Social Workers and Suicidality in Jail: Evidence from Travis County's Mental Health Court*

Jonathan Seward (r) Vivian S. Vigliotti (r)

Baylor University Baylor University

Scott Cunningham (r)

Baylor University

November 2021

Abstract

Suicide is the leading cause of deaths in jails partly due to the high prevalence of mental illness within correctional facilities. The mental health court in Travis County, Texas, assigns people with mental illness to one of two types of indigent defense attorneys within its mental health court: lawyers who employ social workers or those who do not. We instrument for assignment to lawyers-with-social-workers using randomized clinical evaluators' assessment of an inmate's functioning at booking with a leniency design. We find that lawyers-with-social-workers reduce suicide attempts by 9-22%, reduce self-reported suicidal ideation by 1-2%, and improve subsequent functioning scores by 0.5 to 1 points on a 4 point scale for those who re-offend. Marginal treatment effect analysis suggests that the impact is larger for those with higher unobserved mental illness. Finding no effect on recidivism and

*Authorship has been randomized (r). We thank Emily Leslie, Martin Andresen, and Vitor Possebom for helpful input and advice specifically with respect to the monotonicity and marginal treatment effects discussions. We also thank attendees at University of Pittsburgh, Loyola Marymount University, Erasmus University, Cornell University, and the NBER Health Economics group for helpful feedback on an earlier version of this project. For questions or comments please contact Vivian S. Vigliotti at vivian_vigliotti1@baylor.edu, Jonathan Seward at jonathan_seward@baylor.edu, or Scott Cunningham at scott_cunningham@baylor.edu

improved mental health only for those without a history of prior mental illness suggests that social workers, and not lawyer selection, are responsible for improvements in defendant mental health.

Keywords: mental health court; lawyers; suicidality; social workers; instrumental variables

1 Introduction

Suicide in jail is all too common in the US criminal justice system. From 2001 to 2016, it accounted for 6% of all deaths in US state prisons and nearly a third of all jail deaths, making suicide the leading cause of death in US jails and the second leading cause of death in Texas jails (Dillon, 2013; Carson and Cowhig, 2020a,b).[1] Its occurrence is believed to be linked to the high prevalence of mental illness in the inmate population itself.[2] Many counties have adopted diversion courts that redirect individuals with mental illness away from correctional facilities altogether to address the growing share of people with mental illness caught in the criminal justice system. These diversion courts are often called mental health courts, and there are hundreds across the US.

Travis County, the county seat of Austin, Texas, uses its mental health court to redirect mentally ill defendants into treatment and support. Inmates are classified into four categories based on the impact that their mental illness has on daily functioning and then, if eligible for mental health court, to one of two types of attorneys. Individuals displaying the most severe functioning problems are assigned to lawyers-with-social-workers (public attorneys) and those with only moderate functioning problems are assigned to lawyers-without-social-workers (private attorneys from Travis County's "wheel" attorney system).[3] Those with no or only mild functioning problems remain on track to traditional adjudication. Differences in access to social workers across the two types of attorneys are due entirely to the constraints created by Travis County's budget.[4]

[1] In the US in 2019, there were 355 deaths from suicide in local jails.

[2] A previous study found a 77% prevalence rate of mental illness among inmates who attempted suicide (Goss, 2002). Additionally, mental illness is associated with suicide more generally, and almost 90% of all suicides occur with a co-morbid mental illness (Stack, 2014).

[3] The National Drug Use and Health survey in 2008 noted that "Serious mental illness (SMI) among adults is defined in Public Law 102-321 as persons aged 18 or older who currently or at any time in the past year have had a diagnosable mental, behavioral, or emotional disorder (excluding developmental and substance use disorders) of sufficient duration to meet diagnostic criteria specified within DSM-IV that has resulted in functional impairment, which substantially interferes with or limits one or more major life activities. In 2008, there were an estimated 9.8 million adults with SMI, representing 4.4 percent of adults." See `https://dpft.org/resources/NSDUHresults2008.pdf` for more information about the prevalence and experiences of SMI populations.

[4] Wheel attorneys are paid $750 per case. The flat fee does not vary by marginal hour thus potentially blunting wheel attorney effort at the intensive margin. The public defenders office employs two social workers for every one public defender, whereas private indigent defense attorneys are not assigned social workers and typically do not use them due to the low nominal fee paid for a case.

Social workers in Travis County's mental health court help their clients go to scheduled court appointments, maintain their treatment regimen, sign up for social services, find housing, and many other important aspects connected to rehabilitation and improvement. However, scholars acknowledge that it is unknown how or if social workers affect their clients' suicidality (Joe and Niedermeier, 2008). Maple et al. (2017) reviewed 241 papers in a meta-analysis and could not find any evidence that social workers affected suicide risk, either within or outside the correctional setting. Testing this in our context is difficult because the comparisons between lawyers-with-social-workers and lawyers without are contaminated by selection bias since selection into treatment is directly associated with underlying latent mental illness and therefore potential suicide risk. Our contribution to this question is therefore to estimate the causal effect of social worker assignment on suicidality in jails using a leniency design with randomized clinical evaluators as instruments for lawyers-with-social-worker assignment (Imbens and Angrist, 1994).[5]

We present a visual depiction of the sequence of events from booking to counsel assignment that informs our research design in Figure 1. The first encounter an inmate has after their arrest is with a randomized clinical evaluator. Evaluators in our sample are mostly clinical social workers or counselors employed by the correctional facility to evaluate inmate functioning on a high-volume basis. Each day, evaluators are alphabetized by last name and then assigned to each inmate as they arrive, which in turn randomizes clinical evaluators across inmates conditional on shift. The evaluator's only job is to score the severity of the inmate's daily functioning problems.[6] Functioning scores are ordinal and range between 0 (no problems with daily functioning) and 3 (severely low daily functioning). Inmates with a score of 2 or 3 are assigned to the mental health court. Individuals with a 2 are assigned to our control group, which we call lawyers-without-social-workers (wheel attorneys), and those with a 3 are assigned to our treatment group, which we call lawyers-with-social-workers (public attorneys). Using the evaluator's resid-

[5]Regression discontinuity is not possible in our context because the running variable for assignment is only two values: lawyers-with-social-workers or lawyers without.

[6]As this point is crucial for the argument that exclusion holds in our sample, we have had extensive conversations with the director of inmate mental health at Travis to confirm that the evaluation score is not used for treatment in jail. Our contacts within the jail have told us that unambiguously the scores are not used to assign treatment in the jail at the range we are investigating. The only lasting effect of the score in our data is court and lawyer assignment.

ualized leave-one-out mean "score of 3" (hereafter: high score) as an instrument for an inmate's actual score, we estimate the local average treatment effect (LATE) of lawyers-with-social-workers on recidivism, subsequent functioning, and suicidality in jail. We also estimate marginal treatment effects (MTE) to examine whether there exists unobserved heterogeneity in treatment effects as well as to construct more interesting policy-relevant parameters such as the average treatment effect (ATE).

Insofar as our lawyers-with-social-workers cause differential reoffending, then our sample of inmates will suffer from sample bias due to colliders associated with sample selection (Schneider, 2020; Cunningham, 2021). As part of a check on this problem, we estimated whether inmates assigned to lawyers-with-social-workers were more likely to reoffend but could not find evidence for this. Lawyers-with-social-workers in our sample do not appear to cause differential reentry into jail, which suggests our dataset may not suffer from this type of collider problem.

For our main mental health outcomes, we estimate a range of IV models acknowledging that our baseline two-stage least squares (2SLS) model may be biased given that our instrument is many variables collapsed into a single scalar. Since the instrument is in fact a vector of evaluator fixed effects, it is equivalent to an over-identified first stage which can amplify the bias of 2SLS if any instruments are weak and exclusion is not perfectly satisfied. We address the possibility of finite sample bias by exploring four more appropriate IV models: the jackknife IV estimator (JIVE), the unbiased jackknife estimator (UJIVE) and two machine learning based instrument selection models (LASSO and post-LASSO). Our results are robust to all of these modeling choices. Based on our 2SLS results, for instance, we find that lawyers-with-social-workers reduce self-reported suicidal ideation by 1-2% and suicide attempts by 9%. Our other IV models suggest the effect on suicide attempts may be as high as 22%. Lawyers-with-social-workers also appear to cause substantial gains in functioning upon re-entry by between 0.5 to 1 point improvements on the jail's 4-point scoring system.

We explore possible mechanisms to explain these results and believe that evidence is tipped towards the social worker component of the treatment and not the lawyer selection component. For instance, as explained, lawyers-with-social-workers do not differentially

impact reoffending. The effects are only on mental health. Second, the effects only exist for those inmates without a prior diagnosis or history of mental illness. While speculative, we interpret these patterns as consistent with a hypothesis that social workers are improving defendants' mental health by helping them to receive much-needed treatment as well as connecting them to social services.

Our results show signs of interesting heterogeneities, though, as we move from the highest functioning inmates to the lowest. Estimated marginal treatment effects form a parabola across inmate propensity scores from the highest to lowest functioning inmates. Treatment effects grow larger as we move from higher functioning to lower functioning inmates but then fall when we near the lowest functioning group. Medical noncompliance, homelessness, and other problems exacerbate the struggles of seriously mentally ill individuals, which may explain why MTE are not monotonic across this population. But despite smaller MTE for the lowest functioning inmates in our sample, lawyers-with-social-workers still improved subsequent inmate functioning at a level comparable to our main LATE estimates.

2 Background

The US reduced the number of psychiatric beds in residential mental hospitals at the same time as it escalated arrests, sentencing, and imprisonment of criminal offenders (Harcourt, 2006; Western, 2006; Raphael and Stoll, 2013; Neal and Rick, 2014). Due to the timing of both, growth in the correctional populations may have selected on individuals with mental illnesses (Harcourt, 2006; Raphael and Stoll, 2013). As a result, our correctional facilities have a large population of individuals with mental illnesses. Here we discuss that history, including the emergence of mental health courts, indigent defense, and the problem of suicides in jails and prisons.

2.1 Emptying Psychiatric Beds and Rising Correctional Facility Populations

Prior to the 1960s, the US used state-funded mental hospitals to treat serious mental illness. These hospitals were primarily staffed with custodians, maintained poor condi-

tions, and employed few psychiatrists. Complaints about human rights violations within the hospitals led to a movement for restoring the civil liberties of people with a mental illness. This movement eventually resulted in a successful effort to defund the mental hospitals, which caused a decline in the number of psychiatric beds available in the US. At the time of these policy debates, it was hoped that community-based alternatives to residential hospitalization would materialize, but scholars such as Frank and Glied (2006) suggest that the response was inadequate.

In 1963, President Kennedy established the first of US community-based mental health centers. These centers improved options for treating mental health disorders outside of the residential hospital setting. Combined with the establishment of Medicare and Medicaid in the 1960s, a diversity of mental healthcare options grew in the US, which caused a decrease in mental health hospital populations, particularly the public ones. The creation of Section 8 housing only intensified this process as many of the mentally ill, when symptoms are severe, experience homelessness. The discovery of several pharmacological breakthroughs for treating schizophrenia (Clozaril), bipolar disorder (Lithium), and depression (selective serotonin re-uptake inhibitors (SSRIs)) may have contributed to declining demand for mental health residential hospitalization as well (Biasi, Dahl and Moder, 2020).

The timing of the US experiment with deinstitutionalization was not ideal. As the US moved away from residential hospitals and towards community-based treatment for mental illness, the "war on drugs and crime" was escalating (Western, 2006; Neal and Rick, 2014). This effort to punish crime more intensively caused prison and jail populations to rise steeply. From the mid 1970s to 2000, the number of US citizens living in correctional facilities grew from 100 per 100,000 to over 500 per 100,000 (Patillo, Weidman and Western, 2004; Western, 2006). The coincidental decline in residential hospitalizations with the increase in mass incarceration may have led to movements between psychiatric hospitals and criminal justice facilities. Raphael and Stoll (2013) found that the estimated causal effect of mass incarceration resulted in a decline in the number of mentally ill individuals living openly in society. Torrey et al. (2014) estimates that there may be as many as 20% of inmates in jails suffering from a serious mental illness. Frank and McGuire (2010) notes that a serious mental illness is associated with a 58% higher lifetime risk of arrest

conditional on age, gender, and race.

As a result of the emptying of psychiatric beds and the growth in mass incarceration, US county jails have become like temporary homes for millions of individuals who cycle through corrections suffering from mental illness or a behavioral disorder due to gaps in healthcare coverage, limited access to behavioral healthcare, medical noncompliance and more (Center for Substance Abuse Treatment, 2005). In most states, there is at least one jail or prison that houses more mentally ill individuals than the largest psychiatric hospital in the area (Torrey et al., 2014). Within every US county that has both a county jail and a county psychiatric facility, there are more seriously mentally ill individuals housed in the local jail than there are hospitalized in the local psychiatric facility (Torrey et al., 2014). Consequently, ten times more individuals with serious mental illness are in jails and state prisons than in the remaining state mental hospitals (Torrey et al., 2014).

2.2 Therapeutic Justice and the Mental Health Courts

A new therapeutic movement emerged at the end of the 20th century designed to meet the needs of counties with large mentally ill populations entangled in criminal justice. This jurisprudence-based therapeutic movement led to the creation of a new model of diversion known as the mental health court. There are now hundreds of mental health courts across the US.

While there is substantial variation across US mental health courts, in general, these courts have: (1) a specialized docket of cases of defendants with a mental illness; (2) a collaborative and non-adversarial team comprised of a judge, prosecuting and defense attorneys, and a mental health representative; (3) a link to a local mental health system; and (4) compliance monitoring with sanctions for non-compliance (Wolff, 2002). Clients of mental health courts may forgo criminal processing (i.e., they are not prosecuted) altogether, undergo criminal processing (i.e., they are prosecuted on criminal charges) but forgo sentencing, or receive an alternative favorable sentence for participating and completing the mental health court program (Goldkamp and Irons-Guynn, 2001; Steadman, Davidson and Brown, 2001; Watson et al., 2001).

The eligibility criteria for mental health courts typically requires that defendants have

a mental illness (serious, chronic, or persistent) and non-violent criminal charges that are most often classified as a misdemeanor (Wolff, 2002; Wolff and Pogorzelski, 2005). Potential defendants could be referred to the mental health court by a prosecutor, defense attorney, family member, treatment provider, judge, jail personnel, police officer, and others (Goldkamp and Irons-Guynn, 2001; Redlich et al., 2010). They may be screened formally by the court team or a case coordinator with mental health training using a screening protocol (Wolff, 2002).

Some of the variation in the client pools occurs because of eligibility requirements mandated by criminal justice personnel or the court itself, poor or inconsistent program identification or recruitment practices, mixed or variable support among team members, improper matching of services to the target population, or other forms of structural elements that restrict or expand the identification, recruitment, or selection of seemingly eligible clients. In addition, courts may use different incentives and disincentives to encourage participation in the court and, as such, introduce selection bias (Redlich et al., 2010). These motivational inducements may foster or hinder the interests of potential clients and the support from defense attorneys (Wolff, 2002).

