The Role of Individual Health Patterns and Endogenous Employment on Absenteeism

Kristin Vrona

May 2025

Abstract

This paper examines the question often present in behavioral models of health - the bias that can persist across estimates when unobserved factors drive behavior. In the context of this study, individuals may select into employment non-randomly. For instance, ability to balance health issues while being productive. When these possibilities are ignored, the magnitude of the effect that mental illness has on absenteeism is likely underestimated. A classic Heckman selection model, supported by strong exclusion restrictions, suggests the presence of endogenous sample selection that is validated by a semi-parametric copula-based selection model that accommodates the discrete and overdispersed nature of absence data.

1 Introduction

Mental illness is one of the most prevalent and costly health conditions in the United States, affecting 1 in 5 adults annually. While mental health issues entail direct treatment costs, the substantial economic burden primarily stems from indirect costs, such as reduced productivity and heightened rates of disability. Depression and anxiety disorders alone contribute to an estimated $1 trillion in lost productivity worldwide each year. Despite the importance of early and comprehensive treatment, U.S. adults report an average delay of 11 years between symptom onset and treatment. This gap has severe implications, as untreated symptoms lead to higher rates of substance use, depression, anxiety, and suicide.

Untreated mental health induces substantial economic burdens, such as reduced productivity and heightened rates of disability. Depression and anxiety disorders alone contribute to an estimated $1 trillion in lost productivity worldwide each year (World Health Organization [WHO], 2024). The gap in prevalence rates and treatment access in the United States has severe implications, as untreated symptoms lead to higher rates of substance use, bankruptcy, homelessness, and suicide (NIMH, 2024).

The economic impact of untreated mental illness underscores the need for a shift in focus. Rather than solely viewing mental health care as a rising expense, employers could benefit from recognizing it as a critical investment in workforce productivity. Evidence suggests that access to mental health services positively impacts absenteeism and productivity outcomes (Cseh, 2008; Fletcher, 2013; Ashwood et al., 2016). Employers who implement robust wellness programs and promote timely access to mental healthcare may see improved productivity and reduced long-term costs associated with employee turnover and absenteeism. By fostering a supportive workplace environment and offering comprehensive mental health benefits, employers can not only mitigate the economic toll of mental illness but also contribute to a healthier, more resilient workforce.

A final note, the COVID-19 pandemic has highlighted the misalignment of employer and laborer incentives. Higher rates of job vacancies may be misrepresented due to the increase in under-employment across the US observed during the business recovery period of the COVID-19. A relatively stable portion of the population is underemployed, a practice that may have been necessary during the pandemic, but has persisted through the recovery period, becoming common practice for employers.[1] Some underemployment offers may allow for the reduction of benefits such as cost sharing for insurance. While my dissertation focuses on pre-COVID data, pooled across years 2010 - 2014, future research directions should control for this pandemic.

In this dissertation, I examine the role of mental health on productive time loss in the form of absenteeism to inform the often ignored benefits that come from promotion of worker wellness.

The remainder of this paper is organized as follows: section 2 lays out the theoretical approach; section 3 describes the data source and variables; section 4 describes econometric implementation; section 5 provides results; section 6 concludes.

In this dissertation, I examine the role of mental health on productive time loss in the form of absenteeism to inform the often ignored benefits that come from promotion of worker wellness.

To address potential endogeneity between mental health and labor supply, the study employs several empirical strategies, including time-invariant measures of exogenous diagnosis, a correlated random effects approach to account for unobserved individual-specific confounders, and semiparametric estimators. Additionally, a standard Heckman estimator is used to assess endogenous unobserved sample selection, with employment probability estimated in the first stage for both employed and unemployed individuals.

2 Theoretical Approach

The theoretical approach described in what follows applies the Becker (1965) theory of time allocation and household production and extends the Grossman (1972) health capital model. Time allocation models of labor supply consider the decision making process associated with market goods as well as time, with time modeled as a scarce resource. Individuals in a household are modeled as producers that take intermediate inputs such as time and market goods to produce useful commodities that are then consumed by the household. Without loss of generality, consider final products c_k for $k = 1, .., K$. An individual faces the following utility function:

$$U = u(c_1, c_2, ..., c_K), \text{ with } \frac{\partial U}{\partial c_k} > 0 \ \forall k. \tag{1}$$

Each final good c_k can be defined by its respective production process that turns market goods and time into final units of consumption. For final good c_k, this process is characterized by

$$c_k = f_k(T_k, x_k; e_k), \tag{2}$$

where T_k is a vector of time inputs allocated to the production of final good c_k, x_k is a vector of market goods used in production, and e_k represents the efficiency of the process, characterized by exogenous factors such as one's age or education. In each case, the efficiency factor, e_k, impacts

[1] US Bureau of Labor Statistics. (n.d.). Charts related to the latest "The employment situation" news release: more chart packages. US Bureau of Labor Statistics. Retrieved January 13, 2024, from https://www.bls.gov/charts/employment-situation/alternative-measures-of-labor-underutilization.htm.

how much time and how many market goods are required to achieve a certain level of a final good, and may or may not be equivalent across production processes. Given each production process, f_k, utility function (1) can be rewritten as

$$U = u(f_1, ..., f_K) = u(x_1, ..., x_K; T_1, ..., T_K). \tag{3}$$

The separability between time and market inputs exhibited by (3) demonstrates that household production is bounded from above due to the scarcity of income and time resources.

Of particular interest in this paper is the process of producing health which acts as the final good to be "consumed". I define the production process resulting in health stock, H, in the following manner:

$$H = f_h(T_h, x_h; MH, PH, e_h), \tag{4}$$

where T_h is a choice of time allotted to the development of health, such as time spent exercising and utilizing healthcare services, x_h is a vector of market goods utilized in health production, such as vitamins and supplements or health insurance coverage, MH and PH are innate endowments of mental and physical health, respectively, which characterize an individual's health productive capacity, and e_h is a vector of non-health-related factors driving the efficiency of the production process, such as education and age. Equation (4) is a concave function exhibiting diminishing returns to factor inputs. Function (4) innovates the health production function proposed by Grossman (1972). In this set up, the baseline endowment of health (MH and PH prior to the commencement of an arbitrary time period) characterizes the feasible set of attainable levels of health given available time and market inputs and preferences over how these inputs are allocated.

Define C as a conglomerate final good that nests each of the production functions of final goods besides health, $c_k, ..., c_{K-1}$, within it so that one can write

$$C = f_c(T_c, x_c, e_c), \tag{5}$$

where $T_c = \sum_{k=1}^{K-1} T_k$ and $x_c = \sum_{k=1}^{K-1} x_k$. Then, equation (1) can be rewritten in the following manner:

$$U = u(C, H) \equiv u(f_c, f_h) \equiv u(x_c, x_h; T_c, T_h), \tag{6}$$

so that the level of utility realized depends on choices of market inputs and time allocations. It is assumed that an individual maximizes their utility subject to the time constraint,

$$T = T_c + T_h + N, \tag{7}$$

where T is total time available and N is time allocated to occupational work, as well as a budget constraint,

$$I = wN + V = p_c x_c + p_h x_h, \tag{8}$$

where I is total income, V is non-earned income such as monetary gifts or inheritances, w is the market wage rate and p_i for $i = \{c, h\}$ are vectors of input prices corresponding to the market goods utilized in the production processes (5) and (4). The time constraint can be substituted into (8) to yield a single "full income" constraint,

$$\begin{aligned} w(T - T_c - T_h) + V &= p_c x_c + p_h x_h \\ \implies wT + V &= p_c x_c + p_h x_h + wT_c + wT_h. \end{aligned} \tag{9}$$

The left-hand side of the second line in (9) represents full income – the income received when an individual chooses to allot all available time to occupational labor.

The results of the current utility maximization problem are more intuitive if the production functions (5) and (4) are redefined, noting that $T_c \equiv t_c C$, $x_c \equiv b_c C$, $T_h \equiv t_h H$, $x_h \equiv b_h H$, where t_i and b_i for $i = \{c, h\}$ are vectors of the input time per unit and market goods per unit required to produce levels of final goods C and H, respectively[2]. With this, a single resource constraint can be expressed as

$$(p_c b_c + w t_c)C + (p_h b_h + w t_h)H = wT + V, \tag{10}$$

where the full price of each unit of the final good, C and H, is the sum of both the direct costs (prices of market goods) and indirect costs (time away from work) associated with each unit produced (Becker, 1965, 6). Now the individual maximizes utility by choosing optimal levels of b_i and t_i for $i = \{c, h\}$.

The optimizing individual will allot additional units of expenditure and time up to the point at which the marginal utility resulting from an additional unit of the respective input equals zero; this is equivalent to saying that available resources will continue to be allotted to production processes until the marginal product of the input is zero. It should be noted that choices of market good inputs and time inputs are not independent. A condition of utility maximization is that the marginal rate of substitution (MRS) between these types of inputs be equal to the ratio of per-unit input costs.

2.1 Labor Supply

At this point of the analysis, it is assumed that an individual has enough information to maximize their utility by choosing optimal bundle $\{C^*, H^*\}$. Given this decision, more information on an individual's preferences over home production (and thus, preferences over labor) can be revealed. Decisions on labor supply in the current set up can be intuitively illustrated by examining the demand for "forgone income". Define the right-hand side of (10) S, which is thus full income that would arise if allotting all available time to work. The demand for forgone income $L(C^*, H^*)$, is then

$$L(C^*, H^*) = S - I(C^*, H^*), \tag{11}$$

where I is an individual's observed income. $L(C^*, H^*)$ can be thought of as the indirect cost of utility-seeking; it is the potential earnings lost when an individual allots positive units of time to the production of health and the conglomerate consumption good.

The current paper focuses on labor supply among the employed, exhibited by work absence, so that I henceforth focus on the special case of individuals with a positive optimal level of labor supply at baseline. Equation (11) can be further defined as

$$L^* = w(T_c^* + T_h^*), \tag{12}$$

for short run fixed wage, w, which is greater than or equal to the individual's reservation wage that optimizes their utility in any given period.

$I(C^*, H^*)$ from (11) can be expressed as

$$I = b_c p_c C + b_h p_h H. \tag{13}$$

[2]Note that t_h and/or b_h are decreasing in measures of the efficiency of the health production process, MH, PH, and e_h. t_c and/or b_c are decreasing in efficiency factor e_c.

Plugging $T_c + T_h = T - N$ into (12) and then plugging it and (13) into (11) yields

$$w(T-N) = S - b_c p_c C - b_h p_h H, \tag{14}$$

which can be rearranged and simplified as follows:

$$\begin{aligned} -wN &= V - b_c p_c C - b_h p_h H \\ \implies N &= \frac{b_c p_c C + b_h p_h H - V}{w}. \end{aligned} \tag{15}$$

We now have hours of work expressed as a function of health status. Taking the partial derivative with respect to health, we have

$$\frac{\partial N}{\partial H} = \frac{b_h p_h}{w} > 0. \tag{16}$$

Equation (16) tells us that better health induces more short-term labor supply. In (16), $b_h p_h$ represents the marginal cost of producing an additional unit of health using market goods and w is the per-unit time cost associated with the level of time allotted to producing health rather than working.

The framework thus far provides an intuitive interpretation that individuals with lower health endowments exhibit more variation in labor supply in response to a health shock.

