Evaluation of Utah's Mental Health Courts: Limited Findings and a Review of Research Challenges

THE UNIVERSITY OF UTAH

Utah Criminal Justice Center

COLLEGE OF SOCIAL WORK
COLLEGE OF SOCIAL & BEHAVIORAL SCIENCES
UTAH COMMISSION ON CRIMINAL & JUVENILE JUSTICE
S.J. QUINNEY COLLEGE OF LAW

Evaluation of Utah's Mental Health Courts: Limited Findings and a Review of Research Challenges

Kort Prince, Ph.D
Allyson Walker, Ph.D
Erin B. Worwood, M.C.J.
Jess Shade, MA
April O'Neill, MSW
Robert P. Butters, Ph.D

March 2018

Utah Criminal Justice Center, University of Utah

Table of Contents

Acknowledgments

Funding for this project was provided by the state of Utah's Commission on Criminal & Juvenile Justice (CCJJ). Support for project planning was provided by Ben Peterson, PhD, Director of Research & Data at CCJJ. Data for the project were provided by the Utah Bureau of Criminal Investigation (BCI), the Department of Human Services (DHS), and the Utah Department of Corrections (UDC). Please direct concerns, comments or other correspondence to Kort Prince (UCJC Research Assistant Professor; kort.prince@socwk.utah.edu).

Executive Summary

Individuals with mental illness are disproportionally incarcerated within the United States. In order to combat this problem, Mental Health Courts (MHCs) have been developing nationwide since the 1990s. These courts aim to divert defendants with mental illness onto separate court dockets that allow them to avoid incarceration, while simultaneously steering them toward mental health treatment services and supervision. The goal of this model is to promote mental wellness among defendants, as well as to reduce recidivism among a population who may be engaging in offending behavior due in part to mental illness.

While MHCs intend to reduce recidivism and promote wellbeing among participants, whether they succeed in this endeavor has been difficult to ascertain with empirical research. MHCs are difficult to compare to one another, as they are subject to local standards, structure, and data sources and thus have different requirements for participation and completion of programs. Missing data is another common challenge in MHC research, as the data necessary to adequately study the effects of MHCs include mental health and criminal justice histories for each defendant, as well as risk assessments, and mental health and criminal justice outcomes. The sources of these data come from many locations, including departments of mental health, correctional departments, individual courts, and individual treatment providers, making attainment of these data a challenge. Missing data and lack of ability to compare courts to one another can contribute to an inability to properly assess MHC effectiveness.

The Utah Criminal Justice Center (UCJC) at the University of Utah was tasked with conducting a two-part study of Utah's MHCs. The first part (Worwood, Sarver, Borgia, & Butters, 2015) consisted of a review of MHCs throughout the United States and internationally as well as in Utah; the second part, described in this report, was intended to assess the efficacy of Utah's MHCs using a matched sample of individuals with similar treatment and criminal histories. Four jurisdictions were considered for the evaluation, as these were the only jurisdictions with a sufficiently large sample of cases: Utah, Salt Lake, Cache, and Weber.

Propensity Score Matching (PSM) was used in an attempt to identify comparison groups based on variables that predict the likelihood of being placed in the treatment group (in this case, the likelihood of participation in the MHC). A large number of PSM algorithms were employed in an attempt to obtain a matched comparison sample. Unfortunately, due to a high degree of missing mental health history data, PSM could not identify an appropriate comparison group.

Because the study was unable to implement the intended methods to examine MHC outcomes, the MHCs were evaluated using an alternative method. Using pre-post data for MHC clients only (i.e., data about participants from before and after participation in MHCs), participant data was assessed regarding defendants' offending behavior in all four MHCs, and mental wellness for the Salt Lake and Utah MHCs. Each court was evaluated separately for both recidivism and mental health outcomes. Scores from the Outcome Questionnaire (OQ-45 – a measure of psychological functioning used to evaluate change over the course of treatment) were used to describe defendant wellbeing over time, while Utah Bureau of Criminal Identification (BCI) data were used to describe defendants' offending trajectories.

The results of these analyses showed some promise for the Utah MHCs. Trajectories for OQ-45 scores revealed a trend toward better functioning for both courts examined. While this was a favorable pattern, the trend began prior to MHC participation and merely continued at the same rate after MHC start, making the findings regarding MHC effects on mental wellness unclear. Favorable patterns were found in outcomes of offending: offending trajectories were significantly reduced after MHC participation began, and this result remained true after accounting for available data on incarceration.

The pre-post, within-groups design is less rigorous than the originally proposed PSM matching design, and has greater interpretational limitations. For example, it is possible that, if a non-MHC comparison group had been identified, the comparison groups would have experienced the same decline in offending and trend toward better functioning that MHC participants experienced. Comparison data would be needed in order to make a confident statement about the comparative effects of MHCs studied.

Given the inability to implement the intended PSM design, some of the challenges that future studies would face are discussed. In the case of the current study, it was anticipated by UCJC and stakeholders alike that DHS was, in fact, a centralized source of mental health treatment data, but, at least for the population of MHC cases and those considered as comparison cases, that was not the case. Any future evaluation of MHCs in the state of Utah will have to address this lack of centralized data. Future research will also need to consider the diverse range of clients MHCs serve and the diverse range of service providers MHCs allow for treatment, as these impact the ability to evaluate the efficacy of MHCs. In light of the challenges faced when studying MHCs, as well as lessons learned from the current study, two possible methodologies for future studies are discussed.

Background

Individuals with mental health disorders are highly overrepresented within United States correctional facilities (Prins, 2014). Since the Community Mental Health Act was passed in the 1960s, long-term, residential state hospitals for mental health treatment were shut down across the nation. These state hospitals were viewed as warehouses for individuals with mental illness, and frequently had improper funding and care for those who were "institutionalized" in them. While the aim of this deinstitutionalization was to shift mental health care from these large, state hospitals to communities, instead, the United States prison system has become the number one provider of mental health services: the National Alliance on Mental Illness has estimated that at least 400,000 inmates across the United States have a mental illness (Ford, 2015).

What are Mental Health Courts?

Mental Health Courts (MHCs) are problem-solving courts that have been implemented across the nation in order to ameliorate this issue. These courts have been developed to divert non-violent defendants with mental health disorders away from the criminal justice system and into community-based treatment, monitored by the court. When offending behaviors are attributable to mental illness, the MHC aims to prevent reoffending by treating the issue underlying this behavior, while simultaneously avoiding conventional punishment. In providing court-monitored treatment, MHCs also hold mentally ill defendants accountable for engaging in necessary treatment services (Berman & Feinblatt, 2001; Berman & Fox, 2010; Goldkamp & Irons-Guyunn, 2000).

MHCs vary widely in both populations and procedures, but they do have some similarities, including: a docket for defendants with mental illness that has been separated from the traditional court docket; ability of defendants to opt into the MHC (with the alternative of being processed through the traditional court system); dedicated judges, prosecution, and defense attorneys; collaboration between court staff and mental health providers in determining treatment plans and justice system-related repercussions; court monitoring of treatment processes and outcomes; and justice system-related rewards for completion of program requirements, including potential dismissal of charges or averting incarceration (Campbell et al., 2015; Goldkamp & Irons-Guyunn, 2000).

