explained by symptom differences. This evidence underscores a case for targeted screening to ensure risk misperception does not guide treatment selection.

6 Conclusion

In this paper, I evaluate the role of diagnostic selection in child mental illness diagnoses and estimate the long-term effects of being on treatment by underlying risk. Through linked tax records, I assess income and labour force dynamics on a longer time horizon than has been conducted in the previous literature. I find that, while mental health diagnosis and treatment show adverse effects on human capital outcomes, these adverse effects are concentrated in low-risk individuals. In the short- and long-term, I find attenuating effects of diagnosis and treatment for high-risk kids. The realized benefits of treatment are even greater for longer-term direct health outcomes like injury and suicide prevention. Identifying heterogeneous mental illness treatment effects by risk contextualizes previous findings of mixed treatment effects in mental illness treatment to show that children at high risk for mental illness benefit from appropriate treatment.

In the process, this paper provides a novel application of ensemble machine learning to predict mental health risk. The gradient boosted tree modelling approach allows for systematic model selection and better captures complex human decision-making processes for rare events like mental health disorders than the standard approaches used in the literature. The prediction results underscore the value of well-known theories around the role of health endowment, family environment, and hereditary health factors on child mental health. Using this predicted risk, I quantitatively show how demographics, medical access, and symptom manifestation serve to inflate or deflate predicted risk for kids, leading to under or over-treatment. Notably, being young-for-grade changed predicted risk for kids at-the-margin, providing a plausible causal mechanism by which some patient's risk is over-estimated. These low-risk diagnoses are nonetheless rare. Missed or delayed diagnoses in the top percentiles of the risk distribution pose a more significant issue, with a lower bound under-treatment estimate of two percentage points, based on unmet care needs.

In alignment with previous literature, I also find that estimated mental health risk and treatment effects have substantial sex differences, with stronger effects of females' underlying mental health risk on adverse outcomes. I present evidence that this is not due to females' inherently worse outcomes from mental illness, but because their more internalized symptoms are under-weighted compared to males. Pro-social behaviours

leading to symptom masking and direct differences in how male versus female risk is perceived regardless of symptoms translate to fewer and later diagnoses for females. Females who do receive diagnoses are then more likely to be high-risk and show comparably worse outcomes.

Given the rise in mental illness in childhood and the potential for significant adverse outcomes, understanding misdiagnosis drivers has significant implications for health-related inequities. This paper highlights the nuance required in evaluating the long-term effects of mental health treatment and underpins the role of medical access inequities on long-term adverse outcomes. For high-risk children, I estimate that diagnosis should occur before age 11 to mitigate long-term adverse outcomes. The fact that low-risk diagnoses are rare and that many, especially females and potentially minority children, are diagnosed seemingly too late suggests broader screening before age 11 as a policy intervention to mitigate long-term mental health inequities.

References

Abaluck, J., Agha, L., Kabrhel, C., Raja, A., & Venkatesh, A. (2016). The determinants of productivity in medical testing: Intensity and allocation of care. *American Economic Review*, *106*(12), 3730–64.

Almond, D., & Currie, J. (2011). Killing Me Softly: The Fetal Origins Hypothesis. *J. Econ. Perspect.*, *25*(3), 153–172. doi: 10.1257/jep.25.3.153

Athey, S. (2018). *The Impact of Machine Learning on Economics.*

Batstra, L., & Frances, A. (2012, jun). Diagnostic Inflation. *J. Nerv. Ment. Dis.*, *200*(6), 474–479. doi: 10.1097/NMD.0b013e318257c4a2

Bertakis, K. D., Helms, L. J., Callahan, E. J., Azari, R., Leigh, P., & Robbins, J. A. (2001, sep). Patient Gender Differences in the Diagnosis of Depression in Primary Care. *J. Womens. Health Gend. Based. Med.*, *10*(7), 689–698. doi: 10.1089/15246090152563579

Bitsko, R. H., Holbrook, J. R., Ghandour, R. M., Blumberg, S. J., Visser, S. N., Perou, R., & Walkup, J. T. (2018). Epidemiology and impact of health care provider–diagnosed anxiety and depression among us children. *Journal of developmental and behavioral pediatrics: JDBP*, *39*(5), 395.

Bruchmüller, K., Margraf, J., & Schneider, S. (2012). Is ADHD Diagnosed in Accord With Diagnostic Criteria? Overdiagnosis and Influence of Client Gender on Diagnosis. *J. Consult. Clin. Psychol.*, *80*(1), 128–138.

Busch, S. H., Golberstein, E., & Meara, E. (2014). The fda and abcs unintended consequences of antidepressant warnings on human capital. *Journal of Human Resources*, *49*(3), 540–571.

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2018). RDROBUST: Stata module to provide robust data-driven inference in the regression-discontinuity design. *Statistical Software Components*(S458483).

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., ... others (2014). The p factor: one general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science*, *2*(2), 119–137.

Chan, E., Fogler, J. M., & Hammerness, P. G. (2016, may). Treatment of Attention-Deficit/Hyperactivity Disorder in Adolescents. *JAMA*, *315*(18), 1997. doi: 10.1001/jama.2016.5453

Chandra, A., & Staiger, D. O. (2007). Productivity spillovers in health care: evidence from the treatment of heart attacks. *Journal of Political Economy*, *115*(1), 103–140.

Chorniy, A., Currie, J., & Sonchak, L. (2018, jul). Exploding asthma and ADHD caseloads: The role of medicaid managed care. *J. Health Econ.*, *60*, 1–15. doi: 10.1016/j.jhealeco.2018.04.002

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

Cornaglia, F., Crivellaro, E., & McNally, S. (2015, apr). Mental health and education decisions. *Labour Econ.*, *33*, 1–12. doi: 10.1016/j.labeco.2015.01.005

Cronin, C. J., Forsstrom, M. P., & Papageorge, N. W. (2020). *What good are treatment effects without treatment? mental health and the reluctance to use talk therapy.*

Cuffe, S. P., Moore, C. G., & Mckeown, R. E. (2005). Prevalence and Correlates of ADHD Symptoms in the National Health Interview Survey. *J. Atten. Disord.*, *9*(2), 392–401.

Currie, J., & Schwandt, H. (2013). Within-mother analysis of seasonal patterns in health at birth. *Proceedings of the National Academy of Sciences*, *110*(30), 12265–12270.

Currie, J., & Stabile, M. (2006). Child mental health and human capital accumulation: The case of ADHD. *J. Health Econ.*, *25*(6), 1094–1118.