2.2 Absenteeism

At the commencement of a job, a worker and employer reach a contracted agreement. In the process of reaching this agreement, an employer determines the optimal choice of labor hours per period based on the market wage and the minimum expectations regarding the job. The market wage is assumed to be exogenous and will be equal to the marginal revenue product in equilibrium (Hamermesh, 1993).

A job has minimum expectations for employees, characterized by a predetermined output agenda. Without loss of generality, assume this agenda is characterized by a minimum number of output units produced each period, Y. The minimum number of contracted hours, N_{min}, is increasing in Y. The contract additionally specifies any non-wage compensations available to matched workers, such as health insurance and sick leave.

Without loss of generality, in terms of a time dimension, define D^* as the total labor hours demanded over the span of one year and N^* as the total labor hours supplied over one year. Prior to the start of the year, an employer and employee must come to an implicit agreement on employment. This condition is met when an employee's evaluation of the optimal labor supply at the point prior to the start of the period, N^0, is equal to N_{min}, the minimum number of optimal hours demanded by a potential employer. The base supply, N^0, is a function of the optimal level of health production at the time of evaluation, H^0, and consumption, C^0.

In this dissertation, absenteeism is defined as an absence from work due to physical or mental illness. Work-related absences occur when contractual constraints induce discrepancies between optimal labor demand and optimal labor supply. This can be illustrated in the following manner:

$$\begin{aligned} A &= (D^* - N^*) > 0 \\ &\text{for } D^* > N^*, \end{aligned} \tag{17}$$

where A represents absenteeism which is characterized by a decision making process that occurs

after observing a health shock. The discrepancy between D^* and N^* occur in the short run, as an individual is able to observe their daily health status and reevaluate their optimal labor supply accordingly, while contractual agreements specifying wage, fringe benefits, and minimum time on the job are typically only revisited after a predetermined amount of time.

Substituting functions for D^* and N^* into (17) yields an absenteeism function that is decreasing in health,

$$\begin{aligned} A &= [d(w, N_{min}, B) - n(C^*, H^*)] \\ &= [d(w, N_{min}, B) - n(C, f_h(T_h^*, x_h^*; MH, PH, e_h))]. \end{aligned} \tag{18}$$

Equation (11) results in the hypothesis that absenteeism is inversely related to the endowments of mental and physical health:

$$\frac{\partial A}{\partial MH} < 0, \frac{\partial A}{\partial PH} < 0.$$

In general, illness-related work absenteeism represents the substitutability between time spent earning wages and time spent on improving or maintaining health so that factors which make time inputs in health production more attractive relative to market inputs will exacerbate absenteeism while factors that make market inputs relatively more attractive, such as productivity pay will mitigate absenteeism.

In what follows, I form testable hypotheses derived from this model for certain categories of fringe benefits that workers may receive.

3 Data

The main source of data is provided by public use files of the Medical Expenditure Panel Survey (MEPS) which provides information on demographic and employment characteristics, healthcare utilization, and measures of health and well-being at the individual level. The MEPS consists of several data files. I utilize the Full-Year Consolidated Datafile (FYCD), the Medical Conditions File (MCF), and the Jobs File (JF). I also collect data from the Bureau of Labor Statistics (BLS) on historical unemployment rates per month and across regions. In what follows I describe the data generating process (DGP).

The MEPS follows a one-stage cluster random sampling design. The DGP starts with the random selection of households from the participating household of the most recent National Health Interview Survey (NHIS). The MEPS then collects information on each individual within the selected household units. Each panel of the MEPS consists of five rounds spanning across two consecutive years.

Upon consolidating the data sources, some persons are observed once for a single calendar year while other individuals are observed twice – once for each calendar year they participated in the survey. It should be noted that the reference period for the third interview round of the survey spans across the two consecutive calendar years; fortunately, the data sources used in the analysis ascribe round three data to the appropriate calendar year so that this does not cause a problem. Responses to most of the MEPS interview questions are reported separately for each of the three rounds per year (with the first year's round three variables pertaining to the start of round three up until the end of the calendar year and the next year's round three variables reference the time spent in round three after the start of the new year) so that FYCD annual data files contain three variables per interview prompt.

I use several procedures to annualize variables appearing in groups of three. The MEPS collects

information on each individual in a surveyed household. While late entry of *households* into the survey is not permitted, *individual* participants may enter the survey late if they enter a participating household during the survey period. For example, late entry may be observed for a newly married individual that moves into the residence of his or her spouse, who is a current MEPS participant. A binary variable indicating whether an individual moved into a participating household during the survey period is utilized as a control variable in the analysis.

Aside from the possibility of the late entry of individual participants, the length of each reference period round may vary across individuals due to extenuating circumstances that interfere with an individual's availability on the originally scheduled interview date for a particular round. In such a case, the interview may be rescheduled to occur at an earlier or later date; in the former instance, an individual may exhibit fewer reference period days than the average participant and in the latter, may exhibit a greater than average number of days included in the particular survey round. Information on individual-specific reference period start and end dates for each of the three rounds occurring in a given year is utilized to generate control variables that account for heterogeneous exposure to survey prompts.

Only observations reporting employment at one or more of the three interview dates are retained. Some observations indicate employment as of a particular interview date but list a job start date equal to that interview date. This situation may occur when an individual has been hired but has not yet started working at the time of the interview, meaning they did not work during the corresponding reference period.

These observations are omitted from the sample of actively employed individuals if unemployment is reported in both of the other two interview rounds. This is because any survey prompts related to a period when the individual is not actively working cannot capture absenteeism in the context of this dissertation. The FYCD documentation files support this approach, explaining that the portion of the survey collecting work-loss information is independent of the section collecting detailed job-related measures. As a result, individuals who would logically have unobserved values for the *sickdays* variables in this paper may be recorded as reporting zero absences.[3,4]

Variables indicating job changes that occur between interview rounds are utilized to match the proper job characteristics to absenteeism reports relevant to the correct time period. For observations indicating a job change at some point between interview rounds, only the information for the first reported job is considered to mitigate the risk of matching reported work absences to characteristics of the wrong job. Individuals that have ever retired, have a disability, and military personnel are not included. For this research, I keep individuals observed only once in the sample and use pooled data and conduct cluster-robust inference at the individual level.[5]

After these alterations discussed above have been made, the final sample represents adults who were employed for a positive number of days during the respective calendar year and consists of 31,929 observations. The maximum number of absence days reported over an annual period is 160 days.

[3]Codebook Source: MEPS-HC Panel Design and Collection Process, Agency for Healthcare Research and Quality, Rockville, Md. https://meps.ahrq.gov/mepsweb/survey_comp/hc_data_collection.jsp.

[4]Future research could leverage such indicators to explore differences between individuals reporting long periods of layoffs or leave—due to disability, workers' compensation, or maternity/paternity leave—and the actively working population. These sub-populations are excluded in the current paper.

[5]See the following for more detailed descriptions of models estimating discrete outcomes for panel data: (Cameron and Trivedi, 2015; Cameron and Trivedi, 1986; Mundlak, 1978; Greene, 2002).

3.1 Variables

Table 11 in Appendix A reports variable definitions and descriptive statistics for the study sample. Also in the Appendix are tables that present summary statistics of analytical variables by gender (Table 12) and by gender and diagnosed mental illness (Table 13) In this section, I discuss the variables used in my analyses and in the following section I describe the descriptive statistics.

Dependent Variable (A_i): The dependent variable of interest is a count variable ($sickdays$) representing absences from work due to an injury or physical or mental illness or ailment. The MEPS includes a separate variable counting the days absent from work due to someone else's illness. I include only the variable prompting for a count of own-health related absence.

Mental Health (MH_i): The main explanatory variable of interest in this study is a binary variable indicating diagnosed mental illness ($keyMHdis$). This variable is based on responses reported in the Medical Conditions File (MCF) which provide condition-specific codes for various forms of mental illness. This variable is equal to one for individual's reporting diagnosis(es) of mood, anxiety, personality, or psychotic disorders. These categories of mental illness are chosen due to population prevalence, standard treatment protocols, and comorbidity hazards among them. I henceforth refer to these categories of mental illness as "key disorders." If an individual indicates a diagnosis of one or more of these key disorders, $keyMHdis$ equals one and is zero otherwise.

It should be noted that other classes of mental illness are also reported in the MCF, such as sexual disorders, conduct disorders, and developmental disorders. I choose not to include these diagnostic categories in variable $keyMHdis$ due to the differing nature of the diagnostic criteria associated with these groups of disorders according to the *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.; DSM–5; American Psychiatric Association, 2013). Sexual and conduct disorders are often times underreported and symptoms associated with these disorders are often external in nature. Developmental disorders are associated with highly heterogenous symptoms that may be internalized or externalized and that have a broad range of the degree of limitations associated with symptoms. The MEPS separately categorizes one form of developmental disorder, Attention Deficit/Hyperactivity Disorder (ADHD), "because of [its] relatively high prevalence, and because generally accepted standards for appropriate clinical care have been developed." As ADHD is technically a developmental disorder, the nature of its symptoms are less clear than the symptoms associated with the diagnostic categories included in the formation of $keyMHdis$; however, ADHD is highly comorbid with the diagnoses considered as "key" disorders in this paper so that I include a binary variable equal to one for individuals with ADHD as a control in the analysis.

It is important to further stress that the $keyMHdis$ variable represents only *diagnosed* mental illness across the sample so that mental disorder prevalence rates illustrated by the sample may not be representative of actual population prevalence rates in the US. That is, individuals may exhibit symptoms of mental illness even without a formal diagnosis. There is also likely variation in the degree of symptom severity associated with a specific disorder across individuals and failing to account for this possibility may bias the estimated effect of diagnosed mental illness.

The consideration of potential homogeneous symptoms between individuals with and without a given diagnosis, as well as heterogeneous symptoms across individuals within a particular diagnostic category are highlighted in the DSM-5, which separates itself from earlier DSM editions by its focus on a spectral approach to mental illness. DSM-5 states: "Earlier editions of DSM focused on excluding false-positive results from diagnoses; thus its categories were overly narrow, as is apparent from the widespread need to use NOS [not otherwise specified] diagnoses. Indeed the

once plausible goal of identifying homogeneous populations for treatment and research resulted in narrow diagnostic categories and did not capture clinical reality, symptom heterogeneity within disorders, and significant sharing of symptoms across multiple disorders".

In acknowledgment of within-disorder heterogeneity and that mental health may be, to some degree, independent of a particular diagnosis, I utilize an index variable from the FYCD that measures one's general level of emotional distress over the last 30 days on a Kessler-6 scale with scores ranging from 0-24 and higher values indicating more psychological distress (Kessler et al., 2003; Ashwood et al., 2017). This information is collected as part of the MEPS Self-Administered Questionnaire (SAQ) which is collected at rounds 2 and 4 of the MEPS.

General (Physical) Health (PH_i): Physical and mental health are inherently intertwined. The MEPS has numerous physical health markers. The FYCD files of the MEPS provide self-rated general health scores for each of the three interview rounds of the respective calendar year. Responses in each round range from 1-5 with higher values indicating worse perceived general health. These three variables are summed to create an annual index of general health originally with range 3-15; the index is rescaled by subtracting two from each observation, yielding a poor-health index (*physhlth*) ranging from 1-13 that acts as the main measure of general health.