Although MHCs have been developed to work with a common type of population (justice system-involved individuals who have mental health disorders), the exact populations involved with MHCs, and the procedures with which MHCs handle them, vary widely by individual court. In a systematic review of 38 MHCs targeted for adults (Worwood, Sarver, Borgia, & Butters, 2015), the majority of MHCs accepted defendants who had committed either misdemeanors or felonies. Defendants who had committed violent felonies were excluded from about over half of these courts. Other courts included those with violent charges only on a case-by-case basis. Several courts only allowed defendants who had committed misdemeanor offenses or infractions, excluding defendants who had committed a felony offense entirely. Additionally, one court *only* accepted defendants who had committed felonies.

This review also showed variety in which mental diagnoses were considered acceptable among defendants. The majority of courts accepted defendants with psychosis, schizophrenia, and/or depression. Half of the courts accepted defendants with anxiety disorders. Lower numbers of MHCs accepted defendants with other mental health issues, such as traumatic brain injury, personality disorders, developmental disabilities, attention deficit/hyperactivity disorder, or substance-related disorders, unless they were paired with other accepted disorders (Worwood, Sarver, Borgia, & Butters, 2015). Procedures engaged in by MHCs show no less variety than defendant eligibility criteria. Programs differ widely in terms of when in the court process defendants can enter an MHC (i.e., before or after a plea); types of supervision or case management defendants are placed under (e.g., probation, electronic monitoring, intensive case management); and types of services that MHCs can offer to defendants (e.g., housing, employment, education, financial).

Methodological Challenges in the Study of MHCs

Since the inception of MHCs in the late 1990s, funding sources, judicial agencies, and mental health advocates have sought insight into their efficacy. Similar to the early years of drug court research (Wolff & Pogorzelski, 2005), however, MHC research remains highly dependent on local MHC funding, standards, structure, and data sources. Each MHC operates within a local cultural context without extant standards linking all MHCs together. Despite numerous studies of MHCs, there remains no reliable statistic on their overall effect size in regard to reducing recidivism or improving the mental health of court participants. This is due to pervasive methodological challenges posed to MHC research. These challenges include issues in creating comparative treatment and comparison groups, as well as a host of data-related issues.

Treatment and Comparison Groups

A major methodological challenge within studies of MHCs is establishing comparable treatment and comparison groups. Treatment groups are comprised of individuals participating in a MHC while comparison groups include individuals whose cases are processed through standard courts. One review found that most extant studies lack experimental design based on their inability to adequately determine randomized, matched treatment and comparison groups (Honegger, 2015). Often limited at best to a quasi-experimental design (Sarteschi, Vaughn, & Kim, 2011), researchers must first identify MHC participants to include in the treatment group. Due to the variability between MHCs and frequent lack of standardized MHC participant inclusion criteria (Honegger, 2015; Steadman, 2005), treatment groups between different studies can vary widely. This heterogeneity of clients makes consistent mental health treatment plans, resources, and outcomes very difficult to quantify even with consistent data from individual MHCs.

Once a treatment group has been identified, researchers must then attempt to identify an appropriate comparison group to determine if the MHC treatment has effects on recidivism and mental health outcomes. Studies vary as to which covariates are utilized to match these groups. Most evaluations involve matched designs based on "demographic, sociodemographic, and in some cases, clinical characteristics," which are presumed, not proven, to be linked with recidivism (Wolff & Pogorzelski, 2005, p. 546). The variables used as covariates in these studies can vary considerably depending on the availability of data across both criminal justice and mental health sources (McGaha, Boothroyd, Poythress, Petrila, & Ort, 2002). Due to the frequent lack of available data to use as covariates indicative of mental health for the comparison group, researchers often resort to utilizing a comparison group matched based on those who committed the same crimes (e.g., Lowder, Desmarais, & Baucom, 2016), or sometimes fail to establish a comparison group at all (e.g., Hiday, Wales, & Ray, 2013). Establishing adequate case and comparison groups is difficult due to the number of variables to be matched and availability of the data (Steadman, 2005).

Frequent changes to court dockets, the heterogeneity of mental illnesses, and the severity associated with different forms of referral groups (e.g., more severe clients would be civilly committed to treatment versus those deemed eligible for MHC), serve to further complicate the process of establishing and matching treatment and comparison groups (McGaha et al., 2002). Ultimately, for these reasons, most extant studies of MHCs fall victim to a lack of statistical control between MHC and comparison participants (Sarteschi et al., 2011). The lack of experimental design and variability of MHCs therefore make meta-analyses to determine effect sizes of MHC intervention tenuous at best.

Data

Along with establishing appropriate treatment and control groups, several challenges have been identified in the literature regarding the collection of adequate data in order to analyze the effects of MHCs. Securing reliable, complete, and consistent data across diverse sources poses an ongoing challenge for researchers throughout the evaluation of MHCs. Four main categories of data required for MHC evaluation include: participants (i.e., numbers screened, eligible, and accepted as well as demographic information), services (i.e., type and length of treatment), criminal justice outcomes (i.e., subsequent arrests, charges, and days in correctional facilities), and mental health outcomes (i.e., quality of life, hospitalizations, and access to ongoing treatment) (Steadman, 2005). Given the number of variables and outcomes for each participant, data quality, consistency, and completeness is a critical issue in determining MHC effects (Sarteschi et al., 2011; Wolff & Pogorzelski, 2005).

Treatment Characteristics

Mental health diagnoses, treatment, and outcomes often present the most common areas of missing data. These data are often held by different treatment providers and are frequently not included in official court records. Completing the program requirements of MHC requires compliance to treatment standards which frequently vary between MHCs. The guidelines for completing treatment are not necessarily linked to individualized treatment plans and often include arbitrary timeframes for completion as established by the court (Wolff & Pogorzelski, 2005). Additionally, sample representativeness often could not be determined based on lack of explicit guidelines for inclusion in MHC as well as information about who is excluded or drops out of treatment (Wolff & Pogorzelski, 2005). Therefore, when available, treatment data are typically operationalized in a fairly limited manner (Honegger, 2015). According to one meta-analysis, no studies have measured the actual quality of treatment received across multiple providers within MHCs (Sarteschi et al., 2011).

Outcomes

Ultimately, MHCs operate based on the presumption that a reduction in mental health symptomology yields a reduction in recidivism. Based on the paucity of consistent mental health data across studies, however, this remains hypothetical. According to multiple reviews, outcome data for mental health statuses are frequently unavailable to researchers leading them to rely on recidivism data alone to determine effect sizes (Hiday et al., 2013; Honegger, 2015; Lim & Day, 2014). When mental health outcome data were available, symptomology was often measured in the context of single assessments such as the Brief Psychiatric Rating Scale or based on the participant's number of psychiatric hospital days (Honegger, 2015). These measures present a very limited clinical picture of the study participants as they are often divorced from content and duration of the therapeutic interventions and are unrelated to comprehensive and ongoing measures of clinical improvement.