The collision of all these factors predicts variation in client pools among and between mental health courts, as well as variation in mental health court treatment categories, which can create problems for interpreting estimated treatment effects. There are hundreds of highly unique mental health courts due to the lack of a single recommended model, including variation in size and client pool. Furthermore, not all mental health courts employ both public and private indigent defense. Thus comparisons across courts are likely to create interpretability problems due to the considerable hidden variations of treatments across courts (a violation of SUTVA) (Imbens and Rubin, 2015).

2.3 Travis Mental Health Court

Travis County is the county seat of Austin, Texas. It is home to a large correctional complex that attempts to meet the detention demands of its 1.2 million residents. The jail is a highly sensitive and unusual work environment that requires extensive training and a very unique skill set. Approximately 20% of the inmate population at Travis

County's jail requires treatment for mental illness. On any given day, approximately 7% of inmates with mental illness are experiencing severe symptoms such as psychosis, delusions or suicidal thoughts. It is incumbent upon the large urban county's sheriff's office to make every effort possible to ensure their employees are equipped to serve this vulnerable population.

The Travis County mental health court has both unique features and features common to all mental health courts. The court's purpose is, like other mental health courts, to defer charges or punishments so long as defendants participate in services to treat their mental illness. The hope is that these courts can disentangle defendants from the criminal justice system. The foundation of this hope is the provision of critical mental health services that mentally ill defendants otherwise would not receive (Watson et al., 2001).

The purpose of the county's Misdemeanor Mental Health Diversion Docket is to provide court supervision for defendants diagnosed with mental illness who have entered an agreement with the State to have their criminal case dismissed after a period of treatment and stability.[7] Defendants are released on personal bond with conditions that are agreed to by the State and are supervised by specialized pretrial service officers. Defendants report monthly until their case is dismissed. The eligibility criteria includes: mental health diagnosis, pending misdemeanor offense, and approval by prosecution. For our county, eligibility criteria includes that the defendant is also experiencing significant challenges due to mental health, intellectual, or developmental disabilities.

In addition to the county possessing a mental health court, it also staffs two different indigent defense attorneys: a public defense office and a pool of moonlighting private indigent defense attorneys. In Travis County, the most substantial difference between the two pools of lawyers is the employment of social workers. Due to resource constraints, Travis's public defenders office is the only pool of lawyers-with-social-workers, with two social workers for every one public defender. Public defense attorneys collaborate with these social workers to implement therapeutic interventions that may include medication

[7]Much of the authors' knowledge of the institutional details of the docket come from extensive interviews with the public defenders office, prosecutors, representatives of the private attorney wheel and judges, the director of inmate mental health at Travis County's correctional facility from whom we received the data used in this study as well as studying writings describing the Misdemeanor Mental Health Diversion Docket more generally.

management, substance abuse treatment, securing of housing, job training, psychosocial rehabilitation, and enrollment in various public programs such as Social Security or SNAP. But due to the tight county budgets, these resources are not extended to private attorneys, in part because private attorneys are assigned clients with slightly higher functioning. This is a source of contention with the wheel, though, because they believe they do not have adequate resources to assist their clients.

2.4 Suicide in Jails

Self-harm and suicide attempts in jail and prison represent a double tragedy: human life is jeopardized or lost and the correctional facility failed to protect the inmate. The World Health Organization explains that suicide is a serious public health problem (WHO, 1999), yet the feasibility of suicide prevention must involve many moving parts, including effective treatment of mental disorders and environmental control of risk factors (WHO, 2007). There are various mental health treatment and prevention methods utilized by correctional facilities to decrease self-harm and suicide attempts, but suicide is complex, and predicting or evaluating suicide risk is extremely difficult (Turecki and Brent, 2016).[8]

Suicidal behaviors are more common among those who get imprisoned, resulting in pre-trial detainees having a suicide attempt rate of almost eight times that of the general population (Jenkins et al., 2005). The causes of suicide to persons in custody are difficult to understand since those who break the law inherently have many risk factors for suicide before, during, and after release (Pratt et al., 2006). Any combination of individual and environmental factors may account for the higher rates of suicide in correctional facilities, such as the psychological impact of arrest and incarceration, symptoms of withdrawal experienced by drug addicts, expected long prison sentences, the overall stress of being associated with prison life, poor or no access to mental health professionals or treatments, lack of formal policies and procedures to identify and manage inmates at risk of suicide, individuals with mental disorders, substance or alcohol abuse, socially isolated or socially

[8]Many factors that may influence suicide risk vary across correctional settings including short-term detention, pre-trial offending, sentenced prisoners, harsh sentencing practices, overcrowding, purposeful activity, times spent locked up, sanitation, sociocultural conditions, levels of stress, and access to basic health and mental health services, among other aspects (Huey and McNulty, 2005; Leese, Thomas and Snow, 2006).

disenfranchised individuals, among many other factors (WHO, 2007).

Environmental factors and interventions are critical to consider when assessing risk of self-harm and suicide attempt within criminal justice facilities (Marzano et al., 2016). One study found that the importance of talking with someone was helpful in decreasing self-harm and suicide attempts (Borrill, 2002). More time out of one's cell and/or sharing a cell with another prisoner both were discussed as helpful remedies (Borrill, 2002). One case-control study from Austria identified four specific individual risk factors believed to predict suicidality. These included previous suicide attempts, psychiatric diagnosis, psychotropic medication prescribed during imprisonment, and highly violent index scores. One environmental risk factor was single-cell accommodation. All of these and others unexplored may be used to better understand who may be at risk for suicide within jails and prisons (Fruehwald et al., 2004).

Situational factors that contribute to suicide in correctional facilities are isolation or segregation cells and times when staffing is low, such as nights and weekends. Several factors affect housing assignments/arrangements within correctional facilities, such as capacity, staffing, availability of appropriate facilities, and more. Housing is also widely used as a measure of supervision. For example, certain housing arrangements facilitate specific supervision from officers so that inmates are checked on at decreased time intervals as most suicides in jails and prisons occur by hanging. Housing arrangements have a strong association with inmate suicide, especially when an inmate is placed somewhere they are unable to cope. Such housing assignments can result in an inmate being inside his or her cell for as many as 23 hours per day, which is also associated with greater risk (WHO, 2007). Poor social and family support, history of psychiatric illness and emotional problems, and a prior history of suicidal behavior are all common among inmate suicides. Individual stressors and vulnerabilities, resulting from bullying, recent inmate-to-inmate conflicts, adverse information, or disciplinary infractions, lead to inmates feeling hopeless, with narrowing future prospects, and loss of coping options, which ultimately leads to suicide attempts. Furthermore, as length of stay increases, so does suicide rates for long-term inmates, with "lifers" having especially high-risk for suicide (Blaauw and Winkel, 2001; Frottier et al., 2002; Way et al., 2005; Borrill, 2002; Liebling and Ludlow, 2016).

3 Data, Design and Estimation

3.1 Administrative Data from Travis County Correctional Complex

Our data is from the Travis County correctional complex's administrative records, which encompasses every inmate booked between 2016 and 2019. Travis is a large urban county home to over 1.2 million residents. These data were collected as routine mental and physical health assessments on inmates.[9] These administrative data include information on each inmate's offense type (felony, misdemeanor), demographics, mental health records, charges, suicide attempt, suicide ideation, and more. A unique inmate ID and unique booking ID are jointly used to identify a unique inmate booking event, as well as link inmates over time.

We restrict our sample to inmates who were assigned to the mental health courts based on their heightened symptoms at the time of initial assessment, which reduces our sample to 5,222 observations, with 928 in our treatment group. Table 1 reports summary statistics for these individuals by treatment category. These courts differ considerably along observable dimensions. For example, individuals with moderate symptoms, who were assigned to the private indigent attorneys, had higher suicidality and showed fewer signs of improvement at followup (43% improved versus 55% in our treatment group).

The two groups differed along demographic lines as well. The private attorney group was 73% White people versus 70.4% in the public attorney group. Private attorney inmates were less likely to be Black, tended to be younger, had fewer prior recent offenses, and had a higher share of females than those in the public attorney group. They were more likely to have received mental health treatment and were more likely to be homeless and jobless than those in the public defender group. Clinical evaluator characteristics also differed by group with private attorneys having fewer male evaluators, fewer White evaluators, more Black evaluators, and more Hispanic evaluators.

[9]Institutional review board (IRB) approval was granted from Baylor University in April 2019.

3.2 Randomized Evaluators and Treatment Assignment

Upon booking into Travis County's correctional complex, inmates are met by an officer who makes a cursory check about whether an inmate has any signs of mental illness. This decision is based on whether the individual has a history of mental illness, has ever taken medication, or whether the officer believes the person is showing signs of mental illness. If any of these criteria are met, the inmate will meet with a evaluator to have their symptoms evaluated in greater detail.

Travis uses random assignment of evaluators to inmates due to the large number of inmates meeting criteria for subsequent evaluation and the need to balance the workload. At any point in time, the Travis County correctional complex employs approximately 60 evaluators. The vast majority of these evaluators are clinical social workers and professional counselors. Their interest in working for the correctional complex is sometimes due to the generous benefits of the county and an enjoyment of the work, as well as seeking the hours needed by the state's licensing board.

Inmates are assessed according to their activities of daily living, or ADL. ADL is a term used to describe a person's set of skills considered to be fundamental to independent living and caring for oneself. These skills include such tasks as eating, bathing, and moving around. ADL is used by Travis County's correctional complex as an indicator of a person's functional status. Using a structured survey in combination with their professional judgment, a randomly assigned clinical evaluator meets with the inmate to assess their ADL. The complex uses a scale of 0 to 3 to assess the severity of the inmate's problems functioning. Inmates with a 0 (no perceived problems impeding ADL) or 1 (only mild symptoms) do not meet criteria for the county mental health court and so remain on the normal track into typical courts.

Inmates assessed as a 2 (moderate symptoms impeding ADLs) are assigned to a private attorney by the court for indigent defense. These private attorneys are paid a nominal flat fee of $750 which does not vary with the number of hours devoted to the defendant's case.[10] Inmates with a 3 (severe symptoms impeding ADLs) are perceived as unusually low

[10]Theoretically, since payment is a fixed and low nominal fee, private indigent defense attorneys appointed by the court have distorted incentives associated with representation. For instance, they are not

functioning and are redirected to the county public defender's office. Our sample consists only of those individuals who received either a 2 or 3 in their booking assessments which we consider to be "low" and "high" scores, respectively.

3.3 Identification in Leniency Designs

Leniency designs were first suggested by Imbens and Angrist (1994) as a potential instrumental variables strategy using decision-makers' average tendencies to recommend some treatment under random assignment. The design is commonly used in criminal justice studies and has been used to study the consequences of Chapter 13 bankruptcy on future financial events (Dobbie, Goldsmith-Pinkham and Yang, 2017), racial bias among bail judges (Arnold, Dobbie and Yang, 2018), pretrial detention having higher rates of guilty pleas, conviction, recidivism, and worsened labor market outcomes (Leslie and Pope, 2018; Dobbie, Goldin and Yang, 2018; Stevenson, 2018), juvenile incarceration on high school completion and adult crime (Aizer and Doyle, 2015), and more. To identify the local average treatment effects, we examine the five standard IV assumptions as well as a sixth assumption due to the nature of administrative data: 1) stable unit treatment value assumption (SUTVA), 2) independence, 3) exclusion, 4) monotonicity, 5) non-zero first stages, and 6) no collider bias in the sample.

SUTVA requires that a person's observed outcome be based on their own treatment assignment with no spillover effects from anyone else's treatment assignment. But in the context of IV, it also means that their treatment status be based solely on their own instrument assignment. A violation of SUTVA in our context would occur if Inmate A was assigned a lenient evaluator causing Inmate B's score to change.[11] We have been assured by the Director of Inmate Mental Health at the Travis County correctional complex that

paid for each hour of effort, and since effort is costly, they may seek to minimize their costs by exerting the minimum effort above some personal reservation effort. Furthermore, given the low nominal rate, it is more likely that the labor supply would consist of lawyers whose main practices have low demand, thus creating the need to moonlight. And while altruistic highly competent defense attorneys are likely part of the labor supply of private indigent defense, reduced demand linked to the need to moonlight as well as the perverse incentives at the intensive margin implies at least some negative selection may be present in the pool of private indigent defense attorneys.

[11] We will define a "lenient" evaluator as one who tends to score inmate mental health symptoms higher on average, thus recommending inmates to the public defenders office more often.

such a spillover is impossible in our context due to the sequence of booking. Inmate scores are based on their evaluator and no one else's, and because evaluation happens within such a brief time after booking, there is practically no opportunity for social interactions caused by a peer's evaluator to influence another inmate's functioning even hypothetically.

The independence assumption in our context requires that evaluator assignment be independent of potential outcomes and potential treatment status. Independence of that nature is assured with physical randomization, which is accomplished in our setting because administrators use a quasi-randomization based on evaluators' last names.[12] Each day, evaluators are alphabetized by last name and then throughout the shift they are assigned to inmates as they arrive. Since evaluators work different shifts, and it's possible that inmates with differing latent mental illnesses commit offenses at different times and days throughout the week, we include day-of-week/month fixed effects in all our models. The inclusion of these time fixed effects therefore requires that independence hold within a given day.

The exclusion restriction requires that the instrument be independent of confounders and the unobserved determinants of the outcomes as well as have no direct effect on the outcomes of interest. This can be illustrated with a causal graph, shown in Figure A3, as missing arrows from Z to U and no direct effect of Z on Y. Exclusion might be violated if an evaluator provided mental health treatment in addition to scoring or if the correctional complex used the score to assign other treatments. For instance, if the jail treated those assigned to lawyers-with-social-workers with additional forms of assistance, such as cognitive behavioral therapy or some other unknown treatment during detention, then inmate mental health symptoms might improve which could have an effect on future suicidality and mental health symptoms regardless of whether they met with a social worker later. We examined this by speaking with the Director of Inmate Mental Health on numerous occasions and asking about whether the jail ever used these scores internally to assign other treatments. At the margin of high and low scores for the more mentally ill inmates, scores were not used to make any other treatment assignment except for defense

[12]The assignment of evaluators to inmates used at Travis County is similar to that used in Miguel and Kremer (2004) classic deworming study.

representation in the mental health court. This appears to be because capacity for care within the jail was reserved for low functioning individuals and since high and low scoring inmates met that condition, differences in the score were irrelevant at that margin for in jail treatments. There does not appear to be a plausible mechanism in our particular context whereby instrument assignment can influence future suicidality and mental health except via the instrument's effect on treatment assignment.