As the index just described may be prone to attenuation bias due to self-perceptions of health, various other measures of health are utilized in the analysis. The second analytical variable in this category, named *prtycnds*, represents the number of diagnosed priority conditions as defined by the MEPS. The priority conditions specified by the FYCD are cancer, heart conditions, asthma, stroke, chronic bronchitis, emphysema, high cholesterol, high blood pressure, diabetes, arthritis, joint pain and ADHD. ADHD is not counted in *prtycnds* because it is psychological in nature; instead a dummy variable indicating diagnosed ADHD is utilized as a separate control. The MEPS singles out these priority condition categories "because of their relatively high prevalence, and because generally accepted standards for appropriate clinical care have been developed."

A binary variable equal to one if an individual has suffered an injury or illness (*injury*) in the past year that required immediate medical help (e.g. an unanticipated hospitalization) as well as a variable indicating that they have received a routine checkup within the past year (*routine*) are especially important control variables given the impossibility of these events co-occuring with absenteeism in the current year, eliminating the risk of simultaneity. There is no variable available to identify mental-health related visits over the past year due to confidentiality, but it should be noted that severe mental illness requiring hospitalization in the past year may be represented in variable *injury*.

Behavioral variables equal to one for individuals that exercise at least three days a week (*exercise*) and a binary variable identifying smokers (*smoke*) are other controls. For the sub-sample of women, a variable indicating pregnancy at the beginning of the year is also used as a control variable in analysis.

Job Traits (J_i): Variables indicating the fringe benefits offered to an individual by their employer such as paid sick leave (*sickpay*) and bonuses (*bonus*) are utilized as analytical variables in analysis as these factors likely influence an individual's decisions on short-run labor supply. It should be noted that binary variables indicating whether one receives paid vacation time and whether one receives paid leave to visit the doctor are also analyzed, but are determined to induce issues of collinearity with the *sickpay* variable. In addition, the inclusion of either of these binary variables does not significantly improve model fit so that they are excluded from analysis and I instead only focus on the analytical variable *sickpay*.

The decision making process on whether to go to work when facing illness is very likely to depend on the length of time for which an individual has been at their job; for example, an employee who has only worked at a firm for a year might be more averse to absence when ill than an equivalent worker who has worked at a firm for 15 years because less job-specific experience might make the newer employee relatively more expendable; there may also be varying degrees of rapport between the the firm and the employee based on the length of the worker's tenure, another channel through which tenure may impact labor supply decisions. A variable measuring an individuals job tenure in years ($tenure$) is created using information on job start and end dates provided by the MEPS.

Additional job-related controls include indicators for labor union membership ($union$) and dummy variables identifying seasonal workers ($ssnl$), temporary contracts ($temp$), and part-time employees ($parttime$), with individuals that typically work less than 35 hours per week considered part-time. Other job controls include indicator variables for public sector positions ($pubsect$), industry and occupational categories (see Appendix A: Table 11), and firm size (represented by variables $1to19$, $20to99$, $100to499$, and $500plus$ based on sample quartiles).

Health Insurance (I_i): The FYCD provides a copious amount of information on individuals' health insurance characteristics. An imputed variable that indicates the category of insurance the individual reported for the majority of the calendar year (categories are private, public, and uninsured) as well as variables reporting whether one held insurance through their employer from the raw FYCD files are used to derive a group of binary variables representing one's source of health insurance. A binary variable equal to one for individuals who have insurance through their job ($jobins$) is included in the estimation of Baseline Model 1 ($BM1$) described later.

Variables indicating uninsured individuals ($unins$), individuals insured privately through a source other than their employer ($otherins$), and individuals with a public source of health insurance ($pubins$) are created and the group of uninsured individuals acts as the reference category. A categorical variable from the FYCD is utilized to create a binary variable ($inscostly$) equal to one for individuals who either somewhat agree or strongly agree to the survey prompt, "health insurance is not worth its cost". Variable $inscostly$ acts as a proxy of health insurance generosity.

($\tilde{I}_i$): A model specification henceforth referred to as Baseline Model 2 ($BM2$), to be defined in the next chapter, replaces $jobins$ with a group of three binary variables, $plnchoic$, $nochoic$, and $NR.choic$. Two of these three variables will be analyzed in depth in later sections; these are $plnchoic$ and $nochoic$. $plnchoic$ identifies individuals with a choice between multiple insurance plans offered by an employer and $nochoic$ identifies the individuals that receive insurance through their employer but are only offered one plan option. Variable $NR.choic$ indicates observations that report having insurance through an employer but have missing values for prompts on whether or not the employer offers an array of plans to choose from.

Demographics (X_i): The highest level of educational attainment is controlled for using a group of binary variables ($belowhs$, $hsdeg$, $somecoll$, $bachdeg$, $bachplus$). Other variables control for race ($black$, $asian$), Hispanic ethnicity ($hispanic$), native-status ($bornus$), age (age), family size ($famsz$), the number of young children in the household ($yngchldrn$), marital status ($married$), and socioeconomic status ($poor$, $lowinc$, $midinc$, $highinc$).

Other Control Variables (C_i): Regional indicator variables (NE, MW, W, S) are generated using FYCD variables indicating one's region of residence for the majority of the calendar year. Monthly data from the U.S. Bureau of Labor Statistics on regional unemployment rates is utilized

to estimate the average unemployment rate faced by an individual ($unemp.rt$). Reference period start and end dates are used to generate these estimates for individuals that move to a different region during the year of interest; the unemployment rates are averaged across the months for which an individual reported a certain region of residence, then the average is taken across each of the regions that one resided in during the year. This measure of the average unemployment rate faced by an individual in a particular year provides the benefit of controlling for heterogeneity induced by macroeconomic features of the regional economy.

Following the theoretical framework, it is assumed that the short run decision-making process on work absence is only faced on the days that an individual is employed, as the model inherently assumes that an optimal employment contract has already been reached. Thus, differences in exposure to this decision-making process across individuals should be accounted for. I use information on interview start and end dates as well as job start and end dates and employment status to generate an estimate of the number of days employed during the respective calendar year ($empUB$). The log of this estimate is used as a control variable in analyses and roughly accounts for differences in exposure to work days.

Further considering exposure, as reference period rounds can be longer or shorter for some individuals, and because I keep individuals that may have been unemployed at some point during the calendar year in the sample, I include indicator variables that specify whether the individual was unemployed for one or two survey rounds ($unemp.1, unemp.2$). If an individual became unemployed during the reference period, a binary variable, *partialemp* equals one.

Persons observed for a single year are split into two categories to indicate which of the two years of participation in the panel that they are observed for. Two dummy variables indicate these two sub-groups of individuals, one indicates that the individual is observed in his or her first year of participation (*yearone*). The second identifies individuals observed in the second (*yeartwo*) year of participation. Sample statistics for a third variable, *bothyears*, for individuals observed in the sample for both years in which their household participated, are included in Appendix A: Table 11, but it should be noted that *bothyears* acts as a reference variable in the empirical analyses.

I control for attrition induced by the data generating process. I control for other outside factors that might impact absence decisions such as a move from one region of the US to another at some point in the calendar year ($moved\_US$) and a move from one participating household to another ($moved\_RU$). Finally, binary variables indicating each observation's respective calendar year (2010 - 2014) are used to account for year fixed effects.

The following section will define econometric models and methods utilized in empirical estimation.

4 Baseline Models of Absenteeism

Unless explicitly mentioned *all* of the variables grouped by category labels in Appendix A are included in empirical modeling *with the exception* of the final two variables in the category titled "Exclusion Restrictions (ER)". These last two variables in category ER are not utilized until later on in this paper when endogenous selection into employment is addressed, and should be ignored for now.

In what follows, I define two baseline conditional mean functions of absenteeism.

4.1 Econometric Modeling

The dependent variable used in each of the analyses is a count of the number of days absent from work due to one's own physical or mental illness over the span of one calendar year, A_i for observation i. Due to the count nature of dependent variable A_i, I specify an exponential conditional mean function.

I specify two separate baseline specifications of the conditional mean of absence days. The first baseline conditional mean specification is defined as follows:

$$\begin{aligned} E[A_i|MH_i,PH_i,J_i,I_i,X_i,C_i] = &\exp(\beta^0_{BM1} + MH_i\beta^{MH}_{BM1} + PH_i\beta^{PH}_{BM1} \\ &+ J_i\beta^J_{BM1} + I_i\beta^I + X_i\beta^X_{BM1} + C_i\beta^C_{BM1}), \end{aligned} \tag{19}$$

for every observation, $i = 1,...,n$. Equation (19) will henceforth be referred to as $BM1$ (baseline model one). Matrices MH, PH, J, I, X, and C hold observed values of the corresponding explanatory variables for all $i = 1$, ..., n observations. The previous chapter discussed the explanatory variables included in each of these matrices and Appendix A also groups explanatory variables in terms of these matrices and provides definitions. Each matrix of explanatory variables is of dimension $n \times k_l$ where n is the total number of observations in the sample and k_l is an integer equal to the number of explanatory variables in corresponding matrix l for $l = \{MH, PH, J, I, X, C\}$. In a similar fashion, each β^l for $l = \{MH, PH, J, I, X, C\}$ is a $k_l \times 1$ vector of parameters.[6]

A second specification of the conditional mean of A_i that I reference as $BM2$ (baseline model two) throughout the rest of the paper is defined in a similar manner to (19). $BM2$ is identical to specification $BM1$ except that matrix $\tilde{I}$ replaces matrix I for the $BM2$ specification. As noted in Chapter 4, $\tilde{I}$ in $BM2$ holds variables *inscostly* and the set of dummy variables *nochoic*, *plnchoic*, *NR.choic*, *pubins*, and *otherins*, with *unins* acting as the reference category variable. Variables *plnchoic*, *nochoic*, and *NR.choic* further break down the variable *jobins* which is included in the $BM1$ specification in place of those three variables. The purpose of this break down is to analyze and compare how the implementation of different types of benefit package designs influence worker absence behavior. The $BM2$ specification takes the form

$$\begin{aligned} E[A_i|MH_i,PH_i,J_i,\tilde{I}_i,X_i,C_i] = &\exp(\beta^0_{BM2} + MH_i\beta^{MH}_{BM2} + PH_i\beta^{PH}_{BM2} \\ &+ J_i\beta^J_{BM2} + \tilde{I}_i\beta^{\tilde{I}} + X_i\beta^X_{BM2} + C_i\beta^C_{BM2}), \end{aligned} \tag{20}$$

where each β^j for $j = \{MH, PH, J, \tilde{I}, X, C\}$ is a $k_j \times 1$ vector of parameters, where k_j is an integer equal to the number of explanatory variables held in corresponding matrix j.

Thus far, I have used notation that assigns parameter subscripts that indicate the specific baseline specification in order to highlight that coefficient estimates may differ between $BM1$ and $BM2$. For notational efficiency I henceforth drop these subscripts and work within a more general framework.

Coefficients are not directly interpretable in nonlinear models, especially if the goal of research is to form policy implications (Braumoeller, 2004; Buis, 2010; Chunrong & Norton, 2003; Long & Freese, 2006; Williams, 2009). Instead, I prefer to test hypotheses on the difference in the conditional mean expectation of absences resulting from a discrete change in an explanatory variable; this will allow for more intuitive inference (Wooldridge, 2010, *Econometric Analysis of Cross Section*

[6] For example, β^{MH}_{BM1} is a 3×1 vector holding parameter values for the three explanatory variables represented in matrix MH, which are *keyMHdis*, *distress*, and *adhd*.

and Panel Data, 737).

Define ME_i as the effect of a diagnosed mental illness on the expected value of annual illness-related work absences, which is equal to the change in expected absences when the binary variable $keyMHdis$ changes between zero and one. Following from the theoretical framework laid out in Chapter 3, consider an individual endowed with a low level of mental health who faces the same time constraints, budget constraints, and preferences over health and conglomerate consumption as a second individual endowed with a higher level of mental health.