Additional concerns related to outcome data include smaller sample sizes due to attrition and the length of follow-up. Participant attrition can stem from a variety of sources and presents a common issue for researchers. One study lost over one third of study participants due to the inability to locate them for follow-up, death, and incompletion (Cosden, Ellens, Schnell, Yamini-Diouf, & Wolfe, 2003). Also, with many studies including follow-up periods of less than a year (Honegger, 2015), there is a debate about the generalizability of study results (Wolff & Pogorzelski, 2005).

Solutions to Common Issues

Many studies, as discussed above, did not employ mental health variables as covariates in order to match groups and did not track mental health outcomes based on a lack of mental health data. These studies examined MHCs solely based on recidivism. For researchers who were able to track mental health variables, the data were stored in a centralized state database of all community mental health providers (Lowder et al., 2016). This prevented the mental health data from being fragmented between a variety of public and private mental health providers.

Additionally, several studies used propensity score matching (PSM) to minimize bias between MHC cases and potential comparison group cases. This statistical approach helps address the inherent issues in non-random assignment to treatment and comparison groups. Additional detail on the methodology of PSM matching is discussed below under the section covering the current study.

The Current Study

Utah's first MHC started in 2001 in Salt Lake County. Between 2004 and 2011, five other adult MHCs have been established in the state: Utah, Washington, Cache, Weber, and Davis counties. Despite the length of time these courts have existed in Utah, little is known about outcomes related to MHC participation.

In 2014, the Utah Criminal Justice Center at the University of Utah, in partnership with the Utah Commission on Criminal & Juvenile Justice (CCJJ; a State entity), conducted a review of mental health court outcome studies across the United States and examined available data in Utah specifically. One purpose of the review was to help the State of Utah identify the features of these courts, including those related to eligibility criteria (e.g., accepting only clients with felonies, misdemeanors, or both), participant characteristics (e.g., types of diagnoses accepted), and program characteristics (e.g., mental health treatment, substance use treatment, housing, employment, or educational services).

A second, long-term goal of this preliminary study was to subsequently use the information obtained to help guide a second study, looking at the efficacy of Utah's MHCs. The initial design of the second study, described in this report, involved obtaining a matched comparison sample against which to compare the long-term outcomes of MHC clients.

The original purpose of the second study was to compare MHC and non-MHC cases using propensity score matching (PSM; Rosenbaum & Rubin, 1983). PSM attempts to account for the factors that predict receiving treatment (being an MHC client in this case). A treated case is said to have a certain propensity (or probability) of receiving treatment given their background factors; that is, a treated person has certain factors that make receiving treatment more likely. If one can identify the factors that predict treatment, one can, ideally, identify a group of individuals who are similar in likelihood of receiving treatment, but who did not, in fact, receive treatment. By comparing a group who received treatment to another group, with similar background characteristics, who did not, one can approximate the ability of randomized control trials with respect to assigning causality between treatment and outcomes (Guo & Fraser, 2009) – that is, one can reveal effects of an intervention (in this case, the MHC) by comparing the outcomes between the treatment group and the matched group who did not receive treatment.

Ultimately, the current study failed to obtain a sufficiently similar group against which to compare MHC cases. Reasons for that failure are documented in this report. Given an inability to obtain a representative comparison group, the current study turned to a less ideal (and less rigorous) design, examining changes in client well-being and potential reductions in recidivism using pre-post comparisons only. The limitations of this design are outlined in the pages that follow, as are some recommendations that might benefit future research.

Data

Utah Court records served as the starting point for identifying Mental Health Court (MHC) clients and potential comparison cases. The court data request submitted by the Utah Criminal Justice Center asked for identifiers (e.g., name, date of birth) and all minute entries (i.e., records from court minutes that detail what occurred each time an individual appeared in court) for all District Court cases with a sentence date between January 1st, 2011 and December 31st, 2012[1]. This period was selected because it allowed for inclusion of the most recently established MHCs in Utah and, also, allowed for a sufficient follow up period during which changes in well-being and observation of recidivism could occur. The request resulted in 67,373 cases and 3,755,379 minute entries that were evaluated for reference to MHCs.

[1] Though sentence dates in this timeframe served as the starting point, MHC cases were ultimately selected based on their mental health court agreement date. For that reason, cases with agreement dates falling slightly outside of this two-year period were permitted in the study.

Raw data provided by the District Court data lacked an indicator of which cases were processed through MHCs. In order to identify court data related to MHCs specifically, code was written in the R (R Core Team, 2017) programming language to search minute entries for terms such as "Mental Health Court" or various acronyms and abbreviations of the term (e.g., "MHC", "MH Court" "Health Crt", "MH Crt"). The search code was intentionally left broad to be as inclusive as possible. Cases that were screened but not accepted, deemed not appropriate, did not qualify, etc., were removed later during a manual cleaning phase.

Comparison group cases were identified via a similar process, but the search code allowed for broader terms related to mental health, such as orders for mental health treatment, counseling, therapy, or medication management. Minute entries mentioning counseling or treatment had to include terms related to mental health; that is, orders for substance abuse treatment or counseling without mention of mental health were not deemed sufficient to include an individual in the potential group of comparison cases. Abbreviations (e.g., "Tx" for treatment) were accounted for in the code.

Only four of Utah's Mental Health Courts, preliminarily, had a sufficient number of cases in the designated two-year timeframe to allow for an outcome evaluation: Cache, Salt Lake, Utah, and Weber. Each of these courts was sent the list of identified MHC cases for verification[2] and was asked to provide additional case numbers if their records indicated cases were missing. Only two of the courts, Weber and Salt Lake, responded to the request. While both Weber and Salt Lake were able to provide additional cases, the majority of those were duplicate cases for an individual already included in our case list. This indicates that the process described above for identifying MHC cases from court minute entries was largely successful. However, because Utah and Cache MHCs did not respond to the request, cases may have been missed in those jurisdictions. This seems unlikely, though, because the process of identifying cases from minute entries worked well in Weber and Salt Lake, and there is no reason to suspect it would be less accurate in the Utah and Cache MHCs.

In total, 806 cases were originally identified as possible MHC cases based on court minute entries or individual MHC records. Though most of these cases had mention of mental health court in their court minute entries, only 616 were actually screened for inclusion in the MHC, while 400 were accepted by the MHC program, and 393 entered a plea, and therefore entered the MHC. After removing duplicate cases by taking the first qualifying case, 350 MHC cases remained for analyses: 18 in Cache, 258 in Salt Lake, 51 in Utah, and 23 in Weber. Comparison cases were identified within each court. For example, MHC cases in Cache were compared with non-MHC cases also in Cache. This was necessary because, as will be seen below, the courts were fairly dissimilar in whom they served.

In addition to court data, evaluating the effects of MHC required obtaining data regarding treatment services and outcomes. It also required data regarding criminal histories. Because data regarding the criminal histories and treatment received by MHC participants were not available from the courts themselves, agency-specific data requests were next sent for the 350 MHC cases

[2] The individual MHCs were not used as the primary starting point for identifying MHC cases because MHC records lacked certain variables that were only available in state court records. Additionally, their records could not be used to identify potential comparison cases.

and 1,590 potential comparison cases. Identifiers were sent to the Department of Human Services (DHS), the Utah Bureau of Criminal Identification (BCI), and the Utah Department of Corrections (UDC). DHS was asked to provide records related to substance abuse and mental health treatment services and diagnoses, while both BCI and UDC were asked to provide criminal histories and variables related to offending (e.g., prison sentences, risk scores, case planning notes).