Currie, J., & Stabile, M. (2009). Mental Health in Childhood and Human Capital. *Probl. Disadvantaged Youth An Econ. Perspect.*(December 2014), 115 – 148. doi: 10.1300/J171v04n02

Currie, J., Stabile, M., & Jones, L. (2014). Do stimulant medications improve educational and behavioral outcomes for children with ADHD? *J. Health Econ.*, *37*, 58–69. doi: 10.1016/j.jhealeco.2014.05.002

Daley, D., & Birchwood, J. (2010, jul). ADHD and academic performance: why does ADHD impact on academic performance and what can be done to support ADHD children in the classroom? *Child. Care. Health Dev.*, *36*(4), 455–464.

Dalsgaard, S., Nielsen, H. S., & Simonsen, M. (2014). Consequences of ADHD medication use for children's outcomes. *J. Health Econ.*, *37*, 137–151. doi: 10.1016/j.jhealeco.2014.05.005

Deutscher, N. (2018). Place, jobs, peers and the teenage years: exposure effects and intergenerational mobility. *Tax and Transfer Policy Institute-Working Paper*, *10*.

Dhuey, E., Figlio, D., Karbownik, K., & Roth, J. (2019, jun). School Starting Age and Cognitive Development. *J. Policy Anal. Manag.*, *38*(3), 538–578. doi: 10.1002/pam.22135

Dhuey, E., & Lipscomb, S. (2010). Disabled or Young? Relative Age and Special Education Diagnoses in Schools. *Econ. Educ. Rev.*, *29*(5), 857–872.

Dhuey, E., & Lipscomb, S. (2011, apr). Funding Special Education by Capitation: Evidence from State Finance Reforms. *Educ. Financ. Policy*, *6*(2), 168–201.

Dhuey, E., & Lipscomb, S. (2013, jul). Funding Special Education by Total District Enrollment: Advantages, Disadvantages, and Policy Considerations. *Educ. Financ. Policy*, *8*(3), 316–331. doi: 10.1162/EDFP_a_00098

Dickert-Conlin, S., & Elder, T. (2010, oct). Suburban legend: School cutoff dates and the timing of births. *Econ. Educ. Rev.*, *29*(5), 826–841. doi: 10.1016/j.econedurev.2010.03.004

Elder, T. E. (2010). The importance of relative standards in ADHD diagnoses: Evidence based on exact birth dates. *J. Health Econ.*, *29*(5), 641–656.

Evans, W. N., Morrill, M. S., & Parente, S. T. (2010). Measuring Inappropriate Medical Diagnosis and Treatment in Survey Data: The Case of ADHD among School-Age Children. *J. Health Econ.*, *29*(5), 657–673.

Faraone, S. V., & Larsson, H. (2019, apr). *Genetics of attention deficit hyperactivity disorder* (Vol. 24) (No. 4). Nature Publishing Group. doi: 10.1038/s41380-018-0070-0

Fletcher, J. (2008). Adolescent Depression: Diagnosis, Treatment, and Educational Attainment. *Health Econ.*, *17*, 1215–1235. doi: 10.1002/hec.1319

Fletcher, J. (2010). Spillover Effects of Inclusion of Classmates with Emotional Problems. *J. Policy Anal. Manag.*, *29*(1), 69–83.

Fletcher, J. (2013, jul). Adolescent Depression and Adult Labor Market Outcomes. *South. Econ. J.*, *80*(1), 26–49. doi: 10.4284/0038-4038-2011.193

Fletcher, J. (2014). The effects of childhood ADHD on adult labor market outcomes. *Health Econ.*, *23*(2), 159–181.

Frank, R., & McGuire, T. (1999). *Economics and Mental Health.*

Frank, R. G., & Meara, E. (2009). *The effect of maternal depression and substance abuse on child human capital development* (Tech. Rep.). National Bureau of Economic Research.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.*, *29*(5), 1189–1232.

Friedman, J. H. (2002, feb). Stochastic gradient boosting. *Comput. Stat. Data Anal.*, *38*(4), 367–378. doi: 10.1016/S0167-9473(01)00065-2

Friedman, R. A. (2014). Antidepressants' black-box warning—10 years later. *New England Journal of Medicine*, *371*(18), 1666–1668.

Froehlich, T. E., Lanphear, B. P., Epstein, J. N., Barbaresi, W. J., Katusic, S. K., & Kahn, R. S. (2007, sep). Prevalence, Recognition, and Treatment of Attention-Deficit/Hyperactivity Disorder in a National Sample of US Children. *Arch. Pediatr. Adolesc. Med.*, *161*(9), 857.

Furzer, J., Dhuey, E., & Laporte, A. (2019). *The Over/Under on ADHD Misidentification: Relative Age Effects, Direction, and Mitigators.*

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012, jul). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.*, *42*(4), 463–484. doi: 10.1109/TSMCC.2011.2161285

Ginsburg, G. S., Becker-Haimes, E. M., Keeton, C., Kendall, P. C., Iyengar, S., Sakolsky, D., ... Piacentini, J. (2018). Results from the child/adolescent anxiety multimodal extended long-term study (camels): primary anxiety outcomes. *Journal of the American Academy of Child & Adolescent Psychiatry*, *57*(7), 471–480.

Gradus, J. L., Rosellini, A. J., Horváth-Puhó, E., Street, A. E., Galatzer-Levy, I., Jiang, T., ... Sørensen, H. T. (2020, jan). Prediction of Sex-Specific Suicide Risk Using Machine Learning and Single-Payer Health Care Registry Data from Denmark. *JAMA Psychiatry*, *77*(1), 25–34. doi: 10.1001/jamapsychiatry.2019.2905

Grossman, M. (2000). The human capital model. In *Handbook of health economics* (Vol. 1, pp. 347–408). Elsevier.

Insel, T. R. (2014, apr). *The nimh research domain criteria (rdoc) project: Precision medicine for psychiatry* (Vol. 171) (No. 4). American Psychiatric Association. doi: 10.1176/appi.ajp.2014.14020138

Johar, M., & Truong, J. (2014, dec). Direct and indirect effect of depression in adolescence on adult wages. *Appl. Econ.*, *46*(36), 4431–4444. doi: 10.1080/00036846.2014.962227

Kale, M. S., & Korenstein, D. (2018). Overdiagnosis in primary care: framing the problem and finding solutions. *BMJ*, *362*. doi: 10.1136/bmj.k2820

Kapteyn, A., & Ypma, J. Y. (2007). Measurement error and misclassification: A comparison of survey and administrative data. *Journal of Labor Economics*, *25*(3), 513–551.

Kitashima, L., & Chorniy, A. (2017). *Adhd medication effects on primary and secondary students' academic achievement* (Tech. Rep.). Princeton University Working paper.