The former individual will, in theory, exhibit the same optimal level of health production, H^*, as the latter individual; however, the former individual will require more time and market inputs to reach or maintain level H^* compared to the latter individual due to the less-efficient health production process imposed by the low health endowment. It follows that the individual with the lower endowment of mental health will exhibit a lower optimal level of labor supply relative to the individual with high mental health endowment. This result can easily be used to form a testable empirical hypothesis defined as follows:

$$\begin{aligned} E[ME_i] &= E[A_i|keyMHdis_i = 1] - E[A_i|keyMHdis_i = 0] > 0 \\ &= \exp(\beta^{MH_1} + \sum_{l \neq MH_1} x_{i,l}\beta^l) - \exp(\sum_{l \neq MH_1} x_{i,l}\beta^l) > 0, \end{aligned} \tag{21}$$

where β^{MH_1} is the parameter corresponding to the first variable in vector MH, which is assumed to be binary variable $keyMHdis$. Matrix x_l holds observed values for every other explanatory variable $l \neq MH_1$,[7] and β^l is the coefficient corresponding to explanatory variable l. In words, I hypothesize that diagnosed mental illness is associated with higher expected annual absences.

In what follows, I consider how the magnitude of (21) may change across levels of various job, health insurance, and health variables. While this would require the use of the appropriate interaction terms if utilizing a linear model, in nonlinear models of exponential form, the effect of one variable ($keyMHdis$, in this dissertation) may vary with the value of a second variable without the additional consideration of an interaction term between the two variables. Aspects of employment contracts and health insurance coverage are theoretically different in nature than variables characterizing one's mental health endowment because of the different channels through which each of these types of factors influence absence behavior (e.g., sick leave and health insurance generosity influence choices of time and market inputs in health production, while mental illness is tied to an efficiency factor. See Chapter 3 for more detailed discussion).

Conditional marginal effects, where job and insurance characteristics act as the conditioning variables, are more useful in forming testable hypotheses compared to the use of explicit interaction terms (Karaca-Mandic et al., 2012). I later define extended model specifications that include interaction terms between $keyMHdis$ and other variables measuring health and form hypotheses on the sign of these terms.

In the following section, I present hypotheses on the conditional marginal effect of mental illness at different levels of employment contract and health insurance variables.

[7] Note that $x_{i,l}$ is a generalization for notational efficiency. In practice, $E[ME_i]$ will be estimated for both $BM1$ and $BM2$ specifications so that the explanatory variables in matrix x_l for $l \neq MH_1$ will be different based on the model specification used.

4.2 Estimation Strategy

I finally discuss the applied econometric strategy used to generate estimates and test hypotheses.

The full sample resulting from the data-generating process discussed in Chapter 4 is unbalanced; some individuals are observed once while others are observed twice. This poses the issue of potential attrition bias. I include multiple control variables that identify possible sources of attrition for individuals observed once, including variables indicating which year of participation (first year or second) in the MEPS these individual's are observed for and whether these individuals participated for both years of MEPS and thus are observed once as a result of my own data-generating process. The inclusion of these variables is inspired by strategies proposed by Nijman & Verbeek (1992) to roughly control for some of the attrition bias induced by the unbalanced sample design.

Given the significant differences exhibited by the sample between men and women, I conduct analyses for men and women separately. I consider two conditional mean distributions: Poisson and negative binomial, which inherently account for the fact that A_i is restricted to non-negative values. The negative binomial distribution innately allows for overdispersion and includes the Poisson distribution as a special case, allowing for a more flexible specification. I thus choose to utilize a negative binomial model to estimate $BM1$, $BM2$, $I1$, $I2$, and $I3$ conditional mean specifications.

Though estimates are only reported for a subset of explanatory variables for organizational purposes, it should be noted that no variables are dropped from any of these model specifications at any point in the estimation process. The MEPS design is such that information on every individual living within a household is collected. Thus, some individuals in my sample are from the same family unit. Due to possible correlation of unobservables within the same family unit, I cluster standard errors at the family level. Further, there is likely an individual-specific error component that will be correlated within individuals that have two observations in the sample leading to bias in variance estimates for these individuals. Thus, standard errors are additionally clustered at the individual level.

Unconditional average marginal effects (AME) of a subset of explanatory variables are estimated for both of the baseline model specifications. Conditional AME (CAME) estimates for $keyMHdis$ are estimated using the $BM1$ specification for conditioning variables *sickpay*, *bonus*, *jobins*, and *inscostly*. Estimated CAME for $keyMHdis$ conditional on variables *plnchoic* and *nochoic* are generated using the $BM2$ specification. CAME estimates of $keyMHdis$ given conditioning variables *pubins* and *otherins* are estimated for each of the $BM1$ and $BM2$ specifications. Estimates of the moderating effect of each of the conditioning variables on the AME of $keyMHdis$ are calculated as the mean difference in conditional fitted value estimates across levels of the moderating variable and are based on manipulating the sample to assign a particular level of a variable to the entire sample; thus, the moderating effect estimates are based on counterfactual information. Standard errors for all AME, CAME, and moderating effect estimates are computed using the delta method.

Prior to estimation, I consider the possibility of harmful collinearity that may persist when including both $keyMHdis$ and *distress* together in each model. I conduct a VIF test for variables $keyMHdis$ and *distress* for each sex separately and for both $BM1$ and $BM2$ specifications. The results of these tests suggest no statistical evidence of significant multicollinearity by including both of these measures of mental health in a model in tandem. I perform a likelihood ratio test of the null hypothesis that a model specification excluding *distress* provides a superior fit to a model that includes only the diagnosis binary variable. The test indicates rejection of the null hypothesis of the restricted model in favor of the model including both measures of mental health at the 1

percent level.

4.3 Baseline Model ($BM1$, $BM2$)

The estimated dispersion parameter is approximately 0.21 for the baseline negative binomial model applied to the sub-sample of men with an estimated standard error of 0.004; the equivalent values for the baseline estimator for women in the sample are 0.28 wand 0.01, respectively.

Table 2 reports AME for analytical variables. Pseudo R^2 are also reported. AME estimates represent the expected change in the count of absence days per year in response to a unit change in the respective explanatory variable, on average. Estimates are robust across both $BM1$ and $BM2$ specifications.

Both men and women with a diagnosed mental illness are expected to exhibit a higher number of absence days than counterparts without a diagnosed mental illness, on average. These estimates support the empirical hypothesis that the unconditional AME of $keyMHdis$ is positive and are consistent with the literature on mental health and worker productivity. AME estimates for $keyMHdis$ reported in columns one through four are all statistically significant at the one-percent level. Men with a diagnosed mental disorder are estimated to report 1.11 days of additional absence, on average, compared to identical counterparts without a diagnosed mental disorder. The average woman in the sample is expected to report about 0.92 additional annual absences when diagnosed with a mental illness.

Estimates for the other three variables measuring health that are reported in Table 2 are highly statistically significant at the 0.1 percent level for both sexes. A marginal increase in general distress is estimated to increase absence by an average of 0.11 days for men and 0.15 days for women. The generalizability of these estimates should be interpreted with caution; this measure of general distress is based on responses to survey prompts given in only the second of the three interview rounds that occurs per year and so it is unclear how issues of timing may influence these estimates. For example, some individuals in the sample may have been unemployed at the time of the second interview so that the estimated effect of distress on absenteeism is not representative of the true effect of distress on absences from work for the employed.

In other words, the estimates for variable *distress* are only generalizable if most individuals in the sample who are unemployed at the second interview round report levels of distress that are homogeneous to (unobserved) levels of distress exhibited during periods of employment. AME estimates regarding self-reported measures of general physical health suggest that poorer degrees of physical health are associated with higher absences from work, on average, as anticipated.

A one-point increase in the rating of one's own degree of poor physical health is estimated to increase absences by under half a day for men (by a factor of about 0.37) on average, and over half a day (a factor of 0.55) for women in the sample, on average. This discrepancy across the sexes could be driven by differences in preferences over health across sex, but it is unclear to what degree this is the case. AME estimates suggest that men and women respond similarly to an additional priority condition diagnosis on average.

Table 2 reports that AME estimates for variable *sickpay* are significant at the five-percent level and are robust across models for both men and women. On average, receiving compensation for health-related absence from work is anticipated to increase expected days absent by about half a day for the sample of men. Estimates reported in columns three and four suggest that women may be slightly more responsive to a fringe benefit that supplements time lost from work due to illness, on average.

Table 1: Baseline Models 1 and 2: AME Estimates

Dependent Variable:	Men		Women	
sickdays	$BM1$	$BM2$	$BM1$	$BM2$
keyMHdis	1.11 (0.35)***	1.10 (0.34)***	0.92 (0.31)***	0.92 (0.31)***
distress	0.11 (0.06)**	0.11 (0.03)***	0.15 (0.03)***	0.15 (0.03)***
physhlth	0.37 (0.05)***	0.37 (0.05)***	0.55 (0.06)***	0.55 (0.06)***
prtycnds	0.75 (0.09)***	0.67 (0.09)***	0.69 (0.09)***	0.69 (0.09)***
sickpay	0.49 (0.23)**	0.45 (0.23)**	0.63 (0.30)**	0.62 (0.30)**
bonus	-0.19 (0.22)	-0.21 (0.22)	0.08 (0.28)	0.07 (0.28)
jobins	0.82 (0.26)***		1.23 (0.34)***	
plnchoic		1.04 (0.30)***		1.41 (0.38)***
nochoic		0.80 (0.28)***		1.09 (0.37)***
pubins	0.90 (0.45)**	0.87 (0.44)**	1.15 (0.44)***	1.15 (0.43)***
otherins	0.65 (0.33)**	0.68 (0.32)**	0.39 (0.36)	0.41 (0.36)
inscostly	-0.42 (0.19)**	-0.43 (0.19)**	-0.34 (0.24)	-0.32 (0.24)
Observations	15,713		16,216	
Pseudo R^2:	0.164		0.198	

Note: Values in parentheses are standard errors computed using the delta method. *p<0.1; **p<0.05; ***p<0.01

Having health insurance through a job is estimated to have a highly significant, positive association with absenteeism. Men enrolled in health insurance through an employer reporting an average of 0.82 additional days of absence compared to the average of 1.23 additional expected absences estimated for women. Women exhibit larger magnitudes of AME estimates compared to men when having a public source of insurance as a well as some other form of private insurance. This is not unusual considering that the empirical literature suggests that women have a greater propensity to utilize healthcare.

Results for variable *pubins* suggest that having a public source of health insurance slightly amplifies expected absences, on average, relative to having insurance through an employer. This phenomena might be partially driven by the association between poverty and health – low-income individuals who are eligible for public health insurance may face more health issues and thus require more time away from work; public health insurance may also be offered to individuals with debilitating health problems who require more time outside of work to receive care.

Having a source of private insurance provided by an entity other than an employer (represented by variable *otherins*) is significant at the 5 percent level for men; this is not significant for women in the sample. Men reporting agreement to the prompt "my health insurance is not worth its cost" exhibit a decline in absence days of 0.42 to 0.43 days with estimates robust at the 5 percent level across models. One the other hand, women exhibit a similar magnitude and sign of the AME estimate for *inscostly*, however, these are not statistically significant. This result is consistent with literature on the difference between men and women in seeking healthcare when ill.