While both DHS and BCI provided data relatively quickly, data from UDC were delayed considerably owing to legal obstacles pertaining to the drafting of new data sharing agreements. Data from UDC did not arrive until less than one-month before the final report was due. Because of the length of time it takes to clean and aggregate data before analysis, the delay meant UDC data could not be considered in most of the analyses that follow. UDC data were used for two purposes: 1) to account for time in prison for recidivism analyses and 2) to describe the criminal risk levels of clients.

MHC Client Characteristics Pre-MHC

Before describing the results of outcome analyses, it is worth discussing the types of clients served by the four MHCs in the study. This helps explain why it was necessary to examine the courts separately. While the courts all had a common goal, they were quite disparate in terms of the mental health and criminal histories of those they served. Additionally, any analysis that combined the courts would have been heavily dominated by Salt Lake's MHC, its clients, and their outcomes, as it provided 258 of the 350 (73.7%) cases.

Criminal History and Criminal Risk

Table 1 provides a summary of the types of offenses for clients in each of the MHCs during the three-year period immediately preceding MHC participation[3]. Criminal Histories in the table are derived from BCI data, specifically National Crime Information Center (NCIC) offense codes. Because most MHC clients were found in BCI data regardless of the MHC to which they belonged, it is acceptable to compare across the MHCs. Absence of a specific offense likely indicates the person was not charged with the offense rather than representing a discrepancy in reporting across jurisdictions (which occurred for mental health diagnoses and services – discussed later).

Codes in BCI data are grouped into 30 offense categories. For this MHC study, some of these offenses (e.g., smuggling, homicide) are rare and are omitted from the table to make the table easier to read. Any offense for which at least 10% of clients in any MHC were charged is provided in the table (even if other courts had less than 10% of clients committing the offense). The first

[3] For purposes of this report, charges/offenses rather than convictions are used to describe client characteristics. While use of charges (whether ultimately convicted or not) and actual convictions both have pros and cons, it was ultimately decided that charges provided the most accurate representation of MHC client behavior. Because these clients often had extensive criminal histories, misrepresenting that history with the inclusion of unsubstantiated charges was considered less of a threat to validity than was the misrepresentation that can occur with use of only convictions due to factors other than innocence (e.g., pleas and diversion efforts) that impact case dispositions. While findings of guilt or innocence are of central importance with respect to dispositions, their meaning is less clear in the current analyses: simply put, the fact that a charged offense did not result in a conviction does not necessarily indicate that it did not occur in some form.

offense, “Any Charge”, incorporates all rows that follow. Similarly, “Any Felony Charge” incorporates all charges that follow when those charges were at the felony-level.

While all MHCs were similar with respect to the likelihood of being charged with any offense in the three-year period pre-MHC[4], notable discrepancies existed for most other categories. MHC clients in Weber were notably more likely to have charges for felony offenses, as well as controlled substance and liquor offenses. Both Salt Lake and Weber clients were more likely than clients of the two other MHCs to have charges for assault offenses. Salt Lake clients were not different from other jurisdictions on most offenses except privacy, where almost a quarter of all Salt Lake clients were charged with the offense. Clients in Cache were least likely to have been charged with felony offenses, but were more likely than clients of other MHCs to have been charged with DUI offenses. Both Utah and Weber had high rates of theft offenses; fraud offenses were most notable in the Utah MHC.

Table 1: Pre-MHC Three-Year Criminal Offense History by MHC (% Charged)

Offense Category	Cache	Salt Lake	Utah	Weber
Any Charge	88.9	91.9	80.4	91.3
Any Felony	61.1	74.8	70.6	87.0
Controlled Substance	44.4	46.5	41.2	60.9
Theft	22.2	36.8	52.9	47.8
Assault	27.8	35.7	21.6	39.1
Privacy	5.6	24.0	9.8	13.0
Obstruction	0.0	23.3	27.5	13.0
Fraud	5.6	14.3	25.5	8.7
Damaged Property	16.7	13.6	13.7	13.0
Public Order or Public Peace	5.6	13.6	9.8	13.0
Liquor	5.6	13.2	11.8	21.7
Burglary	5.6	10.5	11.8	4.3
Stolen Property	0.0	8.5	11.8	0.0
Forgery	11.1	7.8	17.6	8.7
Traffic	0.0	5.8	13.7	8.7
DUI	11.1	1.9	7.8	4.3

Sample sizes are: 18 in Cache, 258 in Salt Lake, 51 in Utah, and 23 in Weber
The following offense categories are excluded from the table due to a low rate of occurrence (less than 10% in any one MHC: Robbery, Weapons, Escape, Prostitution, Stolen Vehicles, Arson, Family, Morals, Sex Offenses, Kidnapping, Homicide, Pornography, Threats, Smuggling.

In addition to criminal offense histories, any MHC clients who were under supervision of Adult Probation & Parole (AP&P) during the study window had an assessment of recidivism risk (these data were provided by UDC). Though AP&P changed assessment tools in recent years[5], given the historic timeframe of the study, all MHC clients with an assessment prior to MHC participation were given the Level of Service Inventory – Revised (LSI-R).

[4] It is possible to have no charges in the previous three years, but still have a criminal sentence, as certain charges and/or cases can take a considerable amount of time to reach a disposition in the court system.
[5] AP&P has switched from the Level of Service Inventory – Revised (LSI-R) to an updated version, the Level of Service – Risk, Need, Responsivity (LS/RNR)

The LSI-R is a third-generation, theory driven risk and need assessment tool (Andrews & Bonta, 1995) for the criminal offender population. It contains 54 items and 10 subscales: Criminal History, Education and Employment, Financial, Family/Marital, Accommodations, Leisure and Recreation, Companions, Alcohol and Drugs, Emotional and Personal, and Procriminal Attitude Orientation. Items and subscales were derived from a literature review of the factors theoretically related to criminal conduct. The utility of the instrument is supported by three decades of research support (Andrews, 1982; Andrews, Bonta & Wormith, 2006; Gendreau, Goggin & Smith, 2002). Scores on the assessment can range from 0 to 53[6], where higher scores indicate greater risk.

Table 2 shows the number and percentage of MHC clients in each jurisdiction who had an LSI-R assessment within the three-year period prior to MHC participation. These values varied considerably across the courts, with Cache having the fewest LSI-R assessments and Weber having the most. Because AP&P supervision is a more intensive form of supervision, the percentage of cases reaching that level of supervision within a court is partly an indicator of the severity of its clients' criminal histories (not to be confused with severity of risk and need as provided by the LSI-R score).

The table also shows the average risk score and standard deviation (SD) for each court. Across the courts, the highest risk scores were observed in Weber, while the lowest were observed in Utah. The fact that Weber had the highest LSI-R scores is, to some extent, expected given the results of Table 1 that showed Weber had the highest percentage of clients who committed a felony offense.