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, *75*(1), 83–119.

Kodish, I., Rockhill, C., & Varley, C. (2011). Pharmacotherapy for anxiety disorders in children and adolescents. *Dialogues Clin. Neurosci.*, *13*(4), 439–52.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017, aug). The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. *KDD proceedings. Int. Conf. Knowl. Discov. Data Min.*, *2017*, 275–284.

Layton, T. J., Barnett, M. L., Hicks, T. R., & Jena, A. B. (2018, nov). Attention Deficit–Hyperactivity Disorder and Month of School Enrollment. *N. Engl. J. Med.*, *379*(22), 2122–2130.

Link, B. G., & Phelan, J. C. (2010). Labeling and Stigma. In T. L. Scheid & T. N. Brown (Eds.), *A handb. stufy ment. heal. soc. context. theor. syst.* (2nd Editio ed., pp. 571–587). New York: Cambridge University Press.

Maia, C. R. M., Cortese, S., Caye, A., Deakin, T. K., Polanczyk, G. V., Polanczyk, C. A., & Rohde, L. A. P. (2017). Long-term efficacy of methylphenidate immediate-release for the treatment of childhood adhd: a systematic review and meta-analysis. *Journal of attention disorders*, *21*(1), 3–13.

Mitchell, S., & Shepard, M. (1966, nov). A Comparative Study of Children's Behaviour at Home and at School. *Br. J. Educ. Psychol.*, *36*(3), 248–254. doi: 10.1111/j.2044-8279.1966.tb01876.x

Molina, B. S., Hinshaw, S. P., Swanson, J. M., Arnold, L. E., Vitiello, B., Jensen, P. S., ... Houck, P. R. (2009, may). The MTA at 8 Years: Prospective Follow-up of Children Treated for Combined-Type ADHD in a Multisite Study. *J. Am. Acad. Child Adolesc. Psychiatry*, *48*(5), 484–500. doi: 10.1097/CHI.0B013E31819C23D0

Mousteri, V., Daly, M., Delaney, L., Tynelius, P., & Rasmussen, F. (2019, feb). Adolescent mental health and unemployment over the lifespan: Population evidence from Sweden. *Soc. Sci. Med.*, *222*, 305–314. doi: 10.1016/j.socscimed.2018.12.030

Mullainathan, S., & Obermeyer, Z. (2017, may). Does Machine Learning Automate Moral Hazard and Error? *Am. Econ. Rev.*, *107*(5), 476–480.

Mullainathan, S., & Obermeyer, Z. (2019). *A machine learning approach to low-value health care: Wasted tests, missed heart attacks and mis-predictions.*

Nigg, J. T. (2013, mar). Attention-deficit/hyperactivity disorder and adverse health outcomes. *Clin. Psychol. Rev.*, *33*(2), 215–28. doi: 10.1016/j.cpr.2012.11.005

Obermeyer, Z., & Emanuel, E. J. (2016, sep). Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N. Engl. J. Med.*, *375*(13), 1216–9.

Ogundele, M. O. (2018, feb). Behavioural and emotional disorders in childhood: A brief overview for paediatricians. *World J. Clin. Pediatr.*, *7*(1), 9–26. doi: 10.5409/wjcp.v7.i1.9

Owens, J., & Jackson, H. (2017, jan). Attention-deficit/hyperactivity disorder severity, diagnosis, & later academic achievement in a national sample. *Soc. Sci. Res.*, *61*, 251–265. doi: 10.1016/j.ssresearch.2016.06.018

Prinz, D., Chernew, M., Cutler, D., & Frakt, A. (2018, jul). *Health and Economic Activity Over the Lifecycle: Literature Review* (Tech. Rep.). Cambridge, MA: National Bureau of Economic Research. doi: 10.3386/w24865

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

Rambachan, A., Kleinberg, J., Ludwig, J., & Mullainathan, S. (2020). An Economic Perspective on Algorithmic Fairness. *AEA Pap. Proc.*, *110*, 91–95. Retrieved from https://doi.org/10.1257/pandp.20201036 doi: 10.1257/pandp.20201036

Rose, S. (2018, aug). Machine Learning for Prediction in Electronic Health Data. *JAMA Netw. open*, *1*(4), e181404. doi: 10.1001/jamanetworkopen.2018.1404

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

Snowden, L. R. (2003, feb). Bias in mental health assessment and intervention: theory and evidence. *Am. J. Public Health*, *93*(2), 239–43. doi: 10.2105/ajph.93.2.239

Statistics Canada. (2020a). *Canadian community health survey, the daily catalogue no.: 11-001-x* (Tech. Rep.). Ottawa. Retrieved from https://www150.statcan.gc.ca/n1/daily-quotidien/200806/dq200806a-eng.htm?CMP=mstatcan

Statistics Canada. (2020b). *Table 13-10-0392-01 - deaths and age-specific mortality rates, by selected grouped causes* (Tech. Rep.). Ottawa. Retrieved from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310039201

Strickland, J., Hopkins, J., & Keenan, K. (2012, aug). Mother-Teacher Agreement on Preschoolers' Symptoms of ODD and CD: Does Context matter? *J. Abnorm. Child Psychol.*, *40*(6), 933–943. doi: 10.1007/s10802-012-9622-y

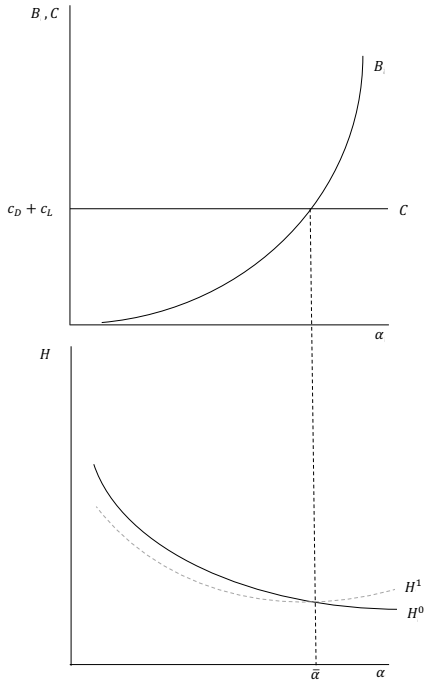
Walkup, J. T. (2017). Antidepressant efficacy for depression in children and adolescents: industry-and nimh-funded studies. *American Journal of Psychiatry*, *174*(5), 430–437.

Wang, P. S., Aguilar-Gaxiola, S., Alonso, J., Angermeyer, M. C., Borges, G., Bromet, E. J., ... others (2007). Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the who world mental health surveys. *The Lancet*, *370*(9590), 841–850.