The $BM2$ specification yields AME estimates for variables *plnchoic*, *nochoic*, and *NR.choic* that further break down the *jobins* variable into three additional categories of insurance. Individuals enrolled through an employer-sponsored plan are grouped based on whether their employer provides a choice between a portfolio of health plans (*plnchoice* = 1) or offers one health plan

($nochoic = 1$). Employed individuals that report being enrolled in an employer-sponsored plan but fail to answer survey prompts on whether or not their employer offers multiple plans to choose from ($NR.choic$). Variable $NR.choic$ is utilized as a necessary control so that the reference insurance source variable is still $unins$. Variable $NR.choic$ is not considered an analytical variable. Therefore, estimates are not reported for this variable in what follows.

Men with employer-provided insurance that chose from a catalog of plans exhibit an average of 1.04 additional days absent relative to uninsured men, and this AME estimate is significant at the 1 percent level. AME estimates of the $plnchoic$ variable for sample women are also significant at the 1 percent level, with results suggesting the ability to select into an insurance plan is associated with an average increase in absenteeism of about 1.41 days for employed women. AME estimates of the additional absence days induced by a single-plan employer-sponsored package (when $nochoic = 1$) are comparatively smaller to the results for variable $plnchoic$.

In what follows I address the potential bias induced by non-random selection into the employed sample that is likely present for a number of reasons, such as sample error, attrition in sample editing, and unobservable data for all except a sub-population of unemployed individuals.

5 Results: Average Marginal Effects (AME)

Prior to estimation, I consider the possibility of harmful collinearity that may persist when including both $keyMHdis$ and $distress$ together in each model. I conduct a VIF test for variables $keyMHdis$ and $distress$ for each sex separately and for both $BM1$ and $BM2$ specifications. The results of these tests suggest no statistical evidence of significant multicollinearity by including both of these measures of mental health in a model in tandem. I perform a likelihood ratio test of the null hypothesis that a model specification excluding $distress$ provides a superior fit to a model that includes only the diagnosis binary variable. The test indicates rejection of the null hypothesis of the restricted model in favor of the model including both measures of mental health at the 1 percent level.

5.1 Baseline Model ($BM1$, $BM2$)

The estimated dispersion parameter is approximately 0.21 for the baseline negative binomial model applied to the sub-sample of men with an estimated standard error of 0.004; the equivalent values for the baseline estimator for women in the sample are 0.28 wand 0.01, respectively.

Table 2 reports AME for analytical variables. Pseudo R^2 are also reported. AME estimates represent the expected change in the count of absence days per year in response to a unit change in the respective explanatory variable, on average. Estimates are robust across both $BM1$ and $BM2$ specifications.

Both men and women with a diagnosed mental illness are expected to exhibit a higher number of absence days than counterparts without a diagnosed mental illness, on average. These estimates support the empirical hypothesis that the unconditional AME of $keyMHdis$ is positive and are consistent with the literature on mental health and worker productivity. AME estimates for $keyMHdis$ reported in columns one through four are all statistically significant at the one-percent level. Men with a diagnosed mental disorder are estimated to report 1.11 days of additional absence, on average, compared to identical counterparts without a diagnosed mental disorder. The average woman in the sample is expected to report about 0.92 additional annual absences when diagnosed with a mental illness.

Estimates for the other three variables measuring health that are reported in Table 2 are highly statistically significant at the 0.1 percent level for both sexes. A marginal increase in general distress is estimated to increase absence by an average of 0.11 days for men and 0.15 days for women. The generalizability of these estimates should be interpreted with caution; this measure of general distress is based on responses to survey prompts given in only the second of the three interview rounds that occurs per year and so it is unclear how issues of timing may influence these estimates. For example, some individuals in the sample may have been unemployed at the time of the second interview so that the estimated effect of distress on absenteeism is not representative of the true effect of distress on absences from work for the employed.

Table 2: Baseline Models 1 and 2: AME Estimates

Dependent Variable:	Men		Women	
sickdays	$BM1$	$BM2$	$BM1$	$BM2$
keyMHdis	1.11 (0.35)***	1.10 (0.34)***	0.92 (0.31)***	0.92 (0.31)***
distress	0.11 (0.06)**	0.11 (0.03)***	0.15 (0.03)***	0.15 (0.03)***
physhlth	0.37 (0.05)***	0.37 (0.05)***	0.55 (0.06)***	0.55 (0.06)***
prtycnds	0.75 (0.09)***	0.67 (0.09)***	0.69 (0.09)***	0.69 (0.09)***
sickpay	0.49 (0.23)**	0.45 (0.23)**	0.63 (0.30)**	0.62 (0.30)**
bonus	-0.19 (0.22)	-0.21 (0.22)	0.08 (0.28)	0.07 (0.28)
jobins	0.82 (0.26)***		1.23 (0.34)***	
plnchoic		1.04 (0.30)***		1.41 (0.38)***
nochoic		0.80 (0.28)***		1.09 (0.37)***
pubins	0.90 (0.45)**	0.87 (0.44)**	1.15 (0.44)***	1.15 (0.43)***
otherins	0.65 (0.33)**	0.68 (0.32)**	0.39 (0.36)	0.41 (0.36)
inscostly	-0.42 (0.19)**	-0.43 (0.19)**	-0.34 (0.24)	-0.32 (0.24)
Observations	15,713		16,216	
Pseudo R^2:	0.164		0.198	

Note: Values in parentheses are standard errors computed using the delta method. *p<0.1; **p<0.05; ***p<0.01

In other words, the estimates for variable *distress* are only generalizable if most individuals in the sample who are unemployed at the second interview round report levels of distress that are homogeneous to (unobserved) levels of distress exhibited during periods of employment. AME estimates regarding self-reported measures of general physical health suggest that poorer degrees of physical health are associated with higher absences from work, on average, as anticipated.

A one-point increase in the rating of one's own degree of poor physical health is estimated to increase absences by under half a day for men (by a factor of about 0.37) on average, and over half a day (a factor of 0.55) for women in the sample, on average. This discrepancy across the sexes could be driven by differences in preferences over health across sex, but it is unclear to what degree this is the case. AME estimates suggest that men and women respond similarly to an additional priority condition diagnosis on average.

Table 2 reports that AME estimates for variable *sickpay* are significant at the five-percent level and are robust across models for both men and women. On average, receiving compensation for health-related absence from work is anticipated to increase expected days absent by about half a day for the sample of men. Estimates reported in columns three and four suggest that women may

be slightly more responsive to a fringe benefit that supplements time lost from work due to illness, on average.

Having health insurance through a job is estimated to have a highly significant, positive association with absenteeism. Men enrolled in health insurance through an employer reporting an average of 0.82 additional days of absence compared to the average of 1.23 additional expected absences estimated for women. Women exhibit larger magnitudes of AME estimates compared to men when having a public source of insurance as a well as some other form of private insurance. This is not unusual considering that the empirical literature suggests that women have a greater propensity to utilize healthcare.

Results for variable *pubins* suggest that having a public source of health insurance slightly amplifies expected absences, on average, relative to having insurance through an employer. This phenomena might be partially driven by the association between poverty and health – low-income individuals who are eligible for public health insurance may face more health issues and thus require more time away from work; public health insurance may also be offered to individuals with debilitating health problems who require more time outside of work to receive care.

Having a source of private insurance provided by an entity other than an employer (represented by variable *otherins*) is significant at the 5 percent level for men; this is not significant for women in the sample. Men reporting agreement to the prompt "my health insurance is not worth its cost" exhibit a decline in absence days of 0.42 to 0.43 days with estimates robust at the 5 percent level across models. One the other hand, women exhibit a similar magnitude and sign of the AME estimate for *inscostly*, however, these are not statistically significant. This result is consistent with literature on the difference between men and women in seeking healthcare when ill.

The $BM2$ specification yields AME estimates for variables *plnchoic*, *nochoic*, and $NR.choic$ that further break down the *jobins* variable into three additional categories of insurance. Individuals enrolled through an employer-sponsored plan are grouped based on whether their employer provides a choice between a portfolio of health plans ($plnchoice = 1$) or offers one health plan ($nochoic = 1$). Employed individuals that report being enrolled in an employer-sponsored plan but fail to answer survey prompts on whether or not their employer offers multiple plans to choose from ($NR.choic$). Variable $NR.choic$ is utilized as a necessary control so that the reference insurance source variable is still *unins*. Variable $NR.choic$ is not considered an analytical variable. Therefore, estimates are not reported for this variable in what follows.

Men with employer-provided insurance that chose from a catalog of plans exhibit an average of 1.04 additional days absent relative to uninsured men, and this AME estimate is significant at the 1 percent level. AME estimates of the *plnchoic* variable for sample women are also significant at the 1 percent level, with results suggesting the ability to select into an insurance plan is associated with an average increase in absenteeism of about 1.41 days for employed women. AME estimates of the additional absence days induced by a single-plan employer-sponsored package (when *nochoic* = 1) are comparatively smaller to the results for variable *plnchoic*.

In what follows I address the potential bias induced by non-random selection into the employed sample that is likely present for a number of reasons, such as sample error, attrition in sample editing, and unobservable data for all except a sub-population of unemployed individuals.

6 Addressing Heterogeneity and Endogenous Sample Selection

I discuss and estimate models of correlated random effects in the next section of this paper. Results indicate that AME estimates and coefficients are highly robust and consistent after including an

estimate of individual heterogeneity. I then expand this analysis to examine when the assumption of conditional exogeneity required to reliably estimate models of correlated random effects is violated. This occurs in the case of sample selection imposed by unobservables that remain even after the indiviudal specific heterogeneity is accounted for. I use a classic two-step sample selection model and then a semi-parametric Copula based approach that has rarely been applied in the literature.

6.1 Correlated Random Effects

Throughout this paper, I have argued and found evidence that symptomatology plays a significant role in labor supply that varies between and within diagnostic status. An equally critical consideration is how people respond to mental health shocks. Ignoring individual-specific patterns within the sample assumes that all individuals respond uniformly to a given shock, holding all else constant. This assumption overlooks the nuanced variability in responses across individuals in the panel data framework.

Sources of heterogeneity may capture worker emotional "grit" and how this may generate omitted variable bias. While a worker's grit is not directly observable until after a contract is undertaken, it fundamentally shapes their production process and must therefore be accounted for. In consideration of these factors, I employ the correlated random effects (CRE) method proposed by Wooldridge (2019).[8] This approach leverages the panel design by incorporating individual-specific means of time-varying explanatory variables. The CRE method accounts for the for individual-level heterogeneity by using these means as control variables (Benson et al., 2022; Heckman, Urzua, and Vytlacil, 2006). The following provides the generalizations of the model and definitions of necessary assumptions.

I utilize the Wooldridge method due to the unbalanced panel I work with, in which some individuals are observed for one year, while others are observed twice – once in each year of participation in the MEPS. For individual i at time t:

$$y_{it}|x_{it}, \alpha_i \sim \text{NegBin}(\mu_{it}, \theta) \quad \text{with} \quad \mu_{it} = \exp(x'_{it}\beta + \alpha_i).$$

For CRE to be feasible, the following assumptions should hold:

$$\alpha_i = \bar{x}'_i\xi + a_i, \quad E[a_i|x_i] = 0.$$

Finally, marginalize over $a_i \sim \text{Gamma} \Rightarrow$ Negative Binomial likelihood:

$$E[y_{it}|x_{it}, \bar{x}_i] = \int \exp(x'_{it}\beta + \bar{x}'_i\xi + a)h(a)da.$$

A final estimate of the average marginal effects (AME) can be estimated simply after accounting for the estimated within-individual effects. Table 3 and Table 4 show estimates for men and women, respectively. The columns labeled $BM1$ in each table report the AME estimates initially reported in Table 2 as well as the coefficient estimates from the $BM1$
model specification. Columns titled "Correlated RE" in Tables 3 and 4 report estimates after including the individual-specific means of explanatory variables as control variables in the analysis.