Table 2: Count and Percentage of Clients with an LSI-R Assessment by MHC; Average LSI-R Score

MHC	Count	Percentage	Average LSI-R Score (SD)
Cache	5	27.8	26.0 (8.6)
Salt Lake	98	38.0	27.7 (8.6)
Utah	23	45.1	23.3 (8.2)
Weber	15	65.2	29.1 (7.1)

Together, Tables 1 and 2 underscore the fact that the MHCs served clients with different criminal histories as well as different levels of risk and need. Using the LSI-R score (and now the LS/RNR score), individuals are categorized into risk levels, which determines the level of supervision they receive. The average score in the Utah MHC would have been categorized as "moderate risk", while average scores in the other regions would have been categorized as "high risk". Owing to these differences in both criminal histories and risk levels, attempts to match clients using propensity score matching (discussed below) were performed within each MHC.

[6] Although the LSI-R has 54 items, two specific items cannot occur together, so the maximum score is 53 rather than 54.

Treatment and Diagnosis History

The next series of tables provides detail on the treatment services and mental health diagnoses of MHC clients' pre-MHC participation. Unlike results for offense histories, treatment services and mental health diagnoses are presented separately for each MHC to emphasize that results should not be compared across the courts. Whereas MHC clients were equally likely to be found in BCI data regardless of their MHC (and cases rising to the level of AP&P supervision had LSI-R scores), this was not the case for DHS data. There were notable discrepancies in the likelihood of being found in DHS data by MHC.

As seen in Table 3, MHC clients in Cache were found in DHS records pre-MHC for only 38.9% of cases. In stark contrast, the vast majority of clients in Utah were found in DHS records. Clients in Salt Lake and Weber were found in DHS records for about half the cases.

Table 3: Count and Percentage of Clients Found in DHS Data by MHC

MHC	Count	Percentage
Cache	7	38.9
Salt Lake	138	53.5
Utah	46	90.2
Weber	12	52.2

Diagnosis categories provided in the tables that follow are from the International Classification of Diseases, Tenth Revision (ICD-10; World Health Organization, 1992) [7]. While not all ICD-10 categories are represented, all of those that occurred in DHS records for MHC clients are. A complete list of codes and categories can be found here: http://www.icd10data.com/ICD10CM/Codes[8]. Fourteen diagnostic categories were reported in the current sample. The categories can be quite broad; for that reason, some exemplars of the diagnoses are provided in Appendix I to aid interpretation of the tables that follow.

Tables 4-11 provide a summary of both diagnoses and services for each MHC. Because one should not compare across MHCs, diagnoses are sorted in descending order of prevalence within each MHC; thus, the ordering of categories is not the same across MHCs (tables).

Notably, the lack of a diagnosis or service does not indicate a person does not have the diagnosis or did not receive a given service. For example, for the Cache MHC, seven of the 18 cases within the court have a diagnosis of "Mental and behavioral disorders due to psychoactive substance use," but only seven cases from Cache were found in DHS data. Some service providers do not report to DHS, and, as seen in Table 3, that lack of reporting differs across jurisdictions. For that reason, the percentage column, labeled "Percentage within DHS", represents the percentage of cases

[7] During the timeframe under study, DHS switched from use of the Diagnostic and Statistical Manual (DSM-IV) to the International Classification of Diseases (ICD-10) system. DSM-IV diagnoses were converted to ICD-10 diagnoses using a UCJC-developed crosswalk table.
[8] The majority of ICD-10 codes are related to physiological conditions; psychological conditions are a small subset of ICD-10 codes.

within an MHC that were also found in DHS records and have a diagnosis or service; the column "Percentage within MHC" represents the percentage of cases with a diagnosis or service out of all cases in the MHC. Lack of a diagnosis or service is, thus, a confounded indicator, indicating a person may not have a disorder or service at all, or simply did not have one that was reported to DHS.

As seen in Table 4, for Cache, the most common diagnosis was "Mental and behavioral disorders due to psychoactive substance use". The next most common category in Cache involved "Mood [affective] disorders". It is important to keep in mind that the low rate at which Cache cases were found in DHS data could be obscuring mental health conditions. To the extent that mental health providers in Cache do not report to DHS, an accurate representation of the mental health of Cache's MHC clients cannot be obtained.

Table 4: Pre-MHC Diagnoses for Cache MHC

Diagnosis	Count	% within DHS	% within MHC
Mental and behavioral disorders due to psychoactive substance use	7	100.0	38.9
Mood [affective] disorders	5	71.4	27.8
Schizophrenia, schizotypal, delusional, and other non-mood psychotic disorders	3	42.9	16.7
Disorders of adult personality and behavior	3	42.9	16.7
Persons encountering health services for examinations	3	42.9	16.7
Anxiety, dissociative, stress-related, somatoform and other nonpsychotic mental disorders	2	28.6	11.1
Behavioral syndromes associated with physiological disturbances and physical factors	1	14.3	5.6
Behavioral and emotional disorders with onset usually occurring in childhood and adolescence	1	14.3	5.6
General symptoms and signs	1	14.3	5.6
Mental disorders due to known physiological conditions	0	0.0	0.0
Intellectual disabilities	0	0.0	0.0
Pervasive and specific developmental disorders	0	0.0	0.0
Persons with potential health hazards related to socioeconomic and psychosocial circumstances	0	0.0	0.0
Persons encountering health services in other circumstances	0	0.0	0.0
Other and unspecified effects of external causes	0	0.0	0.0

Sample size 18 in MHC, 7 found in DHS

Table 5 provides details for some of the mental health services clients of the Cache MHC received prior to MHC participation. Outpatient services were most common, followed by some form of mental health treatment. It is important to note that outpatient services are not necessarily treatment; accordingly, treatment does not subsume outpatient services. Outpatient services may be diagnostic rather than therapeutic. None of the clients received inpatient or residential services pre-MHC.

Table 5: Pre-MHC Services for Cache MHC

Service	Count	% within DHS	% within MHC
Outpatient Service	7	100.0	38.9
Prior Mental Health Treatment	4	57.1	22.2
Any State Hospitalization - No Limiting Date [a]	1	14.3	5.6
Inpatient Service	0	0.0	0
Residential Service	0	0.0	0

Sample size 18 in MHC, 7 found in DHS

[a] The variable representing Any State Hospitalization is not time specific. DHS is not able to attach a date to the variable; the hospitalization could have occurred pre- or post-MHC.

As seen in Table 6, for Salt Lake, the most common diagnosis was “Mental and behavioral disorders due to psychoactive substance use”. Several other categories were similar in terms of prevalence, including psychotic disorders, mood [affective] disorders, general symptoms and signs, persons encountering health services for examinations, and disorders of adult personality and behavior. The Salt Lake MHC appears to serve a diverse population with the presence of multimorbidities.