Williams, D. R., & Mohammed, S. A. (2009). Discrimination and racial disparities in health: evidence and needed research. *J Behav Med*, *32*, 20–47. doi: 10.1007/s10865-008-9185-0

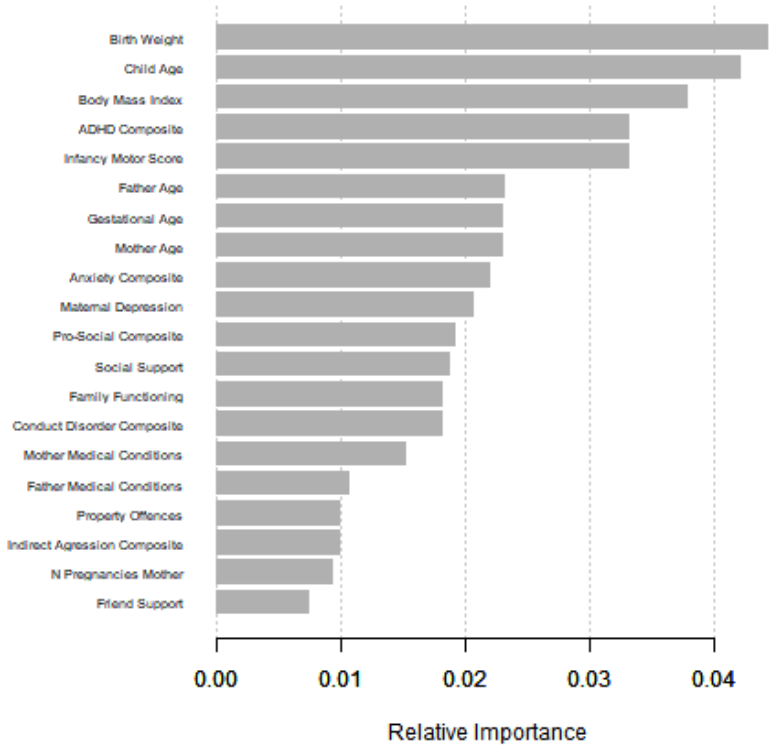
Appendix Figures

Figure A1: Cost-Benefit Diagram of Mental Health Treatment



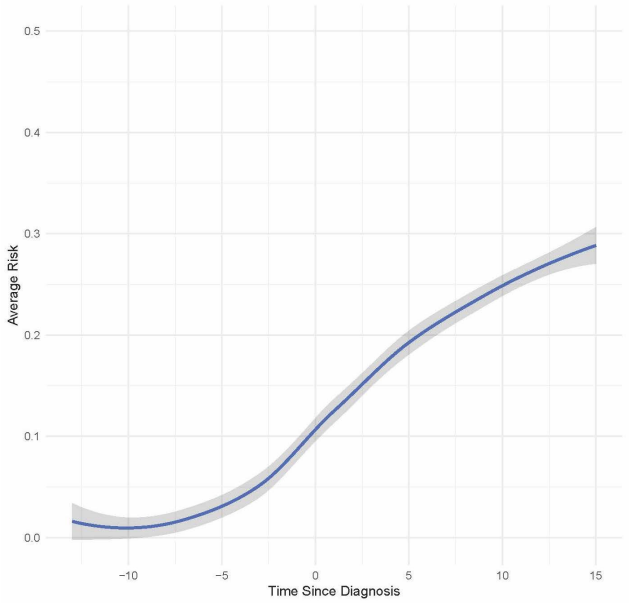
Note: The top panel represents the benefits B and costs C of identification. The bottom panel translates this to human capital outcomes, with a positive increase for identification above some risk threshold $\bar{\alpha}$.

Figure A2: Most Frequently Use Variables for Estimating Health-Based Diagnostic Risk



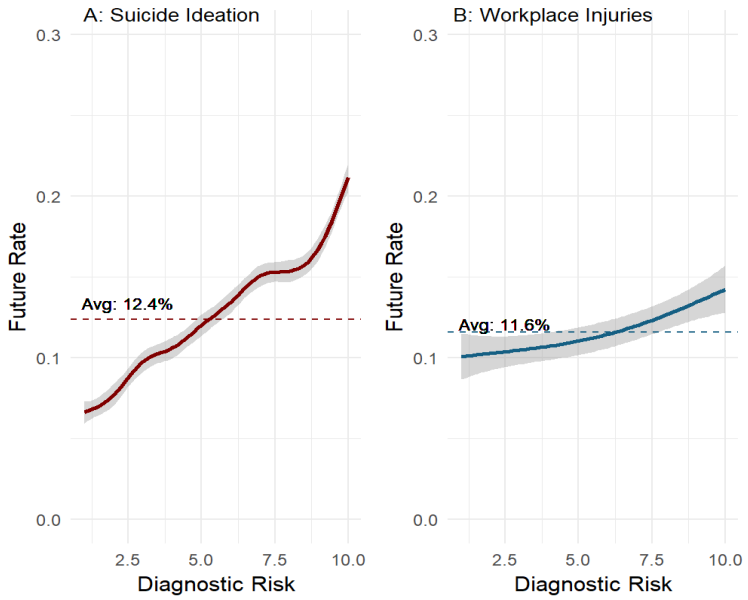
Note: Figure shows the 20 most frequently used variables for predicting health-based diagnostic risk in the gradient boosted tree algorithm. Relative importance measures frequency relative to the most frequently used variable, birth weight.

Figure A3: Average Risk by Time to Diagnosis



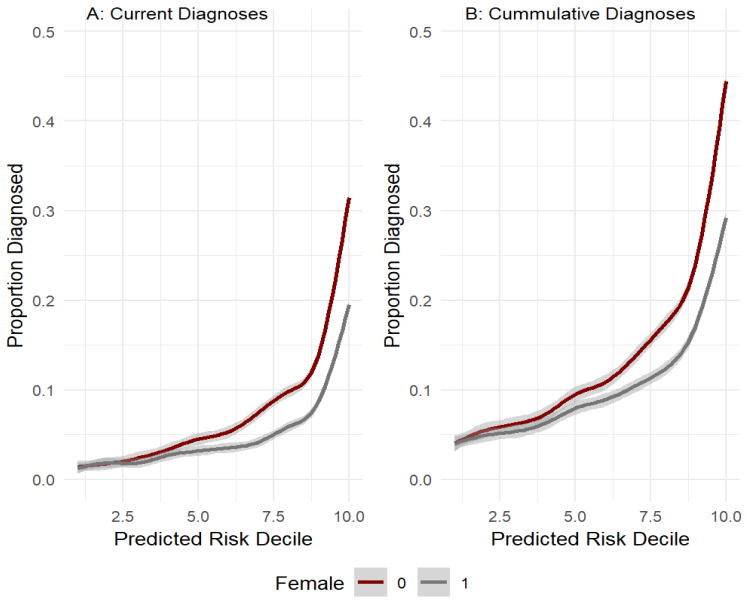
Note: Figure A3 plots the average predicted mental health risk based on time to diagnosis and time since diagnosis (blue). Diagnosis occurs at time zero. Shaded region shows 95 percent confidence interval.