In many leniency design applications, it is the monotonicity assumption that is potentially problematic (Imbens and Angrist, 1994). Violations of monotonicity happen because this is not a true lottery. Inmates are not being randomly offered vouchers for lawyers-with-social-workers – they are randomly offered evaluators with differing tendencies to assign that treatment due to their own heterogeneity. Monotonicity in the instrument requires that it strictly operate in the same direction regardless of the kind of inmate seen by the evaluator. Consider the following situation wherein monotonicity would be violated. John and Sally are evaluators who score inmate functiong (or ADL) using a structured survey. John gives high scores 30% of the time, but Sally only 20% of the time. We therefore consider John to be the more "lenient" evaluator since he recommends the treatment more often. A monotonicity violation is simply a situation in which Sally suddenly became the more lenient of the two. If John gives a high score to every inmate that Sally would have given a high score to (and more), then strict monotonicity holds, but if ever there is a situation where in counterfactual John would not have given a high score to someone Sally had, then strict monotonicity is violated. Monotonicity rules out such "criss crossing" between evaluator assessments. We can test simultaneously for exclusion and strict monotonicity using a test by Frandsen, Lefgren and Leslie (2019). If we reject the null with this test, then we know either exclusion or monotonicity is violated. Since our knowledge of the institutional details in Travis county's correctional complex makes our team comfortable with assuming exclusion, rejection of the null points to a monotonicity violation. We also evaluate montonocity using the more traditional approach of subsample analysis on the first stage. We implement both types of tests in our effort to assess the degree to which monotonicity of any sort holds in our data.

The simplest hurdle of the six assumptions is the non-zero first stage assumption

because unlike the others, it can be directly evaluated with our data as it is not based on potential outcomes. It is simply based on the strength of the association between the evaluator's average high scores excluding the score of the inmate they are seeing and the score they give to that inmate. When such associations are strong enough using conventional statistical hypothesis test, the bias of 2SLS drops.

The last assumption we discuss here does not fall under the traditional IV assumptions but rather is about the data generating process that created our sample of inmates in the first place. Inmates can appear more than once in our data and when they do, it is called recidivism. If they reoffend after they leave jail, then they will appear in the same dataset a second or more times. This creates opportunities as well as challenges if recidivism is a collider (Schneider, 2020; Cunningham, 2021). We therefore need for identification of causal effects to be "no collider" in our sample or a "no collider assumption." A collider situation would be one in which lawyers-with-social-workers cause recidivism, some unknown factor causes recidivism, and that same unknown factor also caused an inmate's suicidality. We illustrate this problem with the directed acyclical graph (DAG) depicting our instrumental variable design presented in Figure A3. When D $\rightarrow$ R, it means that lawyers-with-social-workers cause recidivism. If that is true, then $D \rightarrow R \leftarrow U \rightarrow Y$ forms a causal chain in which R is a collider.[13] Cunningham (2021) and Schneider (2020) provide several examples in which colliders can create spurious correlations simply by creating select samples. But the collider problem is testable because we can directly estimate whether lawyers-with-social-workers cause recidivism using our instrumental variables strategy. If there is no effect of lawyers-with-social-workers on recidivism, then the $D \rightarrow R$ does not exist, and the aforementioned collider chain is not present. We discuss results from this analysis in a later section.

When all six assumptions are met, our estimation strategy can recover consistent and sometimes unbiased estimates of weighted averages of causal effects for the complier subpopulation, or the local average treatment effect (LATE). The LATE parameter is the average treatment effect for the marginal inmates who were only assigned to the

[13]Colliders are variables that are the descendent of two separate variables and have a DAG representation of $D \rightarrow R \leftarrow U$.

lawyer-with-social-worker because he randomly was assigned the more lenient evaluator. Estimates of the LATE parameter provide insight into the relative impact that assignment to lawyers-with-social-workers had on the marginal inmate's mental health upon reentry into jail, but do not tell us anything about the average effect of lawyers-with-social-workers outside of that subpopulation. To calculate aggregate parameters like the average treatment effect, we can use marginal treatment effect analysis, which we discuss in a subsequent section.

3.4 Instrumental Variable Calculation

For our just identified two-stage least squares (2SLS) models, we construct the residualized leave-one-out mean measure of each evaluator's tendency to score an inmate with severe symptoms. As those with the most severe symptoms are assigned to the public defender's office, we convert all scores into a binary treatment variable with 1 being most severe symptoms and 0 being assignment to the private indigent defense attorney for the mental health court sample. We use the residualized leave-one-out mean as an instrument for a evaluator's evaluation that an inmate's symptoms are severe in all of our just identified 2SLS models. For our basic IV models, we present the 2SLS results with the residualized leave-one-out-mean as an instrument in a just identified model, which is similar to the approaches taken by Aizer and Doyle (2015) and Arnold, Dobbie and Yang (2018), as this allows us to visually display the single instrument's distribution, as well as check for balance across the instrument's distribution. Hull (2017) notes that researchers often use the residualized leave-one-out-mean as an instrument for treatment because it is typically simpler than inverting a multidimensional matrix in 2SLS. It can be shown that 2SLS with the residualized leave-one-out-mean as an instrumental variable is comparable to using JIVE with evaulator fixed effects as instruments. Therefore in addition to 2SLS, we also explore four other models more appropriate for multiple instruments: JIVE (Angrist, Imbens and Krueger, 1999), UJIVE (Kolesa'r, 2013) and double-selection (Chernozhukov, Hansen and Spindler, 2015).

Following Arnold, Dobbie and Yang (2018) and Aizer and Doyle (2015), we construct the residualized leave-one-out mean by first regressing an indicator variable equalling one

if the evaluator scored an inmate's symptoms as severe onto a vector of time controls (day-of-week/month fixed effects).[14] This was done as a balancing act between having enough power within each fixed effect to estimate parameters, and wanting to restrict identification to periods in time that control for both seasonality and scheduling. Next, we calculate the residual, $\tilde{D}_{dkt}$, from this regression. Finally, we use the residualized symptom severity evaluation rate to calculate the evaluator recommendation instrument, $\tilde{Z}_{cl}$, as a residualized leave-one-out mean symptom severity evaluation rate associated with each randomly assigned evaluator l and inmate c.

To calculate the leave-one-out mean, we use the following formula which is the same as used by Aizer and Doyle (2015) and others.

$$\begin{aligned} \widetilde{Z}_{cl} &= \left(\frac{1}{n_l - n_c}\right)\left(\sum_{k=0}^{n_l} \widetilde{D}_{dkt} - \sum_{k\in\{c\}} \widetilde{D}_{dkt}\right) \\ &= \frac{1}{n_l - 1}\sum_{k\neq c}^{n_l-1} \widetilde{D}_{dkt} \end{aligned} \tag{1}$$

We overlaid the residualized leave-one-out with the share of individuals scored as having severe symptoms and present it in Figure 1. As can be seen, there is a strong correlation between the average tendency of a evaluator to score other inmates' symptoms as severe and whether they do so in the inmate's own case. Furthermore, there is a large spread in recommendation rates in the first place ranging from -0.6 (normalized) to 0.6.[15]

Table 2 shows the strength of the first stage. A one-point change in the leave-one-out mean is associated with a 0.62-0.64 increase in probability of scoring severe symptoms. We present the robust Kleibergen-Paap first stage F, which is equivalent to the effective F-statistic of Olea and Pflueger (2013) in the case of a single instrument, in the lower panel of Table 2. Our F-statistic from the first stage is 17 and suggestive of a strong non-zero first stage.

[14] We experimented with different time-based fixed effects, but our results never change much.

[15] We also present evidence for systematic differences in Figure A1 and Figure A2, which show effect sizes on evaluator fixed effects as well as the distribution of t-statistics. There is considerable variation in effect sizes and t statistics.

While we cannot directly test the balance of unobservables across our leniency measure, we can test whether observable characteristics are balanced. In Table 3, we present a table of inmate characteristics across the distribution of the residualized leave-one-out-mean instrumental variable with p-values on differences in means for the bottom and middle tercile of the instrument, and between top and bottom, respectively. For the most part, covariates are balanced except in the stratum of Black inmates. But, while there are statistically different effects for shares of Black people across the leniency measure, the difference in magnitudes are not economically meaningful (0.279 vs. 0.253). We conclude that observable characteristics of inmates are balanced. Given the randomization of evaluators, we assert it is most likely the case that unobservables are as well.

Finally, in Table 4, we present evidence on the independence of the instrument from inmate characteristics. Once we construct the instrument, the differences between those with and without severe symptoms shrink to small zeroes, and only one is statistically significant. This is not surprising given that the first stage models a conditional probability of leniency, and Rosenbaum and Rubin (1983) show that propensity score absorbs all information from the covariates.

3.5 Monotonicity

One of the key insights highlighted by Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996) is that instrumental variables models require monotonicity, and without it, the underlying weights are unstable and estimated coefficients do not have a causal interpretation. Monotonicity may fail if scores are based on observable or unobservable inmate characteristics, or due to heterogeneous skills in evaluators themselves (Chan, Gentzkow and Yu, 2019). We examine this using both a traditional approach that checks for qualitative consistency in the first stage across subsamples of the data, as well as a new test by Frandsen, Lefgren and Leslie (2019). We implement the Frandsen, Lefgren and Leslie (2019) test, and can only reject strict monotonicity for our suicide attempt outcome; both suicidal ideation and mental health improvements have large p-values suggesting that strict monotonicity holds (Table A3) for those outcomes, but not suicide attempts.

Since strict monotonicity was rejected for suicide attempt, we turn to evaluating whether average monotoncity might hold. Average monotoncity is sufficient for estimating the local average treatment effect, but is not sufficient for estimating marginal treatment effects. We present evidence for average monotonicity in Table A4 columns 1-7 along the top row. In all race and age subsamples, our instrument is significant and qualitatively in the same direction. Interestingly, the magnitudes on the instrument are considerably larger for Black inmates and female inmates, both of which are over 25-50% larger than the next largest effect. There is a much higher tendency to see lower functioning in Black inmates and female inmates regardless of their presenting symptoms, which may suggest evaluators are even more systematic with respect to these two demographics presenting symptoms than with the rest of the sample (Hampton, 2007). We conclude based on these two tests that strict monotonicity holds for two of our three measures, whereas average monotonicity holds more generally since our leniency measure operates, on average, the same for numerous subsamples. We proceed next to a discussion of our IV estimation and modeling specification.

3.6 IV Estimators

There are two approaches one can take when estimating the first stage: instrument for treatment assignment using the aforementioned leave-one-out-mean high score or using evaluator fixed effects (Hull, 2017). We examine both approaches for the sake of robustness.

The 2SLS model is consistent when all six assumptions hold but suffers from a finite sample bias that is exacerbated with weak instruments. However, Arnold, Dobbie and Yang (2018) and Hull (2017) note that the leave-one-out-mean measure of evaluator leniency is a collapsed scalar corresponding to a series of evaluator fixed effects, much like the propensity score itself because it reduces the dimension of our covariate set into a single scalar (Rosenbaum and Rubin, 1983).[16] Since the bias of 2SLS increases under multiple instruments, 2SLS may not be appropriate in our context as it blows up the bias

[16]Hull (2017) writes that "one should remember ... that the dimensionality of the underlying variation [in the instrument] is K, not one."

associated with 2SLS if exclusion is even slightly violated. Results from 2SLS are presented in Table 5, columns 3-4. Clinical evaluator and inmate robust two-way clustered standard errors and confidence intervals based on inversion of the Anderson-Rubin test are shown in parentheses and brackets below the coefficient estimates, respectively.

One popular alternative to the 2SLS model in these applications has been Angrist, Imbens and Krueger (1999)'s jackknife IV estimator (JIVE), but JIVE is not a panacea because of the challenges it faces in the presence of many covariates. Kolesa'r (2013) notes that in a finite sample, JIVE will be noisy and this estimation error will be correlated with the outcome since it depends on the treatment status of a particular inmate. This will cause JIVE to be biased when the number of covariates is large as is the case in our context. We have 14 covariates and 84 time fixed effects, which means that we must consider the potential bias. Therefore, we face a tradeoff between a set of time fixed effects that ensure conditional randomization and the biases created by large numbers of covariates for our JIVE estimator. Results from our JIVE model are presented in Table 5, columns 5-6.

To resolve this, we accompany our 2SLS and JIVE estimates with models that are more robust under a large number of covariates as well as a large number of instruments. Our first alternative is to estimate LATEs using the UJIVE estimator (Kolesa'r, 2013). UJIVE estimates are robust to a large set of covariates by excluding inmate i from estimation, thus guaranteeing that the aforementioned noisy prediction error be uncorrelated with the outcome (Kolesa'r, 2013). This means that UJIVE estimates are consistent for a convex combination of LATEs even when we have a large number of covariates. UJIVE estimates are presented in Table 6 columns 1 and 2.

We also present two machine learning selection IV models: the post-double-selection model and the post-lasso-orthogonalization method described by Chernozhukov, Hansen and Spindler (2015), which we loosely term our LASSO and post-LASSO models, respectively. LASSO and post-LASSO estimates are presented in Table 6 columns 3-6. These algorithms are designed to minimize the problems of including a large number of instruments (columns 3-4) as well as a large number of controls (columns 5-6). We use the Stata command `lassopack` for its implementation (Ahrens, Hansen and Schaeffer, 2020).

4 LATE Results

In subsection 4.1, we first introduce our OLS and 2SLS results to introduce readers to baseline evidence. We then discuss our recidivism results to test our sixth "no collider" assumption before discussing our main results. Interestingly, we find no effect on recidivism, and our results are consistent across all modeling choices.

4.1 Recidivism

Due to potential collider-based non-random sample selection problems, we cannot estimate the effect of public defense on mental health and suicide risk unless there is no differential effect of public defense on repeat offending (Section A.1, Figure A.3). We first report results for repeat offending using from several measurements. For instance, we examine whether an inmate re-entered the correctional complex (the most common definition of reoffending used by researchers), whether they did so within a year of booking, the number of times they committed another offense, the days to returning, and whether the next offense was a felony. We present estimates from 2SLS in Table A.1 and show that there is no statistically significant impact of lawyers-with-social-workers within the Travis County mental health courts on recidivism regardless of how recidivism is measured. But the point estimates are also imprecise in most cases, which gives some pause in being overly confident.

In addition to being an important test for our design, the lack of a finding on recidivism is interesting given that prior literature found public defenders improved defendant outcomes along criminal justice lines. For instance, a recent study by Shem-Tov (2021) found that public defenders reduced the probability of a prison sentence by 22% and the length of a prison sentence by 10%. However, similar to our paper, they did not examine an effect of public defense on repeat offending. It is important to note that our null result may simply be an artifact of the court we are examining. The Travis County mental health court is a "friendly" court known for its collaboration where prosecutor, defense attorney and judges work together to dismiss the defendant's charges in the hopes of his or her rehabilitation. Thus, the impact that any defense lawyer has on a tendency to

reoffend may be mitigated by the fact that punishment is uncommon in these courts. Given that there is no difference between public and private indigent defense attorneys on recidivism, then the problem of collider bias as shown in Figure A.3 may be safely and plausibly ignored.

This lack of a finding is crucial for our design because it suggests the collider problems we hypothetically lay out in Figure A3 are not present. Therefore, while it is possible that mental health court has some effect on recidivism, it does not appear that there is differential effects by lawyers with and without social workers on recidivism. Since our analysis is only about the lawyer assignment within the courts, we proceed with caution to our main results.

4.2 OLS and 2SLS Results

Next, we consider the results from our analysis of the estimated causal effects of public defenders on mental health outcomes. In columns 1-2 of Table 5, we report our OLS results showing public defenders are associated with a 2 percentage point reduction in suicide attempts at re-entry into jails and slight improvements in mental health scores (around a tenth of a point on a 4-point scale). We find no statistically distinguishable effect, though, on self-reported suicidal ideation.