The third and sixth columns in both Table 3 and Table 4 provide estimates of the AME and coefficients when additionally conditioning on the individual-specific intercept parameters. The

[8] Wooldridge 2019 is a method for nonlinear estimators and unbalanced samples as an extension of the Mundlak approach (1978).

Table 3: Estimates for Pooled Negative Binomial and Correlated Random Effects Negative Binomial: Men

Dependent variable:	*AME*		Coefficients	
sickdays	$BM1$	Correlated RE	$BM1$	Correlated RE
keyMHdis	1.11	0.95	0.37***	0.33***
	(0.43, 1.77)	(0.39, 1.51)	(0.10)	(0.11)
distress	0.11***	0.10***	0.04***	0.03***
	(0.06, 0.16)	(0.05, 0.14)	(0.01)	(0.01)
physhlth	0.37	0.35	0.12***	0.12***
	(0.27, 0.46)	(0.27, 0.43)	(0.01)	(0.01)
prtycnds	0.67***	0.62***	0.22	0.21***
	(0.50, 0.84)	(0.48, 0.75)	(0.02)	(0.02)
sickpay	0.49**	0.44	0.16**	0.15**
	(0.04, 0.93)	(0.02, 0.85)	(0.08)	(0.07)
bonus	-0.19	-0.16	-0.06	-0.06
	(-0.61, 0.24)	(-0.56, 0.23)	(0.07)	(0.07)
jobins	0.82***	0.81***	0.30***	0.36***
	(0.31, 1.32)	(0.32, 1.30)	(0.10)	(0.10)
pubins	0.90**	0.76*	0.33**	0.30**
	(0.003, 1.78)	(-0.05, 1.56)	(0.15)	(0.15)
otherins	0.65**	0.61**	0.25**	0.25**
	(0.01, 1.29)	($-1e^{-3}$, 1.22)	(0.12)	(0.12)
inscostly	-0.42**	-0.38**	-0.14**	-0.13**
	(-0.78, -0.04)	(-0.73, -0.04)	(0.06)	(0.06)
Observations	15,713			

*p<0.1; **p<0.05; ***p<0.01

fringe benefit variables are robust, with *sickpay* coefficients remaining statistically significant and *bonus* consistently not exhibiting statistical nor economic significance.

Access to insurance (job, public, or other) consistently increases sick days for both genders, likely due to reduced financial barriers to taking time off.

The limitations of the CRE method is the main assumption that x_{it} is strictly exogenous after conditioning on the individual level heterogeneity, α_i. Said differently, an individual is observed randomly conditional on their individual-specific intercept and there is no mechanism that induces an individual to be observed after a shock occurs. In what follows, I test for a non-random mechanism of observation.

6.2 Addressing Sample Selection on Unobservable Components

Individuals who are not presently employed do not have observable information for absenteeism. The decision not to work may relate to the factors that affect absenteeism. Thus, I conduct a final analysis using methods to account for unobserved sample selection. First, I use the classic Heckman sample selection model even though the outcome of absenteeism is a count variable. This approach ignores the discrete nature of the outcome variable and the overdispersion in the data. Later on, I invoke an estimator that uses copulas and penalized maximum likelihood to estimate the

Table 4: Estimates for Pooled Negative Binomial and Correlated Random Effects Negative Binomial: Women

Dependent variable:	*AME*		Coefficients	
sickdays	$BM1$	CRE	$BM1$	CRE
keyMHdis	0.92**	0.91***	0.21***	0.21***
	(0.33,1.54)	(0.35, 1.46)	(0.07)	(0.06)
distress	0.15***	0.14***	0.03***	0.03***
	(0.09, 0.21)	(0.09, 0.20)	(0.01)	(0.01)
physhlth	0.55**	0.53***	0.12***	0.12***
	(0.44, 0.66)	(0.43, 0.64)	(0.01)	(0.01)
prtycnds	0.69***	0.66***	0.16***	0.15***
	(0.52, 0.88)	(0.49, 0.82)	(0.02)	(0.02)
sickpay	0.63**	0.55*	0.14**	0.13*
	(0.04, 1.22)	(-0.01, 1.11)	(0.07)	(0.07)
bonus	0.08	0.06	0.02	0.01
	(-0.46, 0.64)	(-0.46, 0.57)	(0.06)	(0.06)
jobins	1.23***	1.24***	0.31***	0.31***
	(0.56, 1.90)	(0.59, 1.88)	(0.09)	(0.09)
pubins	1.15***	1.17***	0.29***	0.30***
	(0.30, 2.01)	(0.35, 1.99)	(0.10)	(0.10)
otherins	0.39	0.39	0.11	0.11
	(-0.31, 1.10)	(-0.30, 1.07)	(0.10)	(0.10)
inscostly	-0.34	-0.31	-0.07	-0.07
	(-0.80, 0.15)	(-0.76, 0.14)	(0.05)	(0.05)
Observations		16,216		

*p<0.1; **p<0.05; ***p<0.01

dependence between the selection and outcome equations, allowing me to use a negative binomial distribution in the second stage.

In what follows, I only am interested in analyzing health-related variables. However, some job factors are utilized as controls in the outcome equations. Because the first stage model estimates the probability of employment, I also include a sample of unemployed persons in this analysis these individuals make up a sample representing adults in the labor force. For consistency, the same data-generating processes are implemented to obtain the sample of unemployed persons so that it excludes individuals who have ever retired, military personnel, disabled individuals reporting significant difficulty completing activities of daily life, and students.

Additionally, I impose the condition of only excluding persons who have never worked and those currently out of the labor force at the time of the MEPS interviews so that I am focusing on behavior of people who could work (Jones et al., 2008; Certo et al., 2016). The inclusion of unemployed individuals yields a full sample size of 23,093 for women and 19,823 for men.

As a final comparison, I consider working-age individuals who persist after the data-generating process previously described and who are outside the labor force (and thus are not unemployed). The extended sample for men consists of 24,562 observations; for women, there are 29,947 observations.

Though estimates of sample selection models can be estimated when the same explanatory variables are included in the first stage equation and second stage equation, the estimation can be improved by including some variables in the selection (first stage) equation that are not included in the outcome equation. These variables are called exclusion restrictions, and strong exclusion restrictions are variables that directly impact the likelihood of employment, but do not directly impact the count of health-related workplace absenteeism.

I propose that two variables may act as exclusion restrictions. The names of these variables are *depout* and *spou.emp*. I will start with the definition and my argument for the variable *depout*. This variable represents the presence of dependents living outside the household, such as college students, non-custodial children, or elderly parents in assisted care. A binary variable equal to one if a person's spouse in employed, *spou.emp*, is also considered as an exclusion restriction.

Dependents inside the household may directly affect daily labor supply decisions, as their health or needs could influence a worker's ability to work. In contrast, dependents outside the household are unlikely to affect daily health-related labor supply decisions but may increase the need for income, thereby influencing employment decisions. Therefore, I argue that the variable *depout* will increase the probability of employment due to higher income needs but is not directly related to health-specific absenteeism.

A similar argument applies to the variable *spou.emp*: while the employment status of one's spouse is likely to strongly impact the individual's employment status, it does not directly influence an individual's decision to be absent from work after experiencing a negative health shock.[9]

The first stage probit includes all categories of explanatory variables as defined in Section 4.1 and also defined in Appendix A, except for the job characteristics variables (those grouped into variable category J). Variables *jobins* and *otherins* in category I of explanatory variables are additionally merged into a single variable representing whether one has a private source of insurance information. Additionally the two exclusion restrictions discussed previously are included in this probit equation. The binary variable EMP, equal to one if an individual is employed, zero otherwise, is the dependent variable in the probit equation. The second stage outcome equation is estimated using OLS and the equation includes both category J, and the inverse mills ratio, which controls for the dependence between selection into employment and the associated outcome.

The probit model estimating probability of employment in the first stage is:

$$P(Y=(EMP=1)|X=\{MH,PH,I,X,C,ER\}) = F(\beta_0 + MH_i\beta^{MH} + PH_i\beta^{PH} + I_i\beta^I + X_i\beta^X + C_i\beta^C + ER_i\beta^{ER}),$$

for every observation, $i = 1,...,n$. Matrices MH, PH, I, X, C, and ER hold observed values of the corresponding explanatory variables for all $i = 1$, ..., n observations. Matrix ER holds the exclusion restrictions *depout* and *spou.emp*. The function $P(.)$ represents the probability of being employed, and F represents the Normal distribution.

Each matrix of explanatory variables is of dimension $n \times k_l$ where n is the total number of observations in the sample and k_l is an integer equal to the number of explanatory variables in corresponding matrix l for $l = \{MH,PH,I,X,C,ER\}$. In a similar fashion, each β^l for $l = \{MH,PH,I,X,C,ER\}$ is a $k_l \times 1$ vector of parameter to be estimated.

[9] It may be argued that if a spouse is unemployed, a worker will be less likely to be absent when ill to ensure job security. However, I contend that such decisions depend on the severity of the health shock and the individual's ability to report to work—factors that are highly individual-specific.

The outcome equation is:

$$E[A_i|EMP_i > 0; \{MH, PH, J, I, X, C\}; \lambda] = \alpha_0 + MH_i\alpha^{MH} + PH_i\alpha^{PH} + I_i\alpha^I + X_i\alpha^X + C_i\alpha^C + \lambda_i\alpha^{bias},$$

for every observation, $i = 1, ..., n$. Matrices MH, PH, J, I, X, and C hold observed values of the corresponding explanatory variables for all $i = 1$, ..., n observations.

The additional conditioning variable, λ is equal to the inverse Mills ratio, and $\lambda_i\alpha^{bias}$ is the estimated bias imposed by endogenous sample selection. Each matrix of explanatory variables is of dimension $n \times k_m$ where n is the total number of observations in the sample and k_m is an integer equal to the number of explanatory variables in corresponding matrix m. Each α^m for $m = \{MH, PH, J, I, X, C, \lambda\}$ is a $k_m \times 1$ vector of parameter to be estimated. The results of the Heckman process are presented and discussed below.

Table 5 reports the estimates of the equation of the first stage of the probit model and the estimates of the marginal effect of the outcome equation for a select group of analytical variables after employing the Heckman procedure for men.

Table 5: Employment and Outcome Equation Estimates (Heckman Procedure), Men

	Labor Force		Extended Sample	
	Probit	Outcome	Probit	Outcome
keyMHdis	-0.27*** (0.04)	0.91*** (0.15)	-0.40*** (0.05)	0.22 (0.29)
physhlth	-0.92** (0.04)	0.25** (0.20)	-0.03** (0.01)	0.35** (0.03)
distress	-0.56*** (0.05)	0.33** (0.12)	-0.01** (0.004)	0.09*** (0.02)
prmrycnds	-0.05*** (0.01)	0.51*** (0.01)	-0.05*** (0.01)	0.62*** (0.06)
depout	0.45*** (0.06)		0.85** (0.10)	
spou.emp	0.19*** (0.03)		0.21*** (0.04)	
Inverse Mills Ratio	-1.32*** (0.37)		0.81*** (0.16)	
Observations	19,823	15,713	24,562	15,713

*p<0.1; **p<0.05; ***p<0.01

Table 5 reports that the coefficient for *keyMHdis* is approximately 0.10 lower than the average marginal effect (see Table 2 for these estimates), while the coefficient for *physhlth* is lower by a similar magnitude. In contrast, the coefficient for *distress* is about two times higher, and the coefficient for *prtycnds* is also higher. Both *depout* and *spou.emp* are statistically significant at the 5 percent level and have positive coefficients, indicating that external dependents and the status of spousal employment significantly influence employment decisions. All estimates for men are robust, suggesting reliable results.