Table 6: Pre-MHC Diagnoses for Salt Lake MHC

Diagnosis	Count	% within DHS	% within MHC
Mental and behavioral disorders due to psychoactive substance use	93	67.4	36.0
Schizophrenia, schizotypal, delusional, and other non-mood psychotic disorders	66	47.8	25.6
Mood [affective] disorders	66	47.8	25.6
General symptoms and signs	63	45.7	24.4
Persons encountering health services for examinations	61	44.2	23.6
Disorders of adult personality and behavior	60	43.5	23.3
Anxiety, dissociative, stress-related, somatoform and other nonpsychotic mental disorders	41	29.7	15.9
Behavioral and emotional disorders with onset usually occurring in childhood and adolescence	21	15.2	8.1
Persons with potential health hazards related to socioeconomic and psychosocial circumstances	12	8.7	4.7
Persons encountering health services in other circumstances	10	7.2	3.9
Behavioral syndromes associated with physiological disturbances and physical factors	8	5.8	3.1
Mental disorders due to known physiological conditions	7	5.1	2.7
Intellectual disabilities	5	3.6	1.9
Pervasive and specific developmental disorders	4	2.9	1.6
Other and unspecified effects of external causes	1	0.7	0.4

Sample size 258 in MHC, 138 found in DHS

Table 7 provides details for some of the mental health services clients of the Salt Lake MHC received prior to MHC participation. Outpatient services were most common, followed by some form of mental health treatment. None of the clients received residential services pre-MHC.

Table 7: Pre-MHC Services for Salt Lake MHC

Service	Count	% within DHS	% within MHC
Outpatient Service	111	80.4	43.0
Prior Mental Health Treatment	91	65.9	35.3
Inpatient Service	25	18.1	9.7
Any State Hospitalization - No Limiting Date	15	10.9	5.8
Residential Service	0	0.0	0.0

Sample size 258 in MHC, 138 found in DHS

[a] The variable representing Any State Hospitalization is not time specific. DHS is not able to attach a date to the variable; the hospitalization could have occurred pre- or post-MHC.

The Utah MHC clients were by far the most likely to be found in DHS data. As a consequence, relatively more can be ascertained regarding the type of clients the Utah MHC serves. Table 8 shows that, for the Utah MHC, the most common diagnosis was "Mood [affective] disorders", followed by "Anxiety, dissociative, stress-related, somatoform and other nonpsychotic mental disorders".

Table 8: Pre-MHC Diagnoses for Utah MHC

Diagnosis	Count	% within DHS	% within MHC
Mood [affective] disorders	42	91.3	82.4
Anxiety, dissociative, stress-related, somatoform and other nonpsychotic mental disorders	38	82.6	74.5
General symptoms and signs	29	63.0	56.9
Mental and behavioral disorders due to psychoactive substance use	25	54.3	49.0
Persons encountering health services for examinations	23	50.0	45.1
Behavioral and emotional disorders with onset usually occurring in childhood and adolescence	13	28.3	25.5
Disorders of adult personality and behavior	11	23.9	21.6
Schizophrenia, schizotypal, delusional, and other non-mood psychotic disorders	6	13.0	11.8
Persons with potential health hazards related to socioeconomic and psychosocial circumstances	6	13.0	11.8
Behavioral syndromes associated with physiological disturbances and physical factors	5	10.9	9.8
Other and unspecified effects of external causes	5	10.9	9.8
Mental disorders due to known physiological conditions	0	0.0	0.0
Intellectual disabilities	0	0.0	0.0
Pervasive and specific developmental disorders	0	0.0	0.0
Persons encountering health services in other circumstances	0	0.0	0.0

Sample size 51 in MHC, 46 found in DHS

Table 9 provides details for some of the mental health services clients of the Utah MHC received prior to MHC participation. Outpatient services were most common, followed by some form of mental health treatment. None of the clients received residential services pre-MHC or had state hospitalizations.

Table 9: Pre-MHC Services for Utah MHC

Service	Count	% within DHS	% within MHC
Outpatient Service	45	97.8	88.2
Prior Mental Health Treatment	35	76.1	68.6
Inpatient Service	3	6.5	5.9
Residential Service	0	0.0	0.0
Any State Hospitalization - No Limiting Date	0	0.0	0.0

Sample size 51 in MHC, 46 found in DHS

[a] The variable representing Any State Hospitalization is not time specific. DHS is not able to attach a date to the variable; the hospitalization could have occurred pre- or post-MHC.

Table 10 shows that, for the Weber MHC, the most common diagnosis was "Anxiety, dissociative, stress-related, somatoform and other nonpsychotic mental disorders", followed by "Persons encountering health services for examinations". It is important to keep in mind that the low rate at which Weber cases were found in DHS data could be obscuring mental health conditions. To the extent that mental health providers in Weber do not report to DHS, an accurate representation of the mental health of Weber's MHC clients cannot be obtained.

Table 10: Pre-MHC Diagnoses for Weber MHC

Diagnosis	Count	% within DHS	% within MHC
Anxiety, dissociative, stress-related, somatoform and other nonpsychotic mental disorders	10	83.3	43.5
Persons encountering health services for examinations	9	75.0	39.1
Mental and behavioral disorders due to psychoactive substance use	8	66.7	34.8
General symptoms and signs	8	66.7	34.8
Mood [affective] disorders	7	58.3	30.4
Schizophrenia, schizotypal, delusional, and other non-mood psychotic disorders	6	50.0	26.1
Disorders of adult personality and behavior	6	50.0	26.1
Behavioral and emotional disorders with onset usually occurring in childhood and adolescence	4	33.3	17.4
Persons with potential health hazards related to socioeconomic and psychosocial circumstances	3	25.0	13.0
Mental disorders due to known physiological conditions	1	8.3	4.3
Behavioral syndromes associated with physiological disturbances and physical factors	1	8.3	4.3
Intellectual disabilities	1	8.3	4.3
Pervasive and specific developmental disorders	1	8.3	4.3
Other and unspecified effects of external causes	1	8.3	4.3
Persons encountering health services in other circumstances	0	0.0	0.0

Sample size 23 in MHC, 12 found in DHS

Table 11 provides details for some of the mental health services clients of the Weber MHC received prior to MHC participation. Outpatient services and some form of mental health treatment were most common. None of the clients received residential services pre-MHC and state hospitalizations were rare.

Table 11: Pre-MHC Services for Weber MHC

Service	Count	% within DHS	% within MHC
Outpatient Service	12	100.0	52.2
Prior Mental Health Treatment	12	100.0	52.2
Inpatient Service	4	33.3	17.4
Any State Hospitalization - No Limiting Date	1	8.3	4.3
Residential Service	0	0.0	0.0

Sample size 23 in MHC, 12 found in DHS

[a] The variable representing Any State Hospitalization is not time specific. DHS is not able to attach a date to the variable; the hospitalization could have occurred pre- or post-MHC.

While cases in the Utah MHC were likely to be found in DHS data, the same was not true of comparison cases selected within the Utah MHC jurisdiction. Specifically, while 88.2% (45 of 51) of MHC cases were found in DHS data, only 60.7% (142 of 234) of potential comparison cases were found. While this might still have been sufficient to match, cases were dissimilar in other ways within the Utah MHC that made adequate matching untenable given the variables available. These issues are discussed next.