In columns 3-6 of Table 5, we present LATE estimates using both 2SLS and JIVE. Recall that 2SLS bias increases in the number of instruments, and the bias of JIVE increases in the number of covariates, both of which we face in our context (Kolesa'r, 2013). Here we find stronger evidence that lawyers-with-social-workers positively impact the lives of defendants with low ADL functioning. With or without baseline controls, public defenders improve mental health outcomes on an order of magnitude that is 8-10 times larger than that of our OLS estimates. ADL functioning scores improve by almost a full point on the four point scale which corresponds to a 25% improvement. Suicide attempts fall 12-16% (compared to 2% using OLS) and self-reported suicidal ideation falls 1.4-2% (column 3-4) compared to a precisely estimated zero effect using OLS. Precision using JIVE for self-reported suicidal ideation decreases but is similar in magnitude, but estimates on subsequent ADL scores become negative, implausibly large

(as it is impossible to get any higher than a 3) and imprecise.

4.3 JIVE and Optimal Instrument Selection

As discussed above, we report results from UJIVE, LASSO, and post-LASSO to address the potential bias introduced with a large set of covariates. Our UJIVE results are presented in Table 6 columns 1 and 2. Here we find results that are more comparable to our 2SLS models. Suicide attempts decrease by 10-15%, suicidal ideation around 2%, though again imprecise, and subsequent ADL functioning scores improve around 1 point on the 4 point scale.

Our LASSO and Post-LASSO IV selection models do not change much from what we have reported thus far. Using the LASSO model, we find suicide attempts decline by 9-10%, and self reported suicidal ideation declines by 2%. Impacts on subsequent ADL functioning scores, on the other hand, are cut in half (0.5-0.58 points) but are less precise than earlier estimates. Overall, our UJIVE and post-LASSO-orthogonalization models find similar results as that of our other IV models, though with less precision and for some outcomes lower magnitudes, suggesting the large number of covariates included as controls is not introducing substantial bias in our case.

To summarize, we find lawyers-with-social-workers cause suicide attempts to fall by as low as 9% (Table 6, column 4) and as high as 22% (Table 5, column 5), both of which are considerably larger than what we find using OLS but within a range of our baseline 2SLS results. Self-reported suicidal ideation falls by closer to 2% and does not vary substantially from that number across all models. ADL functioning scores, on the other hand, improve by around 1 point on a 4 point scale, though we find smaller effects in our two LASSO models closer to half a point. Overall, our results suggest that lawyers-with-social-workers cause defendants' mental health to substantially improve conditional on re-entry into the jails, including extreme mental health problems – suicidal ideation and suicide attempts.

5 MTE Results

A LATE parameter estimate may lack policy relevance depending on the amount of heterogeneity in the population. The LATE measures the average treatment effect for the complier subpopulation and captures a set of compliers unique to the instrument chosen. Taken together, externally valid parameter estimates may hold but only for some comparably unknown subpopulation of compliers whose returns may be very different than anyone else. However, IV is still useful because when combined with MTE analysis we are able to identify more aggregate parameters, like the ATE, as well as investigate how returns vary with unobserved heterogeneity (Heckman and Vytlacil, 2013; Cornelissen et al., 2016).

5.1 MTE setup

Examining the MTE allows us to explore the heterogeneity in treatment effects across inmates' underlying latent mental illness proxied by the propensity score. Consider the following two equations decomposing an inmate i's potential mental health outcomes into the conditional means of potential outcomes, $\mu^j(X_i)$, based on inmate characteristics X_i as well as deviations from the mean, U_i^j. The variables, Y^1 and Y^0, each measure an individual's potential mental health outcome under lawyers-with-social-workers or lawyers-without-social-workers states of the world.

$$\begin{aligned} Y_i^0 &= \mu^0(X_i) + U_i^0 \\ Y_i^1 &= \mu^1(X_i) + U_i^1 \end{aligned}$$

We can write down a person's treatment selection model as a choice made based on an individual's latent index threshold in which the net benefits of lawyers-with-social-workers are exactly equal to

$$D_i^* = \mu^D(X_i, Z_i) - V_i$$

where X_i and Z_i are the inmate's observed determinants of treatment choice, Z_i are the instruments excluded from the outcome equation, and V_i is the unobserved characteristics that makes treatment choice less likely (i.e., the unobserved resistance to treatment) (Cornelissen et al., 2016). Assignment to lawyers-with-social-workers occurs when $D_i^* > 0$, otherwise they are assigned to lawyers-without-social-workers.

While we can follow the literature and describe D_i^* as the expected net gain to lawyers-with-social-workers, this is a strained interpretation given the treatment assignment to low functioning inmates is made by neither inmates, judges, nor lawyers. It is not even made by clinical evaluators with the possibility that an inmate may benefit from a higher score. The scoring of inmates is only based on the evaluator's assessment of the inmates' ADL functioning. It is meant, in other words, to be a description of the inmate, not a choice meant to improve the inmate's chances of rehabilitation.

Nevertheless, the decision tree followed in Travis does indeed assign each inmate to either treatment or control based on whether their functioning is low or high conditional on being eligible for the mental health court. We therefore can think of the selection of treatment as caused by the inmate's underlying latent disability, as opposed to any perceived benefit from treatment. His disability, in other words, "chose" lawyers-with-social-workers because he was severely disabled with respect to his daily function. Thus the word "choice" and the phrase "low functioning" are in this context synonyms. When the evaluator believes that the inmate's functioning is above the evaluator's own reservation threshold for that inmate, V_i, it causes the evaluator to assign a high score which in turn assigns him to lawyers-with-social-workers. The rule, in other words, can be thought of as "chosen" when the inmate's functioning falls below some threshold where $\mu_D(X, Z) = V$, that is, when $D^* > 0$, $\mu_D(X, Z) > V$ and assignment to lawyers-with-social-workers occurs.

When we apply the cdf of V to this inequality, we get $F_V(\mu^D(X_i, Z_i)) \geq F_V(V_i)$ which bounds both sides between 0 and 1. The left side becomes interpretable as the conditional probability of treatment (i.e., the propensity score) which we write as $p_i(X_i, Z_i)$. The right side becomes the quantiles of the distribution of the unobserved resistance to treatment, U_i^D. We can then rewrite the selection equation as $p_i(X_i, Z_i) \geq U_i^D$ which means an inmate

i selects into treatment when their propensity score is greater than their unwillingness to participate due to their slightly higher functioning.[17]

We can write down the choice of either potential outcome according to which treatment was chosen using a switching equation:

$$\begin{aligned} Y_i &= D_i Y_i^1 + (1 - D_i) Y_i^0 \\ Y_i &= Y_i^0 + D_i(Y_i^1 - Y_i^0) \end{aligned}$$

which is the regression model used by Krueger (1999). Substituting our earlier potential outcomes into the above model, we get:

$$Y_i = \mu^0(X_i) + D_i(\mu^1(X_i) - \mu^0(X_i) + U_i^1 + U_i^0) + U_i^0$$

where the interior term is the coefficient on D_i which measures the unique return to the treatment at the individual i level. This coefficient has two components: the average gain of someone with this person's characteristics and an idiosyncratic individual effect (Cornelissen et al., 2016).

5.2 Estimation and Results

The steps to estimation are straightforward. First, we subset the sample into cells defined by X. Then within each subsample, we estimate the propensity score as a function of the excluded instruments. We then model the outcome parametrically as a flexible function of the propensity score and calculate the predicted outcome using a second order polynomial. Finally, within each sample, we take the derivative of the predicted outcome with respect to the propensity score. Doing this for a grid of values for the propensity score allows us to trace out the marginal treatment effects from 0 to 1 across the propensity score estimate.

[17] The MTE literature often describes this unobserved heterogeneity in terms of "distaste for treatment" (Cornelissen et al., 2016). Here distaste for treatment would simply mean having better functioning at the time of assessment since only those with better functioning would "resist" the treatment assignment.

This requires common support across the sample, which we do not have in the tails of the propensity score distribution as there are so few control group units with the extreme values. Thus, we trimmed the sample by dropping the top 0.1% of the propensity score. We present a histogram of the distribution of the propensity score for our treatment and untreated groups in Figure 3 and show where the distribution is trimmed with the vertical dashed lines.

MTE is defined as a continuum of treatment effects along the full distribution of individual unobserved heterogeneity driving the decision to "choose" lawyers-with-social-workers. This allows us to identify aggregate parameters like the ATE, ATT, and ATUT by summing and weighting the MTE accordingly. To represent these heterogeneous gains from treatment visually, we plot the MTE against the resistance to treatment in Figures 4 to 6. The slope of the MTE is represented as a solid dark line, whereas the ATE is represented in the horizontal dashed line with its 95% confidence intervals shaded. The horizontal axis shows quantiles of unobserved resistance to participate in the treatment recalling that people with higher functioning are by definition more "resistant" to treatment because the treatment is reserved for people with lower functioning. Thus moving from left to right is necessarily a movement from those with less severe problems to more severe which we suspect means moving along a spectrum of unobserved mental illness ranging from the most severe (left portion of the x-axis) to the less severe (right portion of the x-axis). The vertical axis shows the treatment effect with covariates held constant at means.

Positive selection on the unobserved returns to lawyers-with-social-workers is represented by a downward sloping MTE curve whereas a flat MTE curve suggests that the assignment shows no signs of selection. We do not find strong evidence for positive or negative selection in Figure 4 for suicide attempts given the ATE, ATT, and ATU are all so similar, but we also note that since we found evidence that strict monotonicity fails for this outcome, these aggregate parameters should be taken with a grain of salt anyway. Nevertheless, the ATE is comparable to what we found for the LATE, and while the MTE is slightly rising, it is not statistically different from a flat MTE curve or even a slightly falling one. The MTE curve for suicidal ideation can be used to calculate aggregate pa-

rameters, but this is not terribly illuminating either because we do not find evidence for heterogeneity. Our LATE estimates are comparable to the aggregate parameters, suggesting that our LATE parameters may have external validity, but with less precision. The ATE is a weighted average of the MTE over the entire distribution, which given the lack of variability in the MTE curve for both suicidality measures is likely why we don't see much variability in the various aggregate treatment parameters.

But, when we get to Figure 6, we do find evidence for underlying heterogeneity in the MTE for subsequent ADL functioning which had been masked in our LATE estimates. Note how in Figure 6 the MTE curve associated with subsequent ADL functioning rises as we move from the most severely compromised (left portion) to the less compromised. This type of rising MTE curve represents reverse selection on gains with those entering at the margin into the lawyers-with-social-workers assignment being those with less to gain from it with respect to future mental health improvements. This implies that around the highest propensity score levels, the marginal treatment effect of the treatment slightly falls.

But we quickly see that this trend reverses around the 0.24 mark on the horizontal axis prompting a descent that is consistent with selection on unobserved gains. Selection on unobserved gains in our context means that as we move towards the seriously mentally ill, the gains from treatment get larger. This is interesting in light of mental health critic Jaffe (2017) who noted that the number of treatments that can truly improve the mental health functioning of the seriously mentally ill are small because the seriously mentally ill, as a group, are more likely to disengage from complying with prescription medication, such as antipsychotics, and are more likely to habitually engage in substance abuse. But the fact that MTE curve begins to decline as we move closer to those with fewer latent mental health problems suggests that the real returns to lawyers-with-social-workers are among the most seriously mentally ill. This is because the ATT is a weighted average of the MTE that more intensely weights the left than the right side of the x-axis and thus causes it to be larger than the ATE. Our findings suggest that the seriously mentally ill, insofar as they are the bulk of those with low unobserved resistance to treatment, gain the most from this lawyers-with-social-workers intervention, which is impressive given the

poor track record that many efforts to help them have had (Jaffe, 2017).

6 Discussion

In Travis County, the courts use a score based on observable functioning problems to assign inmates to indigent defense that vary with regards to their employment of social workers as well as lawyer type conditional on assignment to the mental health court itself. Using this treatment assignment, we estimate a range of IV models and provide evidence that lawyers-with-social-workers improve inmate mental health, including suicidality, of inmates upon reentry to jail. Given that suicide is the leading cause of death in jails, the large share of mental illness among inmates, and the magnitudes of our estimates, our findings suggest that access to social workers for the mentally ill populations should be a priority for counties with mental health courts.

We tentatively suggest two broad explanations for these findings. First, differential responses to lawyers-with-social-workers may have more to do with the lawyer part of the treatment than the social worker component. Public defenders and private attorneys drawn from different pools of lawyers facing different incentives represent defendants with moderate to severe mental illness. Public defenders, for instance, are salaried and select on attorneys who may or may not be, on average, drawn to the work by a sense of civic-minded commitment to indigent defense. Private attorneys, on the other hand, are engaged in dual job holding and may lack proper incentives at the margin to exert effort given they are paid a flat fee versus one that varies at the margin (Shishko and Rostker, 1976). We are skeptical that our findings are driven by lawyer selection since all of our findings appear in the mental health categories, but not recidivism (Table A.3). Private conversations with Travis County mental health administrative staff and the director of mental health at the correctional complex itself also cast doubt on the negative selection hypothesis, as administrators and judges claim the two attorney pools are roughly comparable. Finally, there are also fewer margins for the lawyer to have much impact given the court as a whole is working together to dismiss. Lawyer talent and skill, in other words, does not appear to have the kind of opportunity to influence the

defendant's life when the entire court is non-antagonistic in the first place. Furthermore, given the lower functioning levels of the severe mentally ill inmates, one would expect that they would have a higher propensity for recidivism. This null finding allows us to explore within court variation because the composition of the recidivism sample will not have the collider issues discussed above. While we cannot completely rule out the lawyer selection mechanism, we do not find it as convincing since the only outcomes affected are mental health ones.

The more plausible mechanism, in our opinion, is the social worker component of the treatment. Public defenders employ a large number of social workers to help inmates with serious mental illness through a variety of daily and necessary tasks such as compliance with medication, signing up for services, getting to court and even finding housing. Conversations with Travis County staff suggest that this disparity is a major political consideration for the county because it is a valuable resource that is unequally distributed across the two pools of indigent defense attorneys. While we cannot conclusively prove that our findings are driven by social workers, we offer the following evidence for that theory. First, the effects only show clear mental health outcomes. This suggests the cause has something more directly to do with mental healthcare, which points to the social workers employed by the public defenders. Second, the results only hold for inmates who had never had been medicated, diagnosed, or hospitalized for a mental illness, suggesting that something new was happening to these inmates as a result of their assignment to lawyers-with-social-workers. Given these large returns, this analysis suggests that given even modest estimates of the statistical value of a life, expanding access to social workers likely pays for itself if it reduces the risk of suicide and improves ADL functioning.