Figure A4: Adverse Health Outcomes in Non-Diagnosed Individuals by Risk



Note: Figure shows the proportion of children and adolescents who are never diagnosed but reporting suicide ideation over the sample period (Panel A) or experienced a workplace injury in adulthood (Panel B), by predicted childhood mental health risk decile in a hold-out test set. Dotted lines are overall averages for each outcome.

Figure A5: Proportion Diagnosed by Age-Standardized Risk Decile, by sex



Note: Figure A5 plots the proportion of individuals diagnosed by age-standardized mental health risk decile, broken down for males (green) and females (red). Risk deciles are based on an algorithm-predicted mental health risk and diagnostic status is based diagnosis at time of risk measurement.

Appendix Tables

Table A1: Current and Subsequent Diagnosis by Risk Decile

Risk	Overall		Males		Females	
Decile	Current	Subsequent	Current	Subsequent	Current	Subsequent
1	0.01	0.04	0.01	0.04	0.01	0.04
2	0.02	0.05	0.02	0.05	0.02	0.05
3	0.02	0.06	0.02	0.06	0.02	0.05
4	0.03	0.07	0.03	0.07	0.03	0.06
5	0.04	0.09	0.05	0.10	0.03	0.08
6	0.04	0.10	0.05	0.11	0.04	0.09
7	0.06	0.12	0.07	0.14	0.04	0.11
8	0.08	0.15	0.19	0.18	0.06	0.012
9	0.11	0.20	0.14	0.24	0.09	0.17
10	0.26	0.37	0.32	0.45	0.20	0.29
			% of Total Diagnoses			
$\leq 5^{th}$ decile	0.18	0.25	0.14	0.22	0.20	0.29
$\geq 9^{th}$ decile	0.55	0.46	0.51	0.48	0.54	0.48

Note: Table A1 presents the proportion of individuals diagnosed by predicted mental health risk decile. Predicted risk is based on gradient boosted decision tree algorithm. Deciles of risk are based on 10 equally sized groups based on age-standardized predicted risk.

Table A2: Machine Predicted Diagnosis vs. Actual Diagnosis

		High Risk			Low Risk		
	Overall	$D_i = 1$	$D_i = 0$	Diff	$D_i = 1$	$D_i = 0$	Diff
Risk Score	0.03	0.73	0.67	0.06***	0.10	0.02	0.08***
Future Diagnosis	0.12	1.00	0.17	0.83***	1.00	0.06	0.94***
ADHD Score	5.78	11.85	11.15	0.70***	8.04	5.55	2.49***
Anx/Depres Score	4.17	7.55	7.05	0.50*	5.81	4.02	1.79***
Child Age	10.49	14.23	14.64	-0.42	14.05	10.21	3.85***
Asthma	0.21	0.48	0.48	0.00	0.34	0.20	0.14***
Female	0.50	0.31	0.32	0.01	0.40	0.51	-0.10***
Log Income	10.87	10.76	10.79	0.03	10.86	10.87	-0.01
Quebec	0.19	0.16	0.20	-0.05*	0.24	0.18	0.06***
Rural	0.23	0.17	0.22	-0.05*	0.20	0.23	-0.03***
Learning Disorder	0.05	0.65	0.28	-0.37***	0.34	0.03	0.31***
P (%)	100	0.97	0.54	-	5.72	92.76	-

Note: **??** presents average values of key covariates by whether a person was predicted to be high risk or low risk, and whether they were diagnosed or not. High risk is determined by having a mental health risk score greater than 0.5. D_i is a diagnostic indicator. All reported numbers are mean values for given classification by subset. Significance of differences are represented as: * p<0.05, **p<0.01, *** p<0.001.

Table A3: Baseline Effect of Mental Health Diagnosis and Treatment on Longer-Term Outcomes

	Direct Health Outcomes				Socioeconomic Outcomes					
	Suicide Ideation		WCB		Education		Low-Income		Social Assistance	
	Panel A: Baseline Treatment Effects									
Diagnosis	0.084***	0.073***	0.024**	0.022*	-0.191***	-0.114***	0.107***	0.072**	0.180***	0.134***
	(0.018)	(0.022)	(0.009)	(0.011)	(0.022)	(0.026)	(0.023)	(0.026)	(0.022)	(0.026)
D*Treatment	0.093***	0.067***	0.002	0.007	-0.266***	-0.189***	0.133***	0.076**	0.232***	0.183***
	(0.016)	(0.019)	(0.005)	(0.008)	(0.019)	(0.024)	(0.019)	(0.024)	(0.018)	(0.024)
Risk		0.534***		0.029		-0.542***		0.470***		0.617***
		(0.062)		(0.021)		(0.062)		(0.064)		(0.066)
D*Risk		-0.476***		0.004		0.179		-0.427***		-0.461***
		(0.114)		(0.053)		(0.117)		(0.121)		(0.119)
D*T*Risk		-0.393***		-0.049^{+}		0.325***		-0.287**		-0.465***
		(0.088)		(0.026)		(0.087)		(0.088)		(0.087)
N	10,400	9,800	11,700	10,900	12,000	11,000	12,000	11,000	12,000	11,000

Note: Table A3 presents the treatment effects of mental health diagnosis and treatment by risk as measured at age 15 on several long term outcomes. Direct health measures include whether the individual ever reported suicidal ideation (from the NLSCY) or made a workers compensation board (WCB) claim, as measured from tax records. Socioeconomic outcomes include whether they ever claimed a education tax credit in adulthood, had income below the low-income cut-off, or received social assistance. Models include only baseline year, child age and province fixed effects. All N are rounded to the nearest 100 for the NLSCY and the nearest 1000 for the tax data set. Standard errors in parentheses and clustered at the neighbourhood level: $^{+}$ $p < 0.10$,* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A4: Mental Health and Short Term Outcomes by Sex