Interestingly, for the Heckman model of the extended sample, the IMR is still significant but the sign is now positive. Physical health index variables and distress score variables are significantly smaller in magnitude for the selection equation for men. However, the exogenous diagnostic variables are more robust to the change in samples for the employment equation (*keyMHdis* and *prmrycnds*). In the outcome equation in Table 5, *keyMHdis* becomes insignificant.

Table 6 reports the first-stage equation estimates of the probit model and the outcome equation marginal effect estimates for a select group of analytical variables after employing the Heckman procedure for women.

The results for women, illustrated in Table 6, report that the coefficient estimate for *keyMHdis*

Table 6: Employment and Outcome Equation Estimates (Heckman Procedure), Women

	Labor Force		Extended Sample	
	Probit	Outcome	Probit	Outcome
keyMHdis	-0.02 (0.03)	0.95*** (0.15)	-0.34 (0.03)	0.68** (0.26)
physhlth	-0.79*** (0.03)	0.91*** (0.14)	-0.01 (0.01)	0.49*** (0.04)
distress	-0.12*** (0.04)	0.25** (0.20)	-0.01** (0.003)	0.14*** (0.02)
prtycnds	-0.03*** (0.01)	0.52*** (0.03)	-0.06*** (0.01)	0.52*** (0.07)
depout	0.44*** (0.05)		1.12*** (0.15)	
spou.emp	-0.46*** (0.02)		0.21*** (0.03)	
Inverse Mills Ratio	-1.85** (0.80)		0.54 (0.40)	
Observations	23,093	16,216	29,947	16,216

*p<0.1; **p<0.05; ***p<0.01

is slightly higher than AME estimates (see Table 2) but remains robust, while the coefficient for *distress* is higher by about 0.10 and also robust. The coefficient for is similarly higher and robust. Notably, the coefficient for *depout* is positive and statistically significant, highlighting the role of external dependents in increasing the likelihood of employment. Interestingly, the coefficient for *spou.emp* is negative and statistically significant, suggesting that spousal employment may reduce the likelihood of employment for women, possibly due to household labor dynamics. The Inverse Mills Ratio (IMR) is highly significant and negative, similar to the results for men but larger in magnitude, indicating a stronger correction for selection bias among women. The extended sample for the Heckman model indicates an IMR that is not significant when including women outside of the labor force.

These results reveal important gender differences in the factors influencing employment decisions. While men and women share some similarities, such as the significance of *depout* and the robustness of estimates, key differences emerge in the magnitude and direction of coefficients for variables like *distress* and *spou.emp*. The significant IMR for both genders confirms the presence of selection bias, with a stronger effect among women. These findings underscore the importance of considering sex-specific factors in labor market analyses and policy design, particularly regarding the role of external dependents and spousal employment status.

Due to the discrete nature of absenteeism outcome variable and the characteristic of overdispersion exhibited by this variable, I consider a sample selection method that allows for a probit specification in the first stage and a negative binomial specification in the second stage. The dependence parameters between the two functions, as well as the coefficient estimates, are estimated using penalized maximum likelihood.[10]

In the semi-parametric model, the dependence between the selection and outcome equations is captured by Kendall's τ and the copula parameter, θ. These values as well as coefficient estimates for analytical variables are reported in Table 7 for men. I emphasize that the results in Table 7 are based on estimates of men in the labor force, including the unemployed. Individuals outside of the labor force are not considered at this point.

For men, Kendall's τ is estimated at 0.124 (95% confidence interval (CI): 0.05, 0.20), and θ is

[10]I use the R software repository to locally rebuild the unsupported package, `SemiParSampleSel`. While some functions were not salvageable, the estimation of model coefficients for the first and second stages of an endogenous sample selection model is attainable. The Open AI source, GitHub Copilot, was utilized to edit the manual code necessary to implement these estimates.

Table 7: Estimates for Each Stage of Semi-parametric Selection Model: Men

	Probit		Outcome (Negative Binomial)	
Variables	Coefficients	Standard Errors	Coefficients	Standard Errors
keyMHdis	-0.15***	(0.04)	0.41***	(0.11)
physhlth	-0.70***	(0.04)	0.14***	(0.01)
distress	-0.81***	(0.10)	0.03***	(0.01)
prtycnds	-0.09***	(0.01)	0.20***	(0.02)
depout	0.57**	(0.06)		
spou.emp	0.28**	(0.03)		
Sample Size	19,823[1]		15,713	
Kendall's τ	0.124 (0.05,0.20)			
θ	0.194 (0.08,0.31)			

[1] Sample size includes a the sample of unemployed persons in the labor force.

0.194 (95% CI: 0.08, 0.31). For women, these estimates are higher, with Kendall's τ at 0.15 (95% CI: 0.07, 0.24) and θ estimates of 0.24 (95% CI: 0.104, 0.37). These results indicate a moderate degree of dependence between the selection and outcome processes. The estimates reported in Table 8 for the extended sample of both unemployed men in the labor force and men not in the labor force are robust when compared to results reported in Table 7.

Tables 9 and 10 indicate that the semi-parametric sample selection model estimates for women are robust to broadening the scope of the sample to include women outside of the labor force. However, the point-estimates are not as consistent as those estimated for men.

Table 8: Estimates for Each Stage of Semi-parametric Selection Model: Men – Including men out of the labor force

	Probit		Outcome (Negative Binomial)	
Variables	Coefficients	Standard Errors	Coefficients	Standard Errors
keyMHdis	-0.15***	(0.04)	0.35***	(0.05)
physhlth	-0.73***	(0.04)	0.14***	(0.01)
distress	-0.52***	(0.06)	0.03***	(0.01)
prtycnds	-0.09***	(0.01)	0.20***	(0.02)
depout	0.55***	(0.06)		
spou.emp	0.25***	(0.03)		
Sample Size	24,562[1]		15,713	
Kendall's τ	0.126 (0.04,0.20)			
θ	0.197 (0.07,0.31)			

[1]Sample size includes unemployed persons and those outside of the labor force.

The discrepancy between the IMR and the copula parameters highlights the strengths and limitations of each method. The Heckman model provides a clear and interpretable measure of selection bias through the IMR, which is robust and statistically significant. In contrast, the semi-parametric model offers flexibility in modeling the dependence structure but produces less precise estimates, as evidenced by the wide confidence intervals for Kendall's τ and θ. This implies that while the semi-parametric approach is useful for exploring the dependence structure, the Heckman

Table 9: Estimates for Each Stage of Semi-parametric Selection Model: Women

	Probit		Outcome (Negative Binomial)	
Variables	Coefficients	Standard Errors	Coefficients	Standard Errors
keyMHdis	-0.10***	(0.02)	0.28***	(0.11)
physhlth	-0.64***	(0.06)	0.09***	(0.01)
distress	-0.70***	(0.10)	0.06***	(0.01)
prtycnds	-0.03**	(0.01)	0.18***	(0.02)
depout	0.45**	(0.06)		
spou.emp	-0.12**	(0.03)		
Sample Size	23,093[1]		16,216	
Kendall's τ	0.154 (0.07,0.24)			
θ	0.24 (0.104,0.37)			

[1] Sample size includes a the sample of unemployed persons in the labor force.

Table 10: Estimates for Each Stage of Semi-parametric Selection Model: Women – Including women out of the labor force

	Probit		Outcome (Negative Binomial)	
Variables	Coefficients	Standard Errors	Coefficients	Standard Errors
keyMHdis	-0.28***	(0.04)	0.33***	(0.05)
physhlth	-0.94***	(0.04)	0.09***	(0.01)
distress	-0.76***	(0.06)	0.06***	(0.01)
prtycnds	-0.03**	(0.01)	0.18***	(0.02)
depout	0.47***	(0.06)		
spou.emp	-0.15***	(0.03)		
Sample Size	29,947[1]		16,216	
Kendall's τ	0.139 (0.05,0.21)			
θ	0.217 (0.08,0.33)			

[1]Sample size includes unemployed persons and those outside of the labor force.

model remains a more reliable tool for correcting selection bias in this context.

The significant IMR in the Heckman model, coupled with the moderate dependence indicated by Kendall's τ and θ in the semi-parametric model, underscores the importance of addressing selection bias in labor market analyses. Both methods confirm that failing to account for non-random selection can lead to biased estimates, particularly for variables that exogenously define one's health and therefore, may not be anticipated to be associated with selection bias at first. As we have seen, *keyMHdis* and *prtrycnds* show significant changes in magnitude and significance after correction.

While still up for debate, I end the empirical portion of my dissertation by acknowledging that each methods of sample selection has its benefits – the Heckman model is better suited for providing precise and actionable insights, while the semi-parametric approach offers complementary information about the underlying dependence structure.

7 Concluding Remarks

Taken together, the findings of this study highlight the significant, often hidden, economic impact of undiagnosed moderate-to-severe mental illness on labor productivity. Individuals without a formal diagnosis may lack the resources or awareness to proactively address worsening symptoms, resulting in higher absenteeism and reduced productivity. Conversely, diagnosed individuals, while potentially facing more frequent health shocks, may have greater access to interventions that help mitigate extreme levels of distress and promote workplace attendance. This dynamic underscores the importance of understanding mental health's influence on labor supply, job retention, and match efficiency within the workforce.

Findings suggest that workers with a mental health diagnosis in severe distress and fair- to-low physical health exhibit better productive outcomes compared to similar but undiagnosed peers. This supports the notion that employers can mitigate productivity losses by promoting working wellness and by adopting health policies that encourage early treatment and support for mental health. Notably, AME estimates suggest that absenteeism for women in the sample may be more sensitive to presently observed health shocks than men. This may be due to a multitude of reasons, including differences over preferences for the sexes. Also of note is that when including the interaction of diagnosed mental illness and self-reported physical health status, women do not exhibit clear evidence that a diagnosed mental illness eventually makes them better off at poor levels of self-reported physical health.

The number of chronic physical health conditions is found to significantly influence absenteeism schedules for both men and women when comparing across groups of diagnosed mental illness and no such diagnosis. This finding is consistent with the logic surrounding the argument that adverse health effects may pose less harm to the labor market outcomes of individuals who have knowledge of their diagnoses and risks; this is especially true if an individual has a typical healthcare provider, which reduces the search costs associated with improving health through the route of utilizing healthcare. This is an area of interest for future research that addresses these relationship more clearly by modeling structural equations of health rather than the reduced form focused on in this paper. Of final note is the consideration of whether individuals facing multiple chronic physical conditions on top of mental illness may supply less labor in general or may be on disability leave that technically may not count as short-run absence in the current context, which can be studied in the future.