The seriously mentally ill are a subpopulation within a larger population of mentally ill individuals who are very hard to help due to the loss of social capital, alienation from caregivers, lost income from unemployment, homelessness, extreme substance abuse, noncompliance with medication, and sometimes, accompanying violence from their untreated psychosis. The social costs for this group are large, both in terms of their own suffering, the costs borne by their under-resourced caregivers, their repeated use of public resources such as hospitals, jails, and police, and much more. Their disproportionate consumption

of public goods places strains on the economic order, often without any return. But even if that was small, the costs nonetheless are potentially quite large for this group because of the impact they have on caregivers, families, and other loved ones with little recourse. For example, the US has very strict protections of the civil liberties of the mentally ill, including the extremely high thresholds needed for involuntary hospitalizations under the deinstitutionalization movement, thereby shifting the costs to caregivers with virtually no aid. Thus when we see evidence of success during an important bottleneck stage of adjudication in mental health courts, it suggests an avenue for successful treatments that may help lower social costs overall and improve lives.

References

Ahrens, Achim, Christian B. Hansen and Mark E. Schaeffer. 2020. "lassopack: Model selection and prediction with regularized regression in Stata." Stata Journal 20.

Aizer, Anna and Joseph J. Doyle. 2015. "Juvenile Incarceration, Human Capital, and Future Crime: Evidence from Randomly Assigned Judges." Quarterly Journal of Economics 130(2):759–803.

Angrist, Joshua D., Guido W. Imbens and Alan B. Krueger. 1999. "Jackknife Instrumental Variables Estimation." Journal of Applied Econometrics 14(1):57–67.

Angrist, Joshua D., Guido W. Imbens and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." Journal of the American Statistical Association 87:328–336.

Angrist, Joshua D. and Jorn-Steffen Pischke. 2009. Mostly Harmless Econometrics. 1st ed. Princeton University Press.

Arnold, David, Will Dobbie and Crystal S. Yang. 2018. "Racial Bias in Bail Decisions." Quarterly Journal of Economics 133(4):1885–1932.

Biasi, Barbara, Michael S. Dahl and Petra Moder. 2020. "Career Effects of Mental Health." Working paper.

Blaauw, E. and F.W. Winkel. 2001. "Bullying and Suicidal Behavior in Jails." Criminal Justice and Behavior .

Borrill, J. 2002. "Self-inflected Deaths of Prisoners Serving Life Sentences 1988-2001." The British Journal of Forensic Practice .

Carson, E. Ann and Mary P. Cowhig. 2020a. "Mortality in Local Jails, 2000-2016." Bureau of Justice Statistics Statistics Tables NCJ 251921.

Carson, E. Ann and Mary P. Cowhig. 2020b. "Mortality in State and Federal Prisons, 2001-2016." Bureau of Justice Statistics Statistics Tables NCJ 251920.

Center for Substance Abuse Treatment. 2005. "8 Treatment Issues Specific to Jails." Substance Abuse Treatment for Adults in the Criminal Justice System Substance Abuse and Mental Health Services Administration.

Chan, David C., Matthew Gentzkow and Chuan Yu. 2019. "Selection with Variation in Diagnostic Skill: Evidence from Radiologists." NBER Working Paper Nol. 26467.

Chernozhukov, Victor, Christian Hansen and Martin Spindler. 2015. "Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments." American Economic Review: AEA Papers and Proceedings 105(5):486–490.

Cornelissen, Thomas, Christian Dustmann, Anna Raute and Uta Schönberg. 2016. "From LATE to MTE: Alternative Methods for the Evaluation of Policy Interventions." Labour Economics 41:47–60.

Cunningham, Scott. 2021. Causal Inference: The Mixtape. Yale University Press.

Dillon, Daniel. 2013. "A Portrait of Suicides in Texas Jails: Who is at Risk and How Do We Stop It?" LBJ Journal of Public Affairs 21(Fall):51–67.

Dobbie, Will, Jaconb Goldin and Crystal S. Yang. 2018. "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges." American Economic Review 108(2):201–240.

Dobbie, Will, Paul Goldsmith-Pinkham and Crystal Yang. 2017. "Consumer Bankruptcy and Financial Health." Review of Economics and Statistics 99(5):853–869.

Frandsen, Brigham R., Lars J. Lefgren and Emily C. Leslie. 2019. "Judging Judge Fixed Effects." NBER Working Paper No. 25528.

Frank, Richard G. and Sherry Glied. 2006. Better But Not Well. Johns Hopkins University Press.

Frank, Richard and Thomas G. McGuire. 2010. "Mental Health Treatment and Criminal Justice Outcomes." Unpublished Manuscript.

Frottier, P., S. Fruhwald, K. Ritter, R. Eher, J. Schwarzler and P. Bauer. 2002. "Jailhouse Blues Revisited." Social Psychiatry and Pscyhiatric Epidemiology 37(2):68–73.

Fruehwald, S., T. Matschnig, F. Koenig, P. Bauer and P. Frottier. 2004. "Suicide in Custody: Case-Control Study." The British Journal of Psychiatry: The Journal of Mental Science 185:494–498.

Fryer, Roland. 2019. "An Empirical Analysis of Racial Differences in Police Use of Force." Journal of Political Economy 127(3).

Goldkamp, John S. and Cheryl Irons-Guynn. 2001. Emerging Judicial Strategies for the Mentally Ill in the Criminal Caseload: Mental Health Courts in Fort Lauderdale, Seattle, San Bernardino and Anchorage. Diane Pub Co.

Goss, J.R., Peterson K. Smith L.W. Kalb K. Brodey B.B. 2002. "Characteristics of suicide attempts in a large urban jail system with an established suicide prevention program." Psychiatric Services 53:574–579.

Hampton, MichelleDeCoux. 2007. "The Role of Treatment Setting and High Acuity in the Overdiagnosis of Schizophrenia in African Americans." Archives of Psychiatric Nursing 21(6).

Harcourt, Bernard E. 2006. "From the Asylum to the Prison: Rethinking the Incarceration Revolution." Texas Law Review 84:1751–1786.

Heckman, James and Ed Vytlacil. 2013. "Structural Equations, Treatment Effects and Econometric Policy Evaluation." Econometrica 73(3):669–738.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." Econometrica 47(1).

Huey, M. P. and T. L. McNulty. 2005. "Institutional Conditions and Prison Suicide: Conditional Effects of Deprivation and Overcrowding." The Prison Journal .

Hull, Peter. 2017. "Examiner Designs and First-Stage F Statistics: A Caution." Unpublished Manuscript.

Imbens, Guide W. and Donald B. Rubin. 2015. Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction. 1st ed. Cambridge University Press.

Imbens, Guideo W. and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." Econometrica 62(2):467–475.

Jaffe, DJ. 2017. Insane Consequences: How the Mental Health Industry Fails the Mentally Ill. Prometheus.

Jenkins, R., D. Bhugra, H. Meltzer, N. Singleton, P. Beggington, T. Brugha, J. Coid, M. Farrell, G. Lewis and J. Paton. 2005. "Psychiatric and Social Aspects of Suicidal Behaviour in Prisons." Psychological Medicine 35(2):257–269.

Joe, Sean and Danielle Niedermeier. 2008. "Preventing Suicide: A Neglected Social Work Research Agenda." British Journal of Social Work 38(3).

Knox, Dean, Will Lowe and Jonathan Mummolo. 2020. "Administrative Records Mask Racially Biased Policing." American Political Science Review Forthcoming.

Kolesa'r, Michal. 2013. "Estimation in an Instrumental Variables Model with Treatment Effect Heterogeneity." Working paper.

Krueger, Alan. 1999. "Experimental Estimates of Education Production Functions." Quarterly Journal of Economics 114(2):497–532.

Leese, M., S. Thomas and L. Snow. 2006. "An Ecological Study of Factors Associated with Rates of Self-inflicted Death in Prisons in England and Wales." International Journal of Law and Psychiatry 29(5):355–360.

Leslie, Emily and Nolan G. Pope. 2018. "The Unintended Impact of Pretrial Detention on Case Outcomes: Evidence from New York City Arraignments." Journal of Law and Economics 60(3):529–557.

Liebling, A. and A. Ludlow. 2016. Suicide, Distress and the Quality of Prison Life.

Maple, Myfanwy, Tania Pearce, Rebecca L. Sanford and Julie Cerel. 2017. "The Role of Social Work in Suicide Prevention, Intervention, and Postvention: Scoping Review." Australian Social Work 70(3).

Marzano, L., K. Hawton, A. Rivlin, E. Smith, N. E., M. Piper and S. Fazel. 2016. "Prevention of Suicidal Behavior in Prisons." Crisis 37(5):323–334.

Miguel, Edward and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." Econometrica 72(1):159–217.

Morgan, Stephen L. and Christopher Winship. 2014. Counterfactuals and Causal Inference: Methods and Principles for Social Research. 2nd ed. Cambridge University Press.

Neal, Derek and Armin Rick. 2014. "The Prison Boom and the Lack of Black Progress after Smith and Welch." NBER Working Paper No. 20283.

Olea, Jose Luis Monteal and Carolin Pflueger. 2013. "A Robust Test for Weak Instruments." Journal of Business and Economic Statistics 31(3).

Patillo, Mary, David Weidman and Bruce Western, eds. 2004. Imprisoning America: The Social Effects of Mass Incarceration. Russell Sage Foundation.

Pearl, Judea. 2009. Causality. 2nd ed. Cambridge University Press.

Pratt, D., M. Piper, L. Appleby, R. Webb and J. Shaw. 2006. "Suicide in Recently Released Prisoners: A Population-based Cohort Study." The Lancet 368(9530):119–123.

Raphael, Steven and Michael A. Stoll. 2013. "Assessing the Contribution of the Deinstitutionalization of the Mentally Ill to Growth in the US Incarceration Rate." Journal of Legal Studies 42(1).

Redlich, Allison D., Steven Hoover, Alicia Summers and Henry J. Steadman. 2010. "Enrollment in Mental Health Courts: Voluntariness, Knowingness, and Adjudicative Competence." Law and Human Behavior 34:91–104.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." Biometrika 70(1):41–55.

Schneider, Eric B. 2020. "Collider Bias in Economic History Research." Explorations of Economic History 78(101356).

Shem-Tov, Yotam. 2021. "Make or Buy? The Provision of Indigent Defense Services in the U.S." Review of Economics and Statistics Forthcoming.

Shishko, R. and B. Rostker. 1976. "The Economics of Multiple Job Holding." American Economic Review 66:298–308.

Stack, Steven John. 2014. "Mental Illness and Suicide." The Wiley Blackwell Encyclopedia of Health, Illness, Behavior, and Society https://doi.org/10.1002/9781118410868.wbehibs067.

Steadman, H. J., S. Davidson and C. Brown. 2001. "Law and Psychiatry: Mental Health Courts: Their Promise and Unanswered Questions." Psychiatric Services 52(4):457–458.

Stevenson, Megan T. 2018. "Distortion of Justice: How the Inability to Pay Bail Affects Case Outcomes." The Journal of Law, Economics and Organization 34(4):511–542.

Torrey, E. Fuller, Mary T. Zdanowiz, Aaron D. Kennard, H. Richard Lamb, Donald F. Eslinger, Michael C. Biasotti and Doris A. Fuller. 2014. "The Treatment of Persons with Mental Illnesses in Prisons and Jails: A State Survey." Treatment Advocacy Report.

Turecki, G. and D. A. Brent. 2016. "Suicide and Suicidal Behaviour." Lancet 387(10024):1227–1239.

Watson, Amy, Patricia Hanrahan, Daniel Luchins and Arthur Lurigio. 2001. "Mental Health Courts and the Complex Issue of Mentally Ill Offenders." Psychiatric Services 52(4):477–481.

Way, B. B., R. Miraglia, D. A. Sayer, R. Beer and J. Eddy. 2005. "Factors Related to Suicide in New York State Prisons." International Journal of Law and Psychiatry 28(3):207–221.

Western, Bruce. 2006. Punishment and Inequality in America. Russell Sage Foundation.

WHO. 1999. Mental and Behavioural Disorders Team. Technical report Worth Health Organization.

WHO. 2007. Preventing Suicide in Jails and Prisons. Technical report World Health Organization.

Wolff, Nancy. 2002. "Courts as Therapeutic Agents: Thinking Past the Novelty of Mental Health Courts." The Journal of the American Academy of Psychiatry and the Law 30(3):431–437.

Wolff, Nancy and Wendy Pogorzelski. 2005. "Measuring the Effectiveness of Mental Health Courts: Challenges and Recommendations." Psychology, Public Policy and Law 11(4):539–569.

7 Tables and Figures

Table 1 Descriptive Statistics by Initial Mental Health Assessment

	Moderate	Severe
Outcomes		
Suicide attempt in next booking	0.051	0.030
Suicide ideation in next booking	0.006	0.004
Next booking ADL score improves	0.431	0.547
Inmate Characteristics		
White	0.731	0.704
Asian	0.009	0.011
Black	0.259	0.284
Race other	0.001	0.001
Hispanic	0.218	0.177
Male	0.630	0.702
Age at booking	35.653	37.204
Prior offense w/in 365 days	0.379	0.449
Number of offenses per booking	1.597	1.654
First time in jail	0.019	0.014
Prior treatment	0.140	0.087
Prior medications	0.129	0.089
Prior hospitalization	0.103	0.080
Homeless	0.055	0.042
Jobless	0.073	0.052
Clinician Characteristics		
Clinician Male	0.185	0.200
Clinician White	0.841	0.903
Clinician Black	0.079	0.042
Clinician Hispanic	0.074	0.045
Observations	4,294	928

Table 2 First Stage Regressions for Initial Assessment of Most Severe Mental Health Rating

	(1)	(2)
Z: Clinician's Leave-Out Mean Mental Health Score	0.635***	0.619***
	(0.152)	(0.150)
Kleibergen-Paap F	17.3653	17.1609
Time Fixed Effects	Yes	Yes
Baseline Controls	No	Yes
Observations	5,215	5,215

These estimates are the first stage results of a linear probability model with outcome of interest being the initial assessment of an inmate's mental health being most severe as opposed to moderately severe. The propensity to assign the most severe score is estimated using data from other cases assigned to the clinician as explained in the text. Column (1) shows the results by controlling only for day-of-week-month fixed effects, whereas Column (2) also includes the baseline controls of the inmates as shown in Table 1. Each column gives the corresponding clinician and inmate robust two-way clustered standard errors in parentheses, and the robust (Kleibergen-Paap) first stage F statistic is reported. * $p<0.10$, ** $p<0.05$, *** $p<0.01$

Table 3 Balance of Instrument and Inmate Characteristics for Most Severe Mental Health Rating

	Bottom Tercile	Middle Tercile	Top Tercile	Middle v. Bottom P-Value	Top v. Bottom P-Value
Z: Clinician's Leave-Out Mean Mental Health Score	-0.088	-0.020	0.107	(0.000)	(0.000)
Inmate Characteristics					
Asian	0.010	0.009	0.009	(0.717)	(0.679)
Black	0.279	0.256	0.253	(0.069)	(0.003)
Race other	0.001	0.001	0.002	(0.365)	(0.768)
Hispanic	0.202	0.227	0.202	(0.248)	(0.795)
Male	0.643	0.639	0.649	(0.976)	(0.892)
Age at booking	36.445	35.793	35.523	(0.372)	(0.133)
Prior offense w/in 365 days	0.380	0.372	0.421	(0.706)	(0.082)
Number of offenses per booking	1.606	1.581	1.637	(0.820)	(0.467)
First time in jail	0.025	0.018	0.011	(0.434)	(0.174)
Prior treatment	0.176	0.118	0.098	(0.420)	(0.363)
Prior medications	0.163	0.112	0.092	(0.451)	(0.380)
Prior hospitalization	0.136	0.089	0.073	(0.413)	(0.339)
Homeless	0.062	0.050	0.045	(0.658)	(0.688)
Jobless	0.090	0.074	0.045	(0.745)	(0.293)

Data is from a large county correctional complex.