The fact that most MHC cases did not have DHS data despite being in MHC was surprising. It was also challenging from a modeling perspective because one cannot model the mental health factors that might play a role in predicting mental health court participation if they are unknown/unreported. The section on propensity score matching outlines some of the problems caused by this lack of data.

Propensity Score Matching

Given differences between MHCs outlined above, PSM was performed separately for each court using comparison cases within the same jurisdiction. This matching strategy is designed to create a sample of individuals not included in MHCs who can be compared to MHC participants based on a similar likelihood of participation in the MHC program. Criminal and treatment histories prior to MHC were established by using the MHC agreement date for MHC cases and the sentence date for comparison cases as a hinge date separating the pre- and post- periods. The following variables, derived from past research (Anestis & Carbonell, 2014; Fiduccia & Rogers, 2012; Luskin, 2012; McNiel & Binder, 2007; Steadman et al., 2014) were considered in the PSM process[9] (criminal outcomes are from BCI, while mental health outcomes, and the homelessness flag, are from DHS):

- Number of drug charges in the 3-year period pre-MHC (agreement date, or sentence date if comparison case)
- Number of violent offense charges in the 3-year period pre-MHC
- Number of felony charges in the 3-year period pre-MHC
- Number of total charges in the 3-year period pre-MHC

[9] Owing to statistical considerations including, appropriateness for the MHC, model fit, and sample size, not all variables were considered in a model for each MHC.

- Whether the person had a *diagnosis* for a substance abuse condition in the pre-period[10]. Substance abuse disorders fall under ICD-10 codes starting with "F1".
- Whether the person had a diagnosis for a severe mental illness in the pre-period. The codes for these disorders fall under ICD-10 codes starting with "F2" (Schizophrenia, schizotypal, delusional, and other non-mood psychotic disorders) and "F6" (Disorders of adult personality and behavior).
- Whether the person had a diagnosis for an adjustment disorder in the pre-period. Adjustment disorders fall under ICD-10 codes starting with "F3", "F4", and "F9"; these include, as examples, anxiety, OCD, severe stress, ADHD, conduct disorders, etc.
- Outcome Questionnaire (OQ-45; explained below) pre-score: This variable was calculated as the average (or sum if only one) OQ-45 score from the pre-period.
- Whether the person had some level of substance abuse *service* in the pre-period.
- Whether the person had any "successful" treatment in DHS during the pre-period.
- Number of prior residential treatments in DHS during the pre-period.
- Number of prior inpatient treatments in DHS during the pre-period.
- Number prior outpatient treatments in DHS during the pre-period.
- Whether the person was white or a minority
- Whether the person was male or female
- Whether the person was ever homeless in the pre-period.
- A DHS missing value indicator identified cases that did and did not have a record in DHS during the pre-period.

The following section outlines the PSM process and explains why the process was not deemed a success. Some of the nonessential details of the process are omitted to avoid obfuscating the most important issues. While the process was not successful, details about the process are likely to benefit future research on Utah's MHCs and are provided for that reason.

PSM Background

PSM, as a tool, derives from the counterfactual framework of Neyman and Rubin (Rosenbaum & Rubin, 1983; Shadish et al., 2002). One can think of a counterfactual as what might have happened under different circumstances. For an MHC participant, the counterfactual is the outcome he or she might have experienced under hypothetical participation in the control condition. The counterfactual for the control participants is the inverse of that just described for treatment participants (Guo & Fraser, 2009).

Essentially, PSM provides a method for addressing a missing outcome. For MHC participants, the missing value is the outcome had they not been in MHC, while, for comparison cases, the missing value is the outcome had they been in MHC. Although this logic is perhaps confusing at first, it is relatively easy to imagine a scenario where a person who was, in fact, assigned to MHC was instead not assigned to MHC by some circumstance. If one can identify a group of individuals who were highly likely to have been assigned to MHC, but who were not assigned owing to, for

[10] For DHS-derived mental health diagnoses and services, only half of all cases had DHS records. To avoid further reducing the number of cases with DHS histories, no time-limit was placed on the window in which a diagnosis or service could occur as long as it was before the hinge date.

example, failure to recognize their need, MHC capacity issues, etc., then one has identified a potential counterfactual group to provide the unobserved outcome mentioned above.

PSM works by first identifying the characteristics that make a person likely to be in treatment (MHC) and then reducing all of those characteristics into a single score representing their probability of being in treatment (Leite, 2016). That is, several factors may influence the likelihood of being in MHC, but all of those factors can be reduced to a single score, the actual probability from 0-100%. From the discussion above, one can see there are circumstances (e.g., MHC was full) that can lead a person with a high probability of being in MHC to not actually be in the MHC condition.

Once one has obtained the individual propensities calculated among treated and non-treated alike, the scores are then matched such that MHC cases with a given propensity are paired with non-MHC cases with a similar or even identical propensity. Using this method, the paired participants are as alike as possible on the factors that lead to a given probability of being in MHC. In theory, PSM removes the selection bias mechanism inherent in non-randomized designs and makes the groups comparable (Leite, 2016).

A couple of limitations of the PSM method clearly follow from the above discussion. First, PSM is extremely sensitive to the identification of the correct covariates. Clearly, PSM cannot correct for variables it is not aware of, that should have been in the model, and that impact both the probability of being in MHC and the outcomes that follow. Second, PSM depends on availability of an untreated group. If the counterfactual group to MHC cannot be identified or does not exist, PSM should not be used to make inferences about treatment efficacy.

The first issue, identification of the correct covariates, seems easy in theory but is quite difficult in practice. There is no way to know whether one has modeled every variable that predicts MHC participation and subsequent outcomes; the argument becomes theoretical, as one must convince an expert audience that all or most of the important variables have been captured. Fortunately, researchers can use theory and previous work as a guide, as was done in the current research.

The second issue, actual availability of an adequate untreated group, is, unfortunately, not always knowable until the analyst has modeled the covariates thought to be important, but failed to identify an accurate comparison group. This could be due to the variables chosen to model the propensity score, the cases selected for comparison, the lack of a true counterfactual group, or all of the above.

PSM Modeling

This section explains the process of PSM that was performed for this study, as well as the issues that were encountered that made PSM problematic. Although the process was performed for each MHC, the issues resulting in a failure to identify a comparison sample were universal. For sake of parsimony, rather than present all the models, an exemplar of the problem common to all courts is shown.

Missing Data

This section briefly describes the methods used to address missing data before PSM matching occurred. One of the most widely-accepted, best practices for addressing missing data is multiple imputation (MI; Guo & Fraser, 2009). Multiple imputation is a data-replacement procedure that "fills in" each missing value with a set of plausible values that represent the uncertainty about the "right" value to impute. Simulation research has shown that MI results in valid statistical inferences as long as its assumptions are sufficiently met (Guo & Fraser, 2009).

One requirement of MI is that other available variables should predict the missing variable. In the case of this project, many of the variables outlined above from DHS were missing in the sense that DHS had no records regarding diagnoses or services for 50.4% of all cases. Importantly, this does not indicate the individuals had no diagnoses or services, but, rather, that they simply were not reported to DHS.