	Injuries		Hospitalization		Marijuana		Pres. Pills		School Welfare	
	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
					Panel A: Short-Term Outcomes					
Diagnosis	0.023	0.064*	0.037*	0.023	-0.022	-0.040	0.024+	-0.031	0.028	0.032*
	(0.023)	(0.027)	(0.017)	(0.017)	(0.052)	(0.052)	(0.014)	(0.020)	(0.018)	(0.015)
D*Treatment	0.011	0.039	0.012	0.048*	-0.101+	0.251**	0.015+	0.042**	0.049**	0.060**
	(0.021)	(0.029)	(0.015)	(0.023)	(0.057)	(0.083)	(0.009)	(0.016)	(0.018)	(0.022)
Diagnosis*Risk	0.094	-0.144	-0.028	0.003	-0.062	-0.047	-0.064	0.122	0.041	-0.360**
	(0.083)	(0.124)	(0.056)	(0.080)	(0.173)	(0.222)	(0.041)	(0.137)	(0.069)	(0.119)
D*T*Risk	-0.112*	-0.087	0.009	-0.015	-0.037	-0.376	-0.052	-0.040	-0.089	-0.112
	(0.056)	(0.087)	(0.040)	(0.057)	(0.154)	(0.258)	(0.032)	(0.067)	(0.058)	(0.075)
Risk	0.066+	0.121*	0.058*	0.059*	0.048	-0.014	0.017	0.115**	0.051+	0.091*
	(0.035)	(0.049)	(0.023)	(0.028)	(0.084)	(0.107)	(0.015)	(0.039)	(0.028)	(0.036)
N	36,000		36,000		10,000		29,000		36,000	
					Panel B: Long-Term Outcomes					
	Injuries		Suicide Ideation		Education		Low-Income		Social Assistance	
Diagnosis	0.141*	0.055	0.050+	0.035	-0.839***	-0.554**	0.552*	-0.007	0.058	0.074
	(0.064)	(0.059)	(0.029)	(0.043)	(0.207)	(0.196)	(0.228)	(0.235)	(0.040)	(0.046)
D*Treatment	0.124**	0.124+	0.004	0.130**	-0.705***	-0.681**	0.476*	0.751**	0.105**	0.127**
	(0.043)	(0.070)	(0.022)	(0.046)	(0.136)	(0.231)	(0.187)	(0.241)	(0.035)	(0.047)
Diagnosis*Risk	-0.500+	-0.232	-0.341*	0.180	2.308**	2.479+	-1.173	1.289	-0.076	-0.385
	(0.261)	(0.383)	(0.155)	(0.292)	(0.793)	(1.317)	(0.947)	(1.727)	(0.177)	(0.273)
D*T*Risk	-0.499**	-0.404+	-0.180+	-0.348+	1.587**	1.346+	-0.344	-2.019*	-0.328**	-0.297+
	(0.177)	(0.220)	(0.108)	(0.192)	(0.537)	(0.782)	(0.670)	(1.009)	(0.119)	(0.179)
Risk	0.278*	0.250+	0.386***	0.541***	-1.890***	-2.360***	1.356***	2.894***	0.334***	0.354***
	(0.132)	(0.144)	(0.079)	(0.109)	(0.396)	(0.513)	(0.407)	(0.709)	(0.081)	(0.096)
N	5,000		5,000		5,000		5,000		5,000	

Note: Table A4 presents mental health diagnosis and treatment effects by sex. All models include a full set of covariates and neighbourhood fixed effects. All N are rounded to the nearest 1000. Standard errors in parentheses and clustered at the neighbourhood level: $^{+}$ $p < 0.10$,* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A5: Effect of Early or Late Diagnosis on Long-term Income

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
			Overall Sample			Males	Females
Early	-0.251**	-0.195*	-0.116	-0.153^{+}	-0.161*	-0.158	-0.162
	(0.080)	(0.085)	(0.078)	(0.083)	(0.071)	(0.108)	(0.108)
Late	-0.321***	-0.361***	-0.339***	-0.305***	-0.177**	-0.161^{+}	-0.189*
	(0.061)	(0.076)	(0.075)	(0.077)	(0.066)	(0.093)	(0.094)
Early*Risk		0.049	-0.171	-0.263	0.440^{+}	0.442	0.423
		(0.339)	(0.363)	(0.433)	(0.236)	(0.328)	(0.496)
Late*Risk		0.603^{+}	0.550^{+}	0.444	0.199	0.280	-0.005
		(0.334)	(0.327)	(0.357)	(0.365)	(0.505)	(0.581)
Risk		-0.456*	-0.246	-0.252	-0.436**	-0.534*	-0.327
		(0.193)	(0.196)	(0.223)	(0.158)	(0.227)	(0.264)
Health Endowments		x	x	x	x	x	x
Controls		x	x	x	x	x	x
Neighbourhood FE				x	x	x	x
LFP and Employment					x	x	x
N	7,000	7,000	7,000	7,000	5,000	2,000	3,000
Adj. R^2	0.138	0.139	0.166	0.155	0.376	0.371	0.349

Note: Table presents the treatment effects of mental health diagnosis by whether a child was diagnosed before or after the median age of diagnosis (11). Baseline models include only year, child age and province fixed effects. All N are rounded to the nearest 1000 in accordance with CRDCN rules for this data set. Standard errors in parentheses and clustered at the neighbourhood level: $^{+}$ $p < 0.10$,* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A6: Diagnosis and Treatment Effects by Specific Diagnosis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Suicide Ideation		Workplace Injury		Social Assistance		Higher Education		Low Income	
	ADHD	EPND	ADHD	EPND	ADHD	EPND	ADHD	EPND	ADHD	EPND
Diagnosis	-0.031	0.064**	0.009	0.072	0.089*	0.036	-0.701***	-0.350**	0.404+	0.212
	(0.037)	(0.024)	(0.073)	(0.044)	(0.039)	(0.028)	(0.197)	(0.136)	(0.231)	(0.152)
D*Treatment	0.063	0.015	0.089	0.069+	0.031	0.092***	-0.102	-0.563***	0.076	0.336*
	(0.040)	(0.023)	(0.075)	(0.038)	(0.042)	(0.028)	(0.197)	(0.115)	(0.242)	(0.151)
Risk	0.448***	0.424***	0.220*	0.251**	0.281***	0.299***	-1.816***	-1.727***	1.614***	1.666***
	(0.056)	(0.054)	(0.091)	(0.093)	(0.058)	(0.057)	(0.273)	(0.273)	(0.335)	(0.337)
Diagnosis*Risk	-0.279	-0.120	-0.072	-0.292*	-0.065	-0.108	2.629***	0.837+	-0.225	-0.483
	(0.183)	(0.099)	(0.226)	(0.127)	(0.159)	(0.105)	(0.529)	(0.472)	(1.020)	(0.581)
D*T*Risk	0.020	-0.168+	-0.349	-0.288*	-0.215	-0.224*	-0.940+	0.917*	-0.416	-0.392
	(0.182)	(0.097)	(0.228)	(0.123)	(0.158)	(0.103)	(0.526)	(0.429)	(1.084)	(0.612)
N	9,000	9,000	10,000	10,000	10,000	10,000	10,000	10,000	10,000	10,000
Adj. R-sq	0.071	0.071	0.058	0.058	0.164	0.163	0.298	0.298	0.338	0.338