Time fixed effects include day-of-week-month fixed effects.

Clinician and inmate two-way clustered standard errors shown in parentheses.

* p<0.10, ** p<0.05, *** p<0.01

Table 4 Test of Randomization for Most Severe Mental Health Rating

	(1) Most High ADL Score Rating	(2) Z: Most High ADL Score Rating
Asian	0.030	-0.006
	(0.054)	(0.008)
Black	0.004	-0.007**
	(0.014)	(0.003)
Race other	-0.067	0.011
	(0.144)	(0.017)
Hispanic	-0.035**	-0.002
	(0.015)	(0.004)
Male	0.040***	0.004
	(0.011)	(0.003)
Age at booking	0.001**	-0.000
	(0.001)	(0.000)
Prior offense w/in 365 days	0.036***	0.003
	(0.013)	(0.003)
Number of offenses per booking	0.005	0.000
	(0.005)	(0.001)
First time in jail	0.041	-0.018
	(0.031)	(0.012)
Prior treatment	-0.141***	-0.014
	(0.041)	(0.018)
Prior medications	0.046	-0.004
	(0.037)	(0.012)
Prior hospitalization	0.048**	0.001
	(0.024)	(0.008)
Homeless	-0.007	-0.005
	(0.029)	(0.010)
Jobless	-0.014	-0.018
	(0.018)	(0.013)
Time fixed effects	Yes	Yes
F-test	6	2
Observations	5,222	5,215

Data is from a large county correctional complex.

Time fixed effects include day-of-week-month fixed effects.

Clinician and inmate two-way clustered standard errors shown in parentheses.

* p<0.10, ** p<0.05, *** p<0.01

Table 5 Effects of Initial Assessment of Most Severe Mental Health Rating on Health Outcomes

	OLS		2SLS		JIVE	
	(1)	(2)	(3)	(4)	(5)	(6)
Suicide attempt in next booking	-0.020***	-0.016***	-0.158**	-0.122**	-0.221***	-0.174***
	(0.006)	(0.006)	(0.064)	(0.060)	(0.055)	(0.061)
			[-0.325, -0.053]	[-0.290, -0.035]		
Suicide ideation in next booking	-0.002	-0.002	-0.019**	-0.014*	-0.034*	-0.029
	(0.003)	(0.003)	(0.008)	(0.008)	(0.018)	(0.021)
			[-0.037, -0.003]	[-0.033, -0.000]		
Next booking ADL score improves	0.115***	0.136***	0.964***	0.981***	-4.848	-1.969*
	(0.036)	(0.037)	(0.274)	(0.249)	(6.378)	(1.173)
			[0.518, 1.619]	[0.577, 1.575]		
Time fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Baseline Controls	No	Yes	No	Yes	No	Yes

This table presents the ordinary least squares (OLS), two-stage least squares (2SLS), and jacknived instrumental variables (JIVE) (Angrist et al 1999) estimates of the impact of a clinician's initial assessment of a most severe mental health rating on inmates' subsequent mental health. The outcome variables of interest are given in each row along with the corresponding estimates of the impacts of an initial assessment of a most severe mental health rating. The 2SLS specifications instrument for severe mental health rating using a clinician leniency measure that is estimated using data from other cases assigned to a clinician. For further explanation, please see the text. We include day-of-week-month fixed effects for all specifications and baseline controls for Columns (2), (4), and (6). The clinician and inmate robust two-way clustered standard errors are shown in parentheses. For the 2SLS estimates, confidence intervals based on inversion of the Anderson-Rubin test are shown in brackets. * $p<0.10$, ** $p<0.05$, *** $p<0.01$

Table 6 IVLASSO Results for Initial Assessment of Most Severe Mental Health Rating and Suicidality Outcomes

	UJIVE		LASSO		Post-LASSO	
	(1)	(2)	(3)	(4)	(5)	(6)
Suicide attempt in next booking	-0.146***	-0.102***	-0.099***	-0.090***	-0.109***	-0.072*
	(0.034)	(0.029)	(0.033)	(0.032)	(0.039)	(0.043)
Suicide ideation in next booking	-0.018*	-0.012	-0.020***	-0.021***	-0.020***	-0.021***
	(0.010)	(0.010)	(0.004)	(0.005)	(0.004)	(0.004)
Next booking ADL score improves	1.012***	1.116***	0.469	0.526*	0.509*	0.577**
	(0.259)	(0.164)	(0.289)	(0.306)	(0.277)	(0.254)
Time fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Baseline controls	No	Yes	No	Yes	No	Yes

This table presents Unbiased Jacknife Instrumental Variables Estimator (UJIVE) and the Instrumental Variables LASSO (IVLASSO) estimates of the impact of a clinician's initial assessment of a most severe mental health rating on inmates' subsequent mental health. The outcome variables of interest are given in each row along with the corresponding estimates of the impacts of an initial assessment of a most severe mental health rating. We include day-of-week-month fixed effects and baseline controls for all specifications; however, the IVLASSO procedure penalizes the controls as well as the instruments and can penalize them to zero as discussed in the text. The IVLASSO procedure is run using two methods: lasso-orthogonalization and post-lasso-orthogonalization as shown in Columns (5) and (6). The clinician and inmate robust two-way clustered standard errors are shown in parentheses for IVLASSO and Kolesar's (2013) robust standard errors are shown in parentheses for UJIVE. * p<0.10, ** p<0.05, *** p<0.01

Table 7 No Prior Treatment Results for Initial Assessment of Most Severe Mental Health Rating and Suicidality Outcomes

	(1) OLS	(2) 2SLS	(3) JIVE	(4) UJIVE	(5) LASSO	(6) Post-LASSO
Suicide attempt in next booking	-0.016***	-0.122**	-0.175***	-0.083***	-0.049	-0.059
	(0.006)	(0.060)	(0.061)	(0.027)	(0.036)	(0.043)
Suicide ideation in next booking	-0.002	-0.014*	-0.029	-0.017*	-0.016***	-0.016***
	(0.003)	(0.008)	(0.021)	(0.009)	(0.004)	(0.003)
Next booking ADL score improves	0.136***	0.981***	-2.100	1.158***	0.511	0.580*
	(0.037)	(0.249)	(1.307)	(0.170)	(0.366)	(0.298)
Time fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Baseline controls	Yes	Yes	Yes	Yes	Yes	Yes

This table is analagous to Tables 5 and 6 except that the sample in this table is limited to individuals who reported no prior treatment as discussed in the text. Here we present the ordinary least squares (OLS), two-stage least squares (2SLS), jacknived instrumental variables (JIVE) (Angrist et al 1999), unbiased jacknived instrumental variables (UJIVE) (Kolesar 2013), and instrumental variables LASSO estimates of the impact of a clinician's initial assessment of a most severe mental health rating on inmates' subsequent mental health. The outcome variables of interest are given in each row along with the corresponding estimates of the impacts of an initial assessment of a most severe mental health rating. The 2SLS specifications instrument for severe mental health rating using a clinician leniency measure that is estimated using data from other cases assigned to a clinician. For further explanation, please see the text. We include day-of-week-month fixed effects and baseline controls for all specifications. The clinician and inmate robust two-way clustered standard errors are shown in parentheses. For the UJIVE estimates, we include Kolesar's (2013) robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01

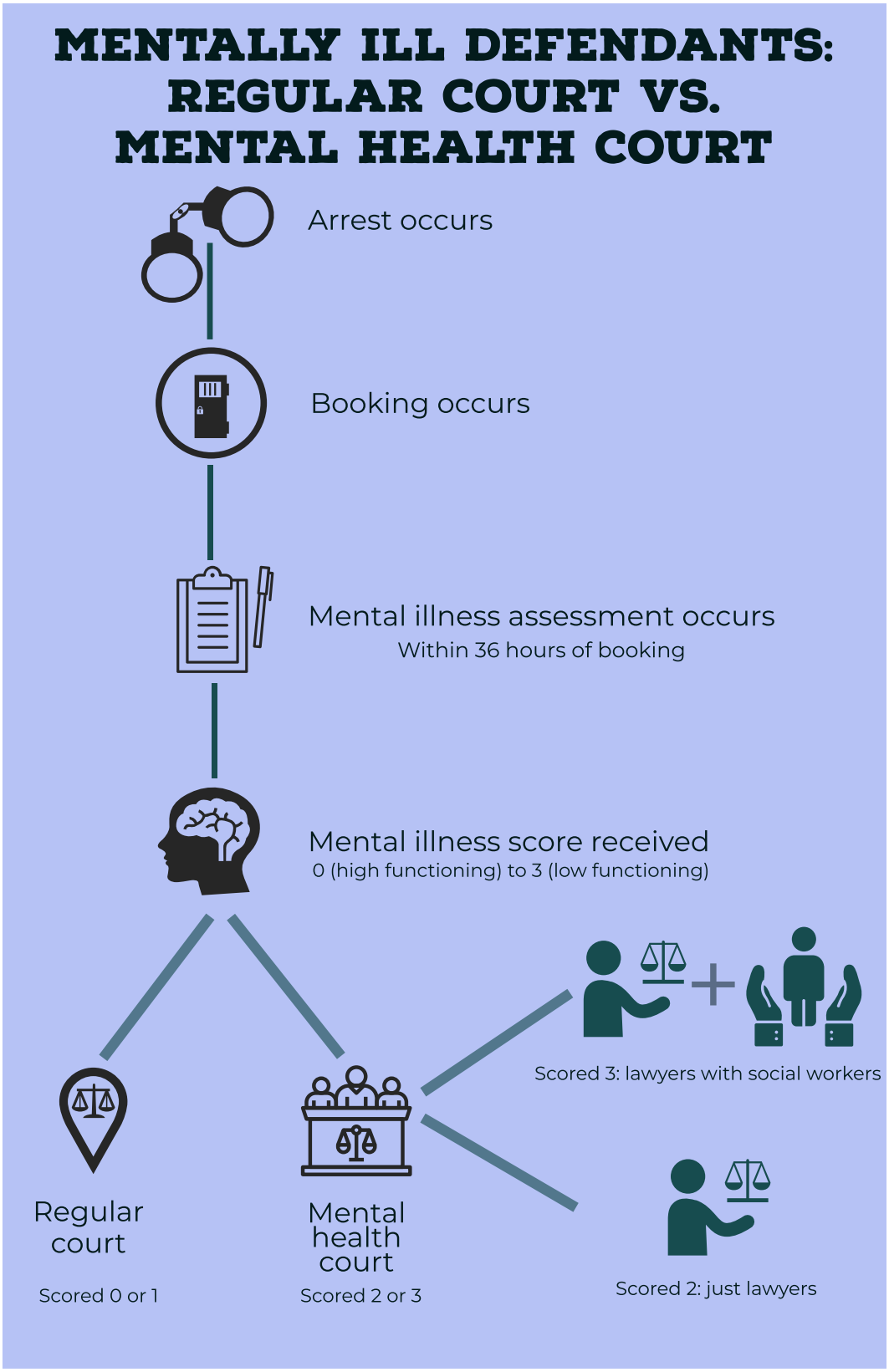

Figure 1 Assignment of inmates to mental health courts and attorney

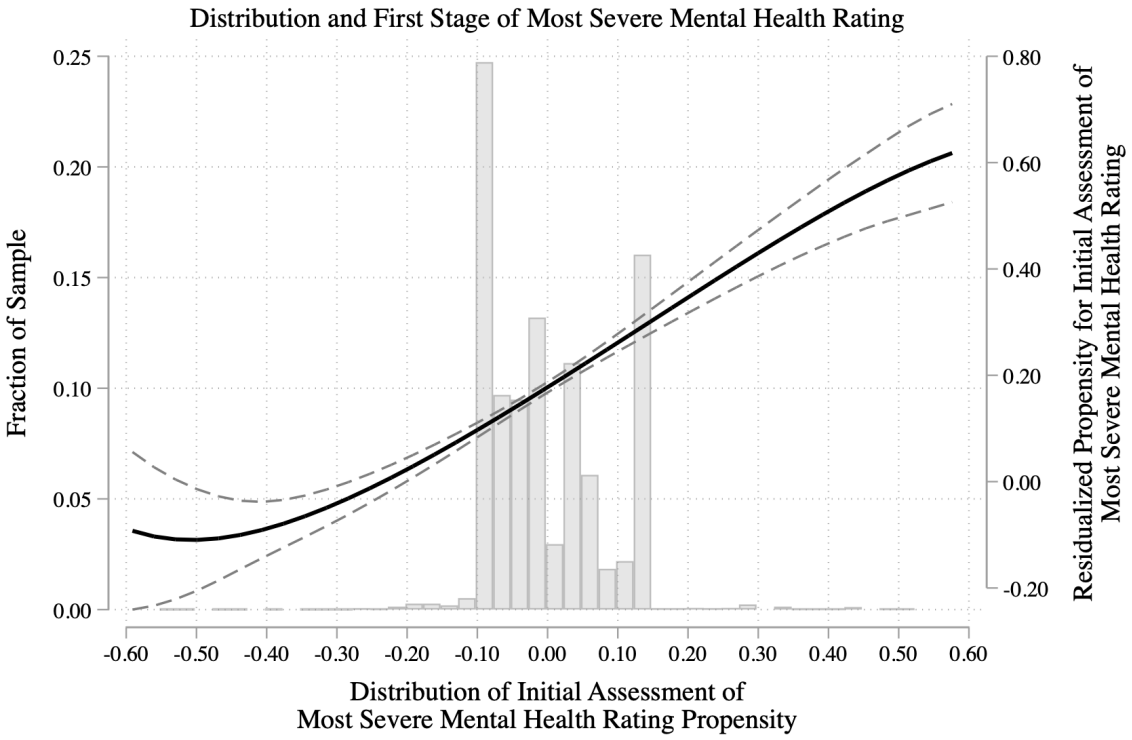

Figure 2 Smoothed fan regression of residualized leave one out against the share of individuals assessed with most serious mental health