This paper has found empirical evidence that diagnosed mental illness leads to deviations from optimal labor supply, particularly in the short run, impacting job performance and, consequently, the labor market. Furthermore, individuals with a known diagnosis may face challenges in job matching and retention due to structural biases and employers' concerns over attendance stability. These factors contribute to limited productive capacity within the labor force and inefficient employment outcomes. Future research into the relationship between mental health, productivity, and job stability is critical, especially in today's economic environment, where high rates of mental illness and inadequate access to mental health resources continue to affect millions. A deeper understanding of these dynamics will inform policies and workplace practices that support better health outcomes and foster a more resilient, productive labor force.

Total Sample Summary Statistics and Variable Definitions

Table 11: Total Sample Summary Statistics

Variable	Description	Mean	Min	Max	SD
Dependent Variable (A):					
sickdays	Dependent variable. Count of the total days an individual has been absent from work due to illness or injury in the past year.	3.18	0	160	9.97
Mental Health (MH):					
keyMHdis	=1 if individual has a diagnosed mental disorder of interest, =0 otherwise.	0.10	0	1	0.30
distress	Discrete scale from 0 to 24 with higher scores indicating greater emotional distress.	2.60	0	24	3.57
adhd	=1 if individual reports an ADHD diagnosis, =0 otherwise.	0.01	0	1	0.08
Physical Health (PH):					
physhlth	A discrete scale score with range 1 to 13 with higher scores indicating poorer general health.	4.66	1	13	2.40
prtycnds	The number of priority condition diagnoses.	1.31	0	9	1.45
routine	=1 if the individual received a routine medical evaluation in the past year, =0 otherwise.	0.03	0	1	0.16
injury	=1 if individual suffered an injury or illness requiring immediate medical care in the past year.	0.24	0	1	0.43

(Continued on following page)

Table 11 Continued					
Variable	Description	Mean	Max	Min	SD
smoke	=1 if individual smokes cigarettes, =0 otherwise.	0.17	0	1	0.38
exercise	=1 if individual exercises at least 3 times per week, =0 otherwise.	0.53	0	1	0.50
pregnt	Sub-sample of women only. =1 if female was pregnant at any point during the year.	0.03	0	1	0.18
Job Traits (J):					
$selfemp$	= 1 for self employed individuals, =0 otherwise.	0.13	0	1	0.02
$one4$	Benchmark group. =1 for individuals with tenure between 1 and 4 years, =0 otherwise.	0.37	0	1	0.48
$five14$	=1 for individuals with tenure between 5 and 14 years, =0 otherwise.	0.35	0	1	0.48
$fifteen24$	=1 for individuals with tenure between 15 and 24 years, =0 otherwise.	0.11	0	1	0.32
$25plus$	=1 for individuals with tenure of 25 years or more, =0 otherwise.	0.06	0	1	0.24
temp	=1 if individual has a temporary employment contract, =0 otherwise.	0.05	0	1	0.22
parttime	=1 if individual reports working 35 hours per week or more, =0 otherwise.	0.17	0	1	0.37

(Continued on following page)

Table 11 Continued					
Variable	Description	Mean	Max	Min	SD
union	=1 if individual is part of a labor union, =0 otherwise.	0.13	0	1	0.33
ssnl	=1 if individual's main job is a seasonal positions, =0 otherwise.	0.05	0	1	0.21
pubsect	=1 if individual works in the public sector, =0 otherwise.	0.18	0	1	0.38
*occ*1	Management, business, and financial operations.	0.13	0	1	0.33
*occ*2	Professional and related occupations.	0.22	0	1	0.41
*occ*3	Service occupations.	0.19	0	1	0.39
*occ*4	Sales and related occupations.	0.08	0	1	0.27
*occ*5	Office and administrative support.	0.14	0	1	0.35
*occ*6	Farming, fishing, and forestry	0.01	0	1	0.10
*occ*7	Construction, extraction, and maintenance.	0.08	0	1	0.27
*occ*8	Benchmark group. Production, transportation, material moving.	0.15	0	1	0.36
*occ*9	Unclassifiable occupation.	0.004	0	1	0.06

(Continued on following page)

Table 11 Continued					
Variable	Description	Mean	Max	Min	SD
$1to19$	=1 if firm has between 1 and 19 workers.	0.30	0	1	0.46
$20to99$	=1 if firm has between 20 and 99 workers.	0.33	0	1	0.47
$100to499$	=1 if firm has between 100 and 499 workers.	0.21	0	1	0.41
$500plus$	Benchmark group. =1 if firm has 500 or more workers.	0.16	0	1	0.37
$NRunemp$	=1 if individual did not respond to questions pertaining to firm size.	0.05	0	1	0.22
Health Insurance (I, $\tilde{I}$):					
$inscostly$	=1 if individual either somewhat agrees or strongly agrees with the statement,"health insurance is not worth the cost", =0 otherwise.	0.28	0	1	0.45
$jobins$	=1 if individual is insured through their job, =0 otherwise.	0.63	0	1	0.48
$plnchoic$	=1 if insured through job AND selected from a catalog of plans, =0 otherwise	0.34	0	1	0.47
$nochoic$	=1 if insured through job AND was only eligible for one plan, =0 otherwise.	0.24	0	1	0.43
$NR.choic$	=1 for observations reporting coverage through job but no response about whether they had a choice of plans, =0 otherwise	0.05	0	1	0.22

(Continued on following page)

Table 11 Continued					
Variable	Description	Mean	Max	Min	SD
otherins	=1 if individual has private health insurance coverage through a non-job source, =0 otherwise.	0.13	0	1	0.34
pubins	=1 if individual has health insurance coverage through a public source, =0 otherwise.	0.07	0	1	0.25
unins	Benchmark group. =1 if individual is uninsured, =0 otherwise	0.17	0	1	0.38
Demographics (X):					
poor	=1 if household income as a % of poverty line puts them into "poor" or "near poor" groups.	0.14	0	1	0.34
lowinc	=1 if household income as a % of poverty line puts them into "low income" group.	0.16	0	1	0.36
midinc	=1 if household income as a % of poverty line puts them into "middle income" group.	0.34	0	1	0.47
highinc	Benchmark group. =1 if household income as a % of poverty line puts them into "high income" group.	0.37	0	1	0.48
married	=1 if individual is married, =0 otherwise.	0.55	0	1	0.50
$famsz$	Number of individuals within the surveyed household.	3.02	0	14	1.66

(Continued on following page)

Table 11 Continued					
Variable	Description	Mean	Max	Min	SD
yngchldrn	Number of children aged 6 and under.	0.38	0	6	0.71
age	Age in years.	41.86	18	84	11.82
belowhs	Individuals with less than high school education.	0.13	0	1	0.34
hsdeg	Benchmark group. Individuals with a high school degree or GED.	0.32	0	1	0.47
somecoll	Individuals with some college or an associate's degree, but no 4-year degree.	0.25	0	1	0.43
bachdeg	Individuals with a bachelor's degree.	0.20	0	1	0.40
bachplus	Individuals with schooling beyond a bachelor's degree.	0.10	0	1	0.31
hispanic	=1 if individual is Hispanic, =0 otherwise	0.28	0	1	0.45
black	=1 if individual is Black, =0 otherwise	0.18	0	1	0.38
asian	=1 if individual is Asian, =0 otherwise	0.08	0	1	0.28

(Continued on following page)

Table 11 Continued					
Variable	Description	Mean	Max	Min	SD
bornus	=1 if individual was born in the US.	0.65	0	1	0.48
Other Controls (C):					
unemp.1	=1 if individual reports unemployment at exactly one of the three interviews.	0.07	0	1	0.26
unemp.2	=1 if individual reports unemployment at exactly two of the three interviews.	0.06	0	1	0.23
partialemp	= 1 if an individual is unemployed at one of the three interviews, but worked for at least part of the reference period prior to employment termination.	0.04	0	1	0.20
empUB	The total number of days for which an individual was employed during the year.	338	365	28	0.19
moved_US	=1 if individual moved within the US during the calendar year.	0.03	0	1	0.18
moved_RU	=1 if individual joined a new reference unit (household) at any point during the calendar year.	0.01	0	1	0.10
NE	=1 if individual resides in the Northeastern region for most of the year.	0.16	0	1	0.36
MW	=1 if individual resides in the Midwestern region for most of the year.	0.20	0	1	0.40

(Continued on following page)

Table 11 Continued					
Variable	Description	Mean	Max	Min	SD
W	=1 if individual resides in the Western region for most of the year.	0.28	0	1	0.45
S	Benchmark group. =1 if individual resides in the Southern region for most of the year.	0.37	0	1	0.48
unemp.rt	The average annual unemployment rate faced by the individual estimated using monthly regional unemployment reports from the Bureau of Labor Statistics and region of residence reported for each round.	8.05	5.5	10.9	1.29
yearone	=1 if individual is only observed for the first year of their designated panel.	0.26	0	1	0.44
yeartwo	=1 if individual is only observed for the second year of their designated panel.	0.20	0	1	0.40
bothyrs	Benchmark. =1 if individual is observed for both years of their panel.	0.54	0	1	0.50
Exclusion Restrictions (ER:)					
depout	=1 if dependent(s) reside outside of household, =0 otherwise.	0.04	0	1	0.01
spou.emp	=1 if individual claims dependents residing outside of the individual's lace of residence.	0.84	0	1	0.12

Table 12: Means of Analytical Variables by Sex

	Variable Name	Men	Women	T Stat
Dependent Variable (A)	*sickdays*	2.50	3.85	-12.11***
Mental Health (MH)	*distress*	2.30	2.90	-15.02***
	physhlth	4.44	4.87	-16.01***
	prtycnds	1.26	1.35	-5.57***
Job Characteristics (J)	*sickpay*	0.61	0.64	-5.04***
	bonus	0.22	0.18	9.92***
Health Insurance Characteristics (I, $\tilde{I}$)	*inscostly*	0.30	0.26	7.69***
	jobins	0.66	0.61	9.38***
	plnchoic	0.34	0.34	-0.30
	nochoic	0.26	0.22	8.76***
	otherins	0.11	0.15	-12.01***
	pubins	0.05	0.09	-14.71***
	unins	0.19	0.15	8.57***
Observations		15,713	16,216	

Note: *p<0.1; **p<0.05; ***p<0.01

Table 13: Means of Analytical Variables by Sex and Presence of Mental Illness

	Men			Women		
Variable Name	$keyMHdis$ =1	$keyMHdis$ =0	T Stat	$keyMHdis$ =1	$keyMHdis$ =0	T Stat
sickdays	4.64	2.35	6.35***	6.14	2.48	9.02***
distress	5.33	2.09	20.53***	5.57	2.46	28.44***
physhlth	5.59	4.36	15.34***	5.78	4.72	18.50***
prtycnds	2.04	1.21	16.47***	1.96	1.25	19.24***
sickpay	0.64	0.61	2.45**	0.66	0.63	2.10**
bonus	0.26	0.22	2.40**	0.19	0.18	1.90*
inscostly	0.27	0.30	-1.61	0.22	0.26	-4.96***
jobins	0.71	0.65	3.87***	0.63	0.60	2.70***
plnchoic	0.38	0.33	3.29***	0.36	0.33	2.87***
nochoic	0.27	0.26	0.47	0.24	0.22	1.63
otherins	0.10	0.11	-0.64	0.16	0.15	1.43
pubins	0.05	0.05	0.40	0.10	0.09	1.42
unins	0.14	0.19	-4.75***	0.11	0.16	-7.05***
Observations	15,713			16,216		

Note: *p<0.1; **p<0.05; ***p<0.01