Note: Table presents the treatment effects of mental health diagnosis by whether a child was diagnosed with ADHD or any emotional, psychological or nervous disorder (EPND). All models are fully specified to include year, child age and province fixed effects, health endowments, family environment, adulthood controls, and neighbourhood fixed effects. All N are rounded to the nearest 1000 in accordance with CRDCN rules for this data set. Standard errors in parentheses and clustered at the neighbourhood level: $^{+}$ $p < 0.10$,* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix A: Blinded Covariate Set for Mental Risk Algorithm

The full set of covariates in the blinded risk model include the child's precise age in months, birth weight and gestational age, other physical disorders, self-reported health, injuries, the individual components of the 50-point CBS scale, as well as specific disorder composite scales, social ratings and current body mass index (BMI). I also include the following information on parent health: number of chronic conditions, depression scales, birth history, including the number of pregnancies and live births, presence of post-partum depression, and parental smoking and drinking frequency. The full list is detailed below.

Demographics
Child Age in months

Diagnoses and Medications
Diagnosed Diabetes; Diagnosed Allergy; Diagnosed Heart Disease; Diagnosed Cerebral Palsy;Diagnosed Kidney Disease; Diagnosed Mental Handicap; Diagnosed Asthma; Diagnosed Other Long Term Condition; Child uses inhaler; Child uses other medications

General Health
Self-reported Health; Parent rating of child health; BMI

Feelings and Behaviour
Child Depression Scale; Self-Image Scale; Friendship Scale Family functioning score (Standardized); Social Support Scale (Standardized); Mood Scale (adolescents) (Standardized); ADHD Composite Score; Anxiety and Depression Composite Score; Conduct Disorder Composite Score; Prosocial Behaviours Composite Score; Property Offenses Composite Score; Indirect Aggression Composite Score

Parent Health
Person most knowledgeable age (mostly mother) ; Parent 1 Gender Indicator; Parent 1 Self-reported health ; Parent 1 Smoking Parent 1 Drinking; Parent 1 Depression score (Standardized) Number of Conditions Parent or Guardian 1; Parent 2 Age (mostly father); Parent 2 Gender Indicator; Parent 2 Self-reported health Parent 2 Smoking; Parent 2 Drinking; Number of Conditions Parent or Guardian 2 ; Mother Experienced Post-partum depression; Mother number of pregnancies (Standardized); Mother number of children (Standardized)

Schooling
Grade; Child changed schools in last year; Contacted by school in last year; Child's feelings about school; Child Attended Juniour Kindergarten; Child Attended Kindergarten

Other Life Events
Child moved in last year; Child experiences and adverse childhood event in past year (divorce, death, abuse, other traumatic event)

Early-Related Health
Birth weight (Standardized); Gestational Age (Standardized) Motor and Social Development (0-47 months

of age) (Standardized)

Children's Behavioural Scale

How often would you say that child:

Clings to adults or is too dependent?; Constantly seeks help?; Doesn't want to sleep alone?; Gets too upset when separated from parents?; Defiant; Doesn't seem to feel guilty after misbehaving?; Punishment doesn't change his/her behaviour?; Has temper tantrums or hot temper?; Has angry moods?; Gives up easily; Miserable, unhappy, tearful or distressed?; Stares into space?; Fidgets?; Is disobedient at school?; Tends to do things on their own - is rather solitary?; Shows sympathy to someone who has made a mistake?; Can-t sit still, is restless or hyperactive?; Destroys his/her own things?; Will try to help someone who has been hurt?; Steals at home?; Seems to be unhappy, sad or depressed?; Gets into many fights?; Volunteers to help clear up a mess someone else has made?; Is distractible, has trouble sticking to any activity?; When mad at someone, tries to get others to dislike that person?; Is not as happy as other children?; Destroys things belonging to his/her family, or other children?; If there is a quarrel or dispute, will try to stop it?; Can-t concentrate, can-t pay attention for long?; Is too fearful or anxious?; When mad at someone, becomes friends with another as revenge?; Is impulsive, acts without thinking?; Tells lies or cheats?; Offers to help other children (friend, brother or sister) who are having difficulty with a task?; Is worried?; Has difficulty awaiting turn in games or groups?; When another child accidentally hurts him/her (such as by bumping into him/her), assumes that the other child meant to do it, and then reacts with anger and fighting?; When mad at someone, says bad things behind the other-s back?

Child self-report:

I physically attack people; I comfort others who are hurt; I vandalize; I threaten people; I help pick up things; I cannot settle to anything; I bully or am mean to others; When I am mad at someone, I say to others: let-s not be with him/her; I am nervous, high-strung or tense; I kick or hit other people my age; I invite bystanders to join in the game; I steal outside my home; I am inattentive; I have trouble enjoying myself; I help others my age who are feeling sick; I tell secrets to a third person; I encourage other people my age

Appendix B: Modeling the Blinded Gradient Boosted Tree

For fitting the gradient boosted tree (GBT) algorithm, I deliberately leave out variables indicating sex, ethnicity, immigrant status, and household income or any welfare receipt, along with province, city and location markers. I remove information on annual medical visits that may propagate medical access bias. I also remove indicators for learning disorders given its high co-morbidity with ADHD and evidence that learning disorders may be purposefully diagnosed in conjunction with some mental health disorders to render a child eligible for specific school resources. Other removed variables include parent work variables, family structure, housing type, parent education, religion and language first spoken. I further exclude the year to remove any upward time trend in the risk assessment. This leaves 109 variables, detailed in Appendix A. To improve prediction, I normalize all continuous variables and factorize any categorical variables, including indicators for missing values by missingness type. The NLSCY includes four missingness types: valid skip, refusal to answer, do not know, or not stated. This process is known as "one-hot encoding". With one-hot encoding I can maintain more observations for the prediction training and test set and glean information from the variable missingness itself. Upon factorization, there are 689 total variables.