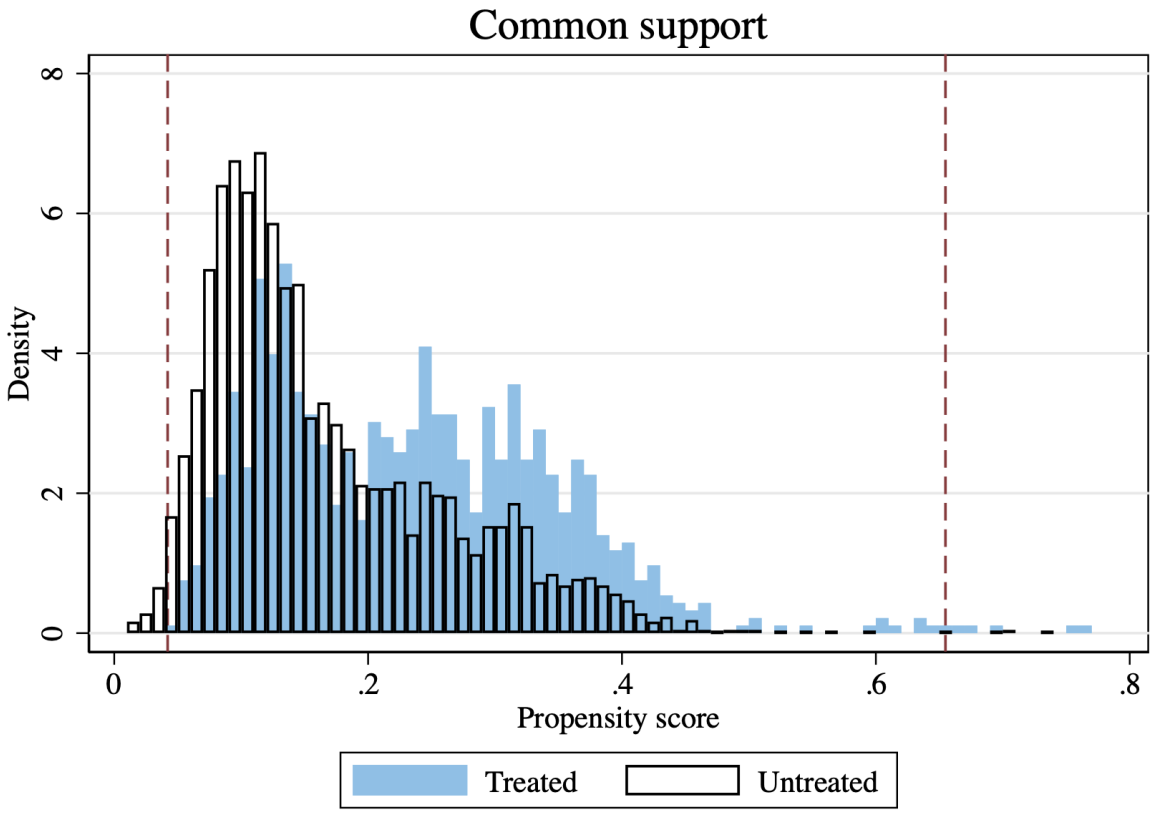

Figure 3 Propensity score distribution for high symptoms score

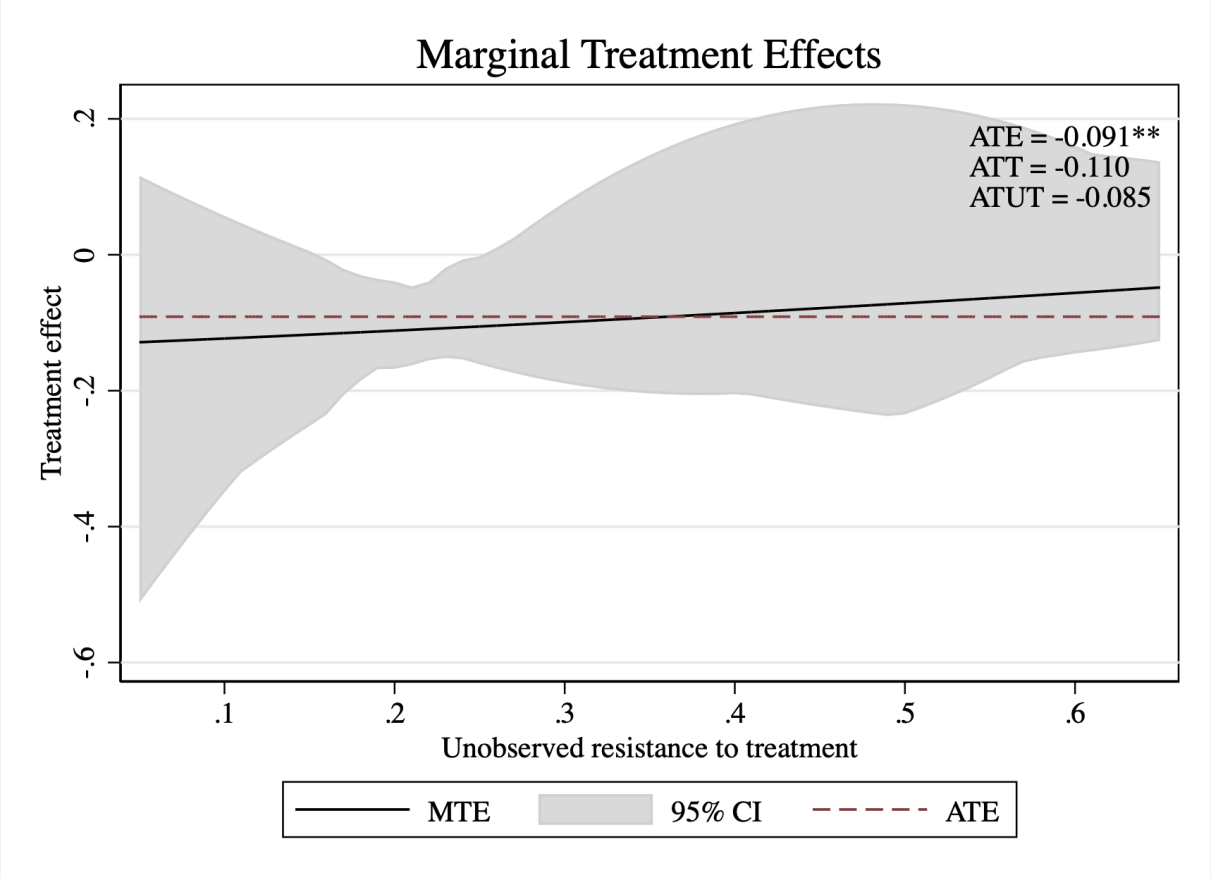

Figure 4 Marginal Treatment Effects for suicide attempt with second degree polynomial

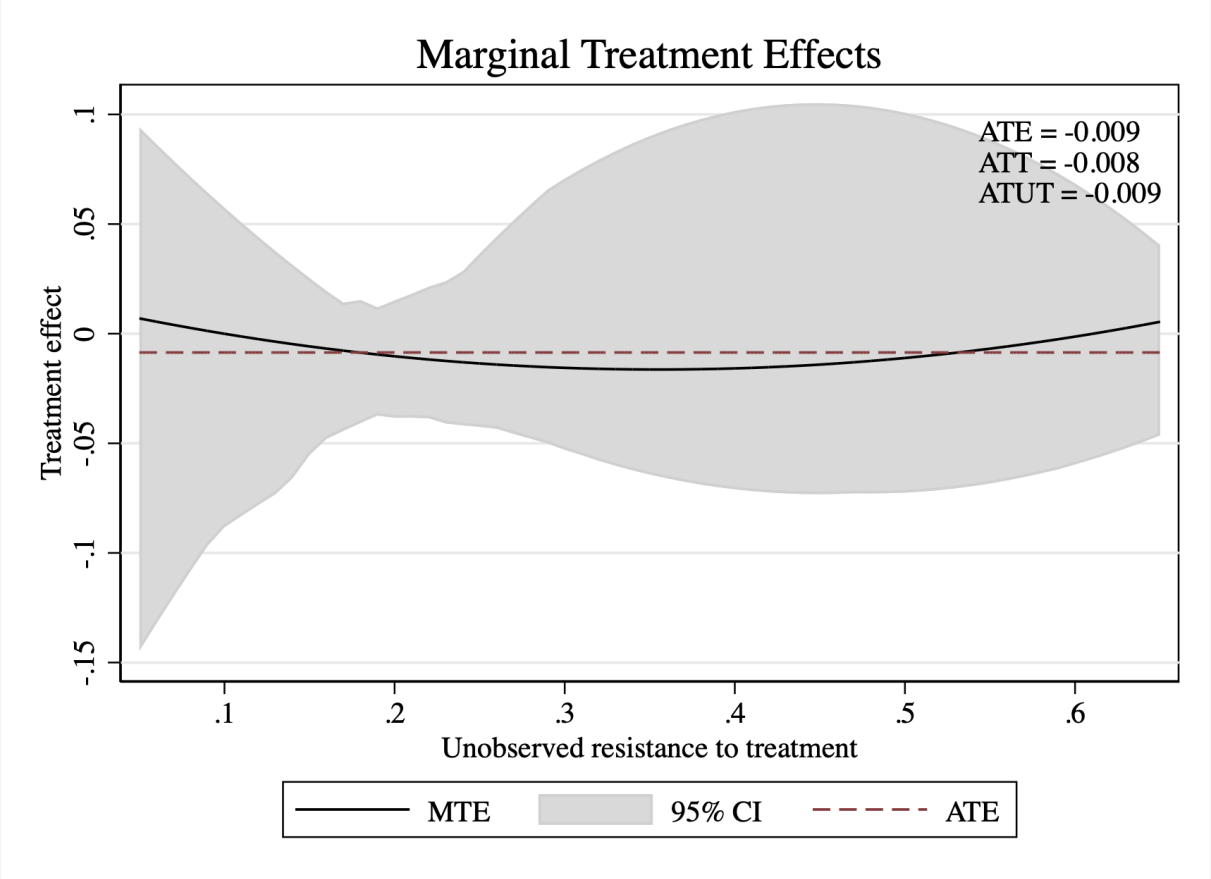

Figure 5 Marginal Treatment Effects for suicidal ideation with second degree polynomial

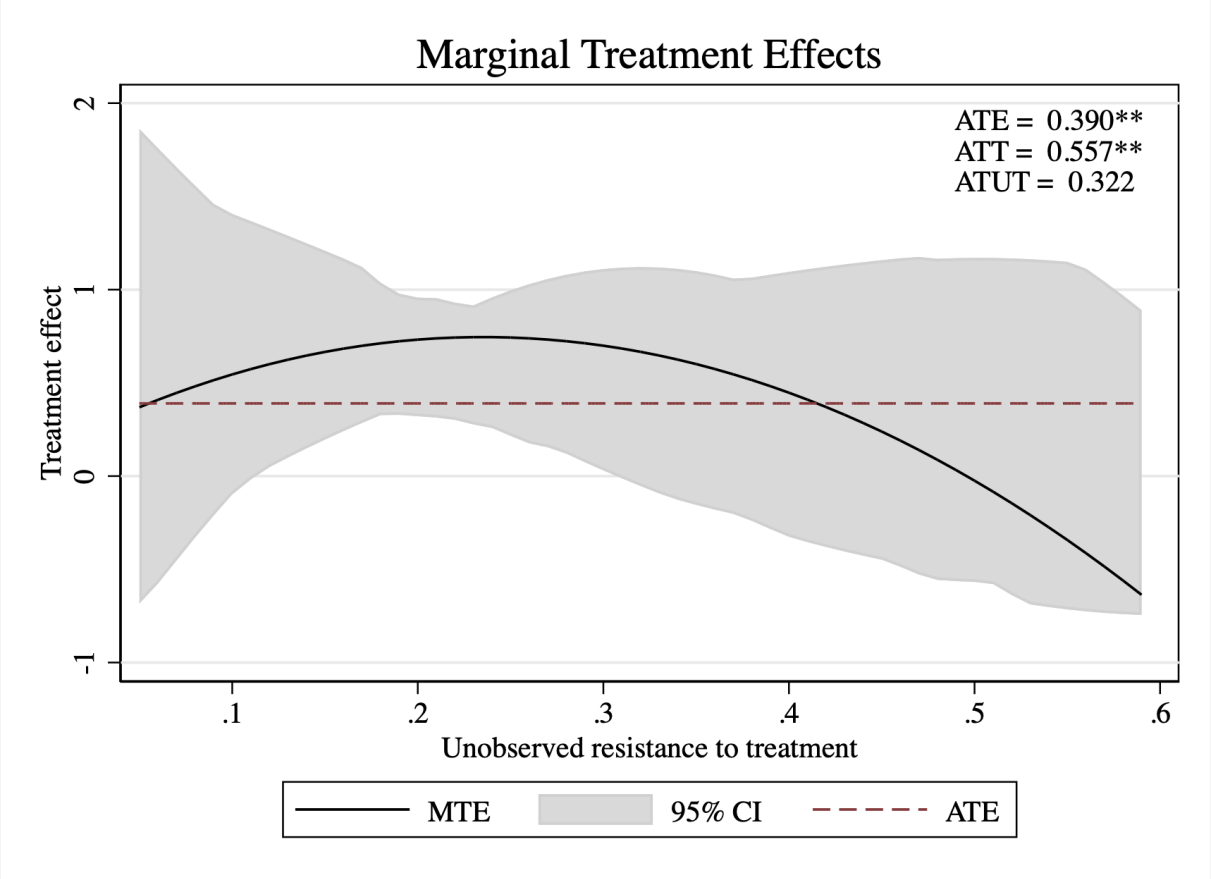

Figure 6 Marginal Treatment Effects for next clinician scoring of symptoms with second degree polynomial

A Appendix

A.1 Sample selection and collider bias

Our administrative data is rich in both outcomes and controls as well providing opportunities for identification using randomized clinicians. For instance measuring recidivism is straightforward since the administrative data assigns each resident in the county with a unique inmate ID. If John Doe was arrested twice, the same inmate ID will be assigned to him. Insofar as all offender in our samples are equally likely to be caught and remain in the county, then anytime someone reoffends, they will be arrested and therefore appear in our data.

Another interesting feature of our dataset is that we have various mental health measurements, such as the clinician reviews of each inmate upon subsequent booking, as well as records as to whether the inmate attempted suicide or displayed any suicidal ideation. But because we only observe suicidality and mental health scores within the administrative data, it means we only observe mental health outcomes for those inmates who reoffend. This is a potential problem because insofar as recidivism is endogenous to mental health court, then our sample – which is based on recidivism – may suffer from what is sometimes called "collider bias" (Pearl, 2009; Schneider, 2020; Cunningham, 2021), nonrandom sample selection (Heckman, 1979) and "bad controls" (Angrist and Pischke, 2009). We will illustrate this problem in Figure 2 using a directed acyclic graphical (DAG) (Morgan and Winship, 2014) describing a plausible data generating process creating complex relationships within our administrative data.[18]

Assume that assignment to the public defenders office due to having severe symptoms (D) has some causal effect on recidivism (R). Assume, too, that mental health court can affect mental health outcomes, such as suicidality (Y). This effect of serious mental illness on suicidality can happen both for those who reoffend (the mediated edge, $D \rightarrow R \rightarrow Y$) as well as for those who do not reoffend ($D \rightarrow Y$). We instrument for D with the residualized leave-one-out mean (Z), which alone is sufficient to block the backdoor path

[18] Our problem resembles the problem identified by Knox, Lowe and Mummolo (2020) in their critique of Fryer (2019). Sometimes administrative data can suffer from collider bias insofar as certain conditions hold.

between D and Y via controls ($ZMHC \leftarrow X \rightarrow Y$). This is because D when instrumented by Z is a "collider" and colliders by design eliminate spurious correlations due to backdoor paths (Schneider, 2020; Cunningham, 2021).