For the final gradient boosted decision tree model, through cross-validated training over five randomly sampled folds and hyper-grid searching for optimal parameters, I find an optimal learning rate of 0.05; a maximum allowed tree depth of 6 splits; and a minimum number of observations in a terminal node of 1. I use a sampling rate of 0.90, where only 90 percent of the observations are selected for each iterative run. Finally, I use an 100 percent bagging rule, where all features are used in each iteration. The final model is fit using 640 trees.

To compare the value of using machine learning compared to more traditional approaches, I estimate parallel predictions using a standard linear probability model and a logistic regression model. I estimate the linear probability and logistic regression models on a restricted set of health endowment and current health variables that are in the top 20 variables within the gradient boosted decision tree algorithm. These are the composite scores for ADHD, anxiety and depression, pro-social behaviours, and conduct disorder. For health endowments, I include birth weight, gestational age, infancy motor scores, father and mother age. For family health factors I include, number of chronic conditions reported by the mother, maternal depression score, a social support score, and a family functioning score. I additionally include controls for child BMI, age and sex. I estimate each model on a complete sample of 71,100 observations.

I find linear probability models (LPM) are particularly unhelpful in predicting mental health incidence. In the estimated model, if using a 0.5 threshold to indicate diagnosis, they predict a mental health disorder rate of zero. Logistic methods are notably similar. Training a logit model with cross-validation only marginally improves the disorder prediction rate to 1.40 percent. The area under the curve (AUC), a standard measure of model performance, is exceedingly high using linear probably (0.83) or logistic regression (0.84) because mental health disorders in childhood are rare (the AUC for the gradient boosted tree model is 0.83). The tuned and blinded gradient boosted tree algorithm performs better than the logit model, with a predicted diagnosis rate of 1.43 percent if using a strict 0.5 threshold. In Table A7, I also compare results of predicted high risk using the mean predicted risk of those in the top decile of risk (0.25 for the baseline GBT). The baseline GBT also outperforms the LPM and logit using this more relaxed threshold approach. The GBT prediction thus remains the preferred specification, firstly because of improvements in prediction capability, particularly for those in the top decile of risk. Secondly, computationally this method allows for the inclusion of many variables that may offer new insights into mental health pathways. Finally, because of the generally messy nature of survey data, which is prone to missingness and attrition, the use of gradient boosted trees

allows for a high degree of data retention. This is especially important given the sample size available to estimate long-term income effects through the matched tax file (T1FF) data set is relatively small, with just over 13,000 individuals.

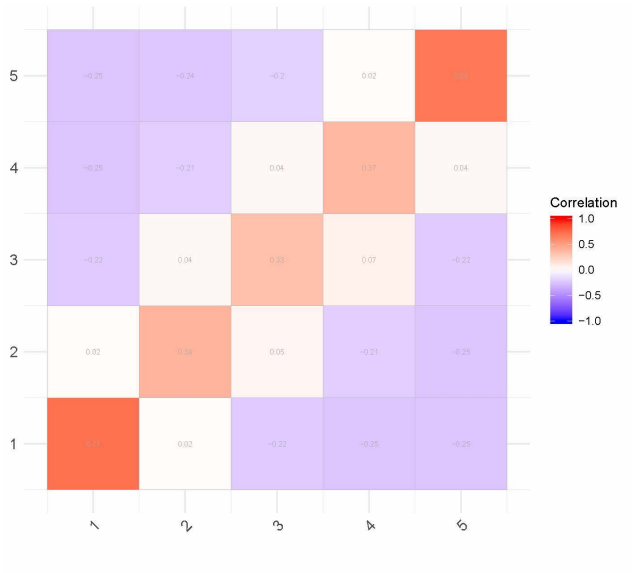
Table A7: Performance parameters across three model specifications

Outcome	Current Diagnosis					Suicide
Model	LPM	Logit	GBT Baseline	GBT Male	GBT Female	GBT
Mean Squared Error	0.046	0.043	0.067	0.076	0.068	0.129
Area under the curve	0.830	0.839	0.830	0.822	0.811	0.724
Average Predicted Risk	0.055	0.055	0.045	0.045	0.023	0.053
Males	0.057	0.057	0.054	0.052	0.027	0.059
Females	0.053	0.053	0.035	0.038	0.018	0.048
Top Decile Future Diagnoses (%)	0.352	0.326	0.390	0.364	0.338	0.280
Males	0.387	0.358	0.414	0.450	0.400	0.278
Females	0.290	0.273	0.338	0.280	0.270	0.284
Predicted High Risk (%)	0.023	0.039	0.078	0.048	0.021	0.078
Male	0.033	0.051	0.083	0.058	0.026	0.059
Female	0.014	0.027	0.048	0.038	0.015	0.097

Note: Table presents model fit comparisons across a linear probability model (LPM), logit model, and gradient boosted tree (GBT) models using wither current diagnosis as the dependent variable of interest or suicide ideation. Two additional GBT models are fit using only males or females from the randomly selected training sample, but fitting test-sample prediction of the full set of males and females. The suicide GBT model uses a truncated sample of individuals with later adolescent suicide information available.

I additionally build a "non-blinded" model, incorporating all demographic, socioeconomic and location factors that might lead to selection into treatment. Comparing this to the baseline model, the full model has a higher prediction accuracy than the blinded model, with a diagnostic rate of 2.3 percent above a 0.5 threshold versus 1.43 percent. However, the full model increased the base rate identification error between males and females by inflating risk for males and deflating risk for females. The average risk decile for females is 5.18 in the full model but increases to 5.25 in the blinded model. Male risk is 5.82 in the full model versus 5.75 in the blinded model. I generate risk quintiles for both the blinded and full prediction models and compare how much each model overlaps in Appendix Figure A6. This figure shows substantial overlap in both the bottom and top quintiles of risk, representing the lowest risk and highest risk individuals. Both models thus differentiate high and low-risk individuals accurately. The inclusion of biasing factors in the full model only serves to increase risk assignment in the center of the risk distribution. Variables relevant to the non-blinded prediction included the presence of a learning disorder and annual physician visits. Both these variables might indicate one's ability to access medical care more than a true underlying health risk. Thus, I conclude that the full model biases risk status for children and adolescents at the margin of diagnosis, but by factors seemingly unrelated to the underlying disorder. In blinding the algorithm, the base rate error in using identification as an outcome variable can be improved while not throwing out information useful for predicting health-specific identification risk.

Figure A6: Blinded versus Naive Model Overlap



Note: Predicted risk for blinded (x-axis) and non-blinded (y-axis) risk predictions were divided into quintiles of risk ranging from one (low-risk) to five (high risk). Correlation coefficients shown in each square represent the level of overlap for those given quintiles of risk. High levels of correlation correspond to higher overlap.