But notice the unobserved variables, U, which cause a person to reoffend for reasons other than mental illness. Insofar as these U unobserved variables also cause mental health outcomes in the jails, then working with an administrative dataset consisting only of reoffending individuals, R, will create collider bias between D and Y along a backdoor path represented by a chain of variables, $D \rightarrow R \leftarrow U \rightarrow Y$ (Morgan and Winship, 2014).

Such situations only occur, though, if D does cause recidivism. If public defenders have no effect on recidivism, then there is no collider problem because $MHC \rightarrow R$ does not exist. In such a situation, the direct edge $MHC \rightarrow Y$ will capture potentially the general effects of mental health court on mental health, including suicidality. The implications of this DAG is that insofar as there are ever any recidivism results, we cannot estimate the effect of public defense due to rated severe symptoms on mental health outcomes using the administrative data because the administrative data suffers from a sample selection version of collider bias. But, if there is no effect on recidivism, then public and private defender representation are equally likely to reoffend and therefore analyzing jail outcomes conditional on reoffending could be informative.

We discuss results from our analysis exploring the potential for this problem statistically in Table A1. None of our 2SLS models show statistical differences at conventional levels in recidivism probabilities across the lawyers with and without social workers treatment assignment. We conclude that while a collider problem with these day is possible, we can find no evidence for it in Travis county's mental health courts.

A.2 Figures

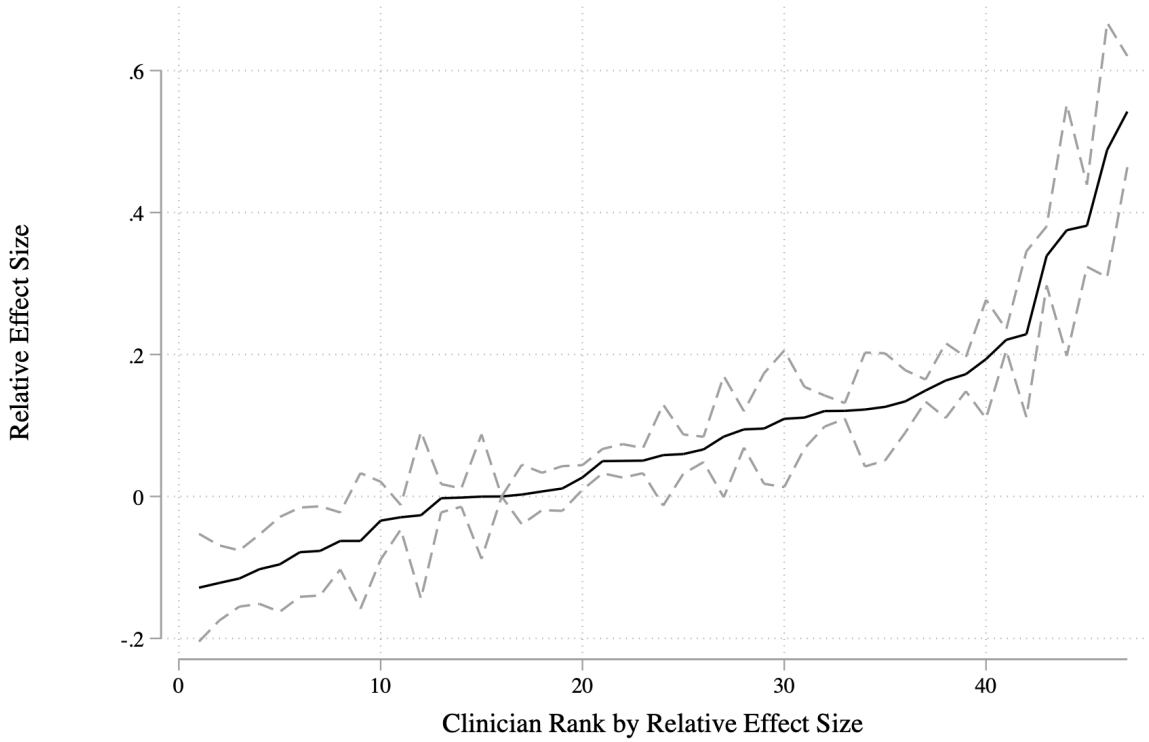

Figure A1 Evaluator fixed effects with unique clinician sample for inmates with high scores

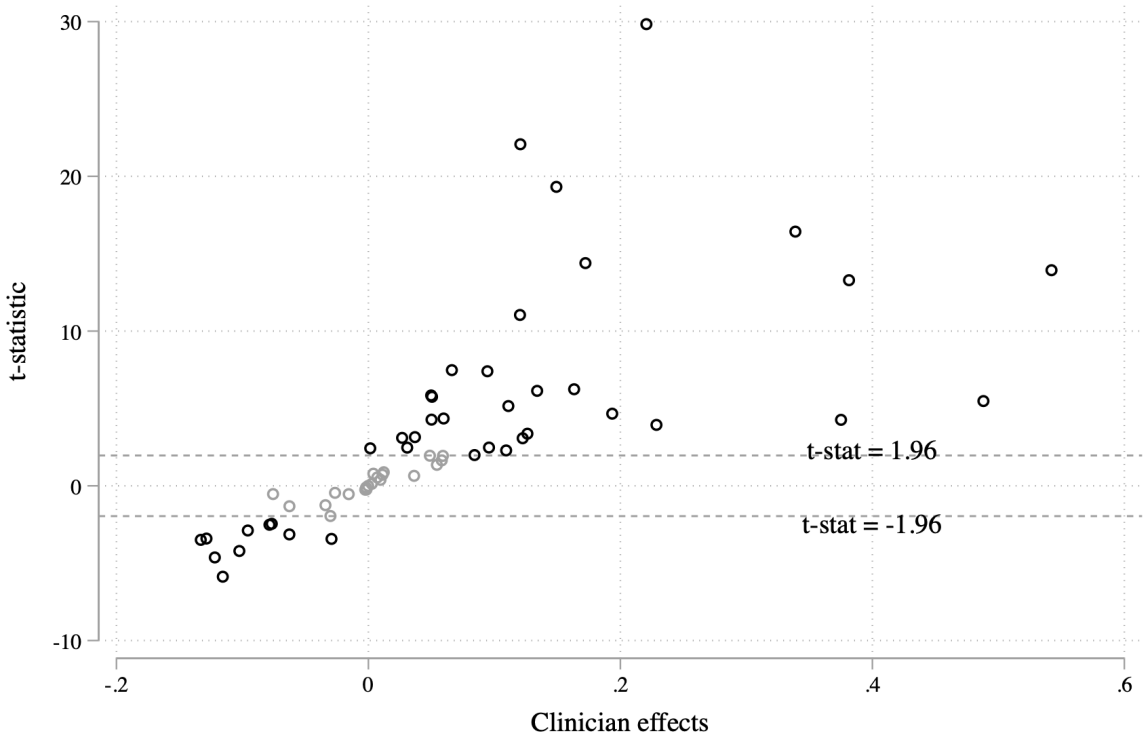

Figure A2 Distribution of t-statistics on individual evaluator fixed effects for unique clinician sample for individuals assessed with most serious mental health

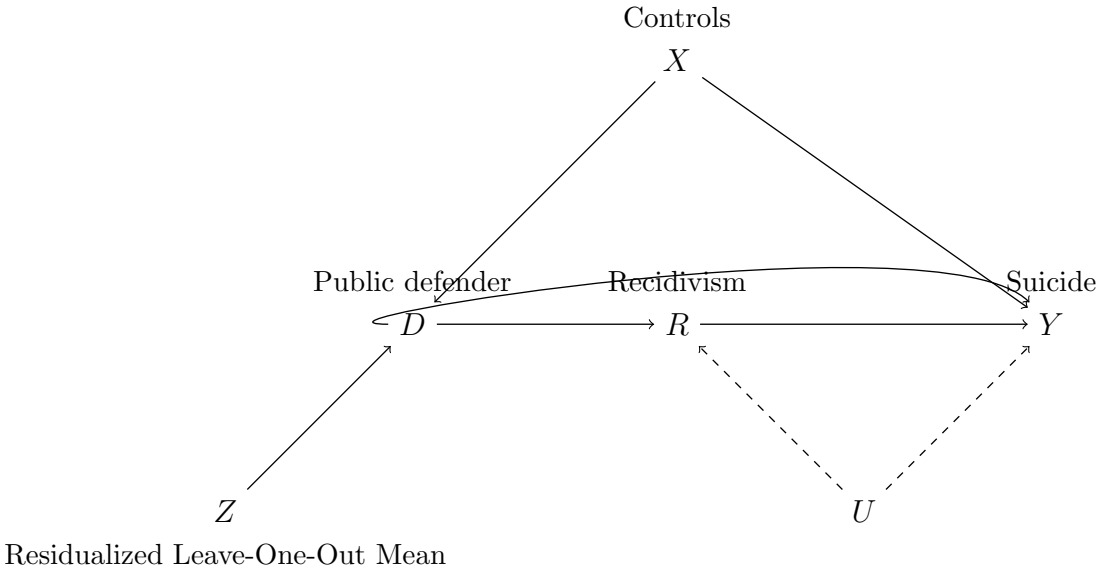

Figure A3 DAG showing sample based collider bias

A.3 Tables

Table A1 Effects of Initial Assessment of Most Severe Mental Health Rating on Recidivism Outcomes

	OLS results		2SLS results	
	(1)	(2)	(3)	(4)
Recid after current booking	0.053**	0.024	0.123	-0.016
	(0.020)	(0.019)	(0.182)	(0.145)
			[-0.283, 0.458]	[-0.311, 0.279]
Recid within 1 year	0.029	0.023	0.006	-0.064
	(0.022)	(0.023)	(0.156)	(0.143)
			[-0.400, 0.293]	[-0.437, 0.199]
Count of future recidivism	0.135	0.043	0.700	0.278
	(0.082)	(0.083)	(0.589)	(0.469)
			[-0.501, 1.901]	[-0.677, 1.324]
LOS	13.442***	12.344***	16.056	17.292
	(1.905)	(1.648)	(15.244)	(14.792)
			[-9.105, 58.977]	[-4.222, 58.885]
Days to recidivism	-20.884*	-18.330	29.171	-17.727
	(12.060)	(11.705)	(90.727)	(77.893)
			[-118.927, 229.537]	[-144.483, 153.767]
Next offense felony	-0.011	-0.020*	0.034	-0.020
	(0.010)	(0.010)	(0.083)	(0.075)
			[-0.151, 0.202]	[-0.171, 0.132]
Time fixed effects	Yes	Yes	Yes	Yes
Baseline Controls	No	Yes	No	Yes

This table reports the ordinary least squares and two-stage least squares estimates of the impact of a clinician's initial assessment of a most severe mental health rating on inmates' subsequent mental health. The outcomes are given in each row along with the corresponding estimates of the impacts of an initial assessment of a most severe mental health rating. Two-stage least squares specifications instrument for severe mental health rating using a clinician leniency measure that is estimated using data from other cases assigned to a clinician as described in the text. We include day-of-week-month fixed effects for all specifications and baseline controls for Columns (2) and (4)-(6). The clinician and inmate robust two-way clustered standard errors are shown in parentheses. For the IV estimates, confidence intervals based on inversion of the Anderson-Rubin test are shown in brackets. * p<0.10, ** p<0.05, *** p<0.01

Table A2 Initial Assessment of Most Severe Mental Health Rating and Heterogeneity in Outcomes

	Prior treatment		Prior medications		Prior hospitalization	
	(1) No	(2) Yes	(3) No	(4) Yes	(5) No	(6) Yes
Suicide attempt in next booking	-0.085*	0.486	-0.075	0.592	-0.090*	0.267
	(0.049)	(0.504)	(0.048)	(0.574)	(0.049)	(0.326)
Suicide ideation in next booking	-0.013**	0.051	-0.013**	0.060	-0.012**	0.040
	(0.006)	(0.087)	(0.006)	(0.095)	(0.006)	(0.066)
Next booking ADL score improves	0.903***	-6.008	0.904***	-4.811	0.875***	-7.145
	(0.223)	(24.778)	(0.222)	(23.310)	(0.226)	(24.786)
Time fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Baseline controls	Yes	Yes	Yes	Yes	Yes	Yes

This table explores the heterogeneity in effects on whether an inmate admitted to having prior treatment, prior medications, or prior hospitalization. The outcomes are given in each row along with the corresponding two-stage least squares estimates of the impacts of an initial assessment of a most severe mental health rating. We include day-of-week-month fixed effects and baseline controls, and the clinician and inmate robust two-way clustered standard errors are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Table A3 Initial Assessment of Most Severe Mental Health Rating and Heterogeneity in Outcomes

	Homeless		Jobless	
	(1) No	(2) Yes	(3) No	(4) Yes
Suicide attempt in next booking	-0.119**	-0.021	-0.107**	-0.004
	(0.053)	(0.165)	(0.054)	(0.255)
Suicide ideation in next booking	-0.013**	0.076	-0.018**	-0.101
	(0.007)	(0.077)	(0.008)	(0.079)
Next booking ADL score improves	0.945***	-1.189	0.925***	0.309
	(0.220)	(1.415)	(0.219)	(1.039)
Time fixed effects	Yes	Yes	Yes	Yes
Baseline controls	Yes	Yes	Yes	Yes

This table explores the heterogeneity in effects on whether an inmate had a lower subsequent mental health score during the booking or admitted to being homeless or jobless. The outcomes are given in each row along with the corresponding two-stage least squares estimates of the impacts of an initial assessment of a most severe mental health rating. We include day-of-week-month fixed effects and baseline controls, and the clinician and inmate robust two-way clustered standard errors are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Table A4 Joint Test of Exclusion and Monotonicity

Outcome	FLL P-Value
Suicide attempt in next booking	0.000
Suicide ideation in next booking	1.000
Next booking mental health score improves	0.267

This table presents results from the Frandsen, Lefgren, and Leslie (2020) test for the joint null hypothesis that the monotonicity and exclusion assumptions hold. Where we fail to reject the null hypothesis, then we cannot conclude that the monotonicity and exclusion assumptions jointly hold. We used the package testjfe in Stata to test these assumptions (Frandsen, 2020). The outcomes and corresponding p-values are given in each row.

Table A5 Average Montonicity for Initial Assessment of Most Severe Mental Health Rating

	Male	Female	Black	White	Hispanic	Age < 25	Age > 45
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Z: Clinician's Leave-Out Mean Mental Health Score	0.562*** (0.164)	0.766*** (0.152)	0.899*** (0.179)	0.547*** (0.147)	0.556*** (0.200)	0.598** (0.254)	0.568*** (0.170)
Observations	3,355	1,860	1,371	3,790	1,097	1,031	1,219
Time Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes

This table reports the first stage results by subsamples as listed in the column headers, which serves as informal evidence of average monotonicity if the estimate is significant across all subsamples. * p<0.10, ** p<0.05, *** p<0.01