The relatively diffuse problem of missing data meant that certain variables, such as race (minority) and sex, could not be accurately imputed because only criminal history variables were available (only 9.5% of cases had no data in BCI for the three years pre- study hinge date). Not surprisingly, criminal history variables were not particularly good predictors of race or sex. For that reason, missing data on these variables was addressed by alternative means.

Race (Minority)

Although data were aggregated across different systems to provide as much complete information as possible, Race was missing for 48.2% (936) of the 1,940 cases. Missingness on this variable was addressed by using orthogonal contrasts. While the mathematics of using orthogonal contrasts is somewhat complicated, fortunately, understanding their purpose is not. Essentially, these contrasts allow one to account for missing values on the "minority" variable, but in a way that makes statistical models more stable.

Sex

Another variable with missing values was client sex, missing for 25.4 percent of cases. R software has a package called 'gender' (Blevins & Mullen, 2015) that can be used to complete missing values for sex on a probability basis. The package's authors devised a rather ingenious use of Social Security Administration (SSA) data. Using 91,320 names collected from SSA data from 1880 to 2012, one can assign a probability, within a specific year, that a name was either male or female. For the purposes of this project, the package was used to "fill in" sex when it was missing, but only if the value to be filled was associated with one sex at 99% probability or higher.

Because some names are quite unique, and because the criterion for matching (99% probability) was quite strict, the above method left 6.3% of values for sex still missing. These were addressed using a missing value indicator variable as described for "minority" above.

OQ-45 Score

The OQ-45 is a measure of psychological functioning used to evaluate change over the course of treatment (Beckstead et al., 2003). It measures mental health and overall functioning using three subscales: symptom distress (depression and anxiety); interpersonal relations (loneliness, conflict with others, and marriage and family difficulties); and social roles (difficulties in the workplace, school or home duties). It takes only a few minutes to administer and can be given either pre- and post-intervention or after each treatment session. The total score ranges from 0 to 180.

Missing values on the OQ-45 were addressed in a different manner than missing values for other variables. Data were first divided into cases with and without DHS records and then multiple imputation was used to complete OQ-45 scores for those cases with DHS variables available to use in prediction. Ten imputed datasets were created for the missing OQ-45 scores. Subsequent PSM analysis made use of these imputed values as well as all other variables previously described. For cases with no DHS records, OQ-45 scores were not imputed. Instead, a missing value indicator was included, indicating whether a case had an OQ-45 (imputed or original) or not.

Predictor Problems and Transformations

A few additional methodological concerns are worth noting as they are likely to be of use to future research. First, the variable representing the number of residential services was dropped because it had too little variance to be useful in prediction; only one person was flagged as having residential services in DHS data. The two other DHS count variables (number of inpatient and number of outpatient services) were highly over-dispersed, which means that some people had no such services and some people had an extremely high number. Variables of this nature are often called “ill-scaled”, and they can cause problems in model estimation.

To examine whether they would be problematic in the PSM models, the two DHS count variables were entered individually as predictors of Case Type (that is, MHC or non-MHC) in separate binary logistic regression models (the same type of analysis that PSM would later use). As anticipated, they both produced highly unstable estimates and impossible predicted probabilities. Less technically, at the extremes, a person can have a predicted probability of being in MHC of either 0 or 1. In no case can a probability be less than 0 or greater than 1. Yet, with these problematic count variables, predicted probabilities exceeding one were found. To address the issue, the variables were log transformed and the relationships were modeled again. The transformed versions provided stable estimates as well as predicted probabilities within the allowable range.

Limiting Cases Based on Data Availability

Before the PSM process could begin, the issue concerning a lack of DHS data needed to be addressed. For cases without DHS data, matching would have been based purely on criminal history. Proceeding with matching based only on criminal history is tantamount to inferring that mental health factors are not relevant to how people arrived in MHC. This would likely be unreasonable because MHC clients arrive in MHC due to the special mental health needs of the population. Because the absence of a DHS record was not regarded as evidence that no mental health diagnoses or services existed (instead, it was posited that MHC clients and many of the comparisons likely had received diagnoses and services not reported to DHS), the absence of mental health records needed to be addressed.

To provide a more appropriate basis for matching, cases without DHS histories were removed from the matching process. This analytic decision alone would require justification if the PSM process had worked, as one would expect clients with DHS services might differ from clients receiving services from agencies not reporting to DHS. This might, for example, indicate a greater use of private mental health services relative to public ones, from which one may infer differences in the two populations' respective socioeconomic statuses. Because the matching did not work, this issue can be ignored, but the public/private distinction would need to be considered in future research if private records could be obtained.

PSM Methods

Cases were next matched using several matching algorithms. Best practice in PSM matching (Leite, 2016) recommends using several methods, as methods will vary in their degree of matching success depending on features of the data. That is, there is no one universally best matching method. In the present study, the following algorithms were implemented: nearest neighbor; optimal matching; coarsened exact matching; full matching; and genetic matching. Descriptions of each of these algorithms are available in Appendix II.

Matches were created using multiple versions of each of the above algorithms. For example, nearest neighbor matching was implemented with different caliper sizes and different ratios of treatment to comparison cases. In total, 9 matching algorithms were implemented within each of the MHCs, for a total of 36 matching attempts.

PSM Diagnostics

Recall that, in PSM, one seeks to identify a group of individuals who were highly likely to have been assigned to MHC (i.e., treatment), but who were not assigned to MHC for some reason. Here, "likely" is defined by a propensity score, where a value closer to 1 indicates a high probability of being in MHC. Fortunately, we can create a visual of this probability to aid in interpretation. Figure 1 below provides a visualization (a histogram) of what is known as "The Area of Common Support." Data in this figure are hypothetical and represent a relatively ideal case that can (later) be compared to what was found in the present study; therefore the graph in Figure 1 is merely an exemplar. In reality, the obtained common support varies depending on how propensity scores were calculated and the method of matching.

The x-axis (horizontal) represents the probability of being an MHC case based on historic variables, such as mental health and criminal history. The axis ranges from 0 (no chance conditional on modeled variables) to 1 (100% chance conditional on modeled variables). The y-axis is a count of cases, indicating how many cases have a particular propensity for being in MHC. Because this is an example, the actual counts on the y-axis have been removed. Transparency has been added to the model's coloring so the reader can see the MHC cases "through" the non-MHC, comparison cases: grey in the figures indicates where the comparison cases and MHC cases overlap. All of the MHC cases in Figure 1 are grey; none are the pink shade indicated in the key because all MHC cases overlapped with the comparison cases.

A few features of an ideal case, as represented in Figure 1, are worth noting. First, the hypothetical comparison cases cover the full range of hypothetical MHC cases. In this ideal model, there are no instances, of either low or high probability, where hypothetical MHC cases are not overlapped by hypothetical comparison cases. Second, the number of comparison cases outnumbers the number of hypothetical MHC cases. While this is not a requirement, it does typically provide better overall matching success in terms of diagnostics. Third, the hypothetical MHC cases tend to group toward the high probability, which can serve as an indicator that variables that predict being in MHC have, indeed, been identified.

Figure 1: Hypothetical/More Ideal Area of Common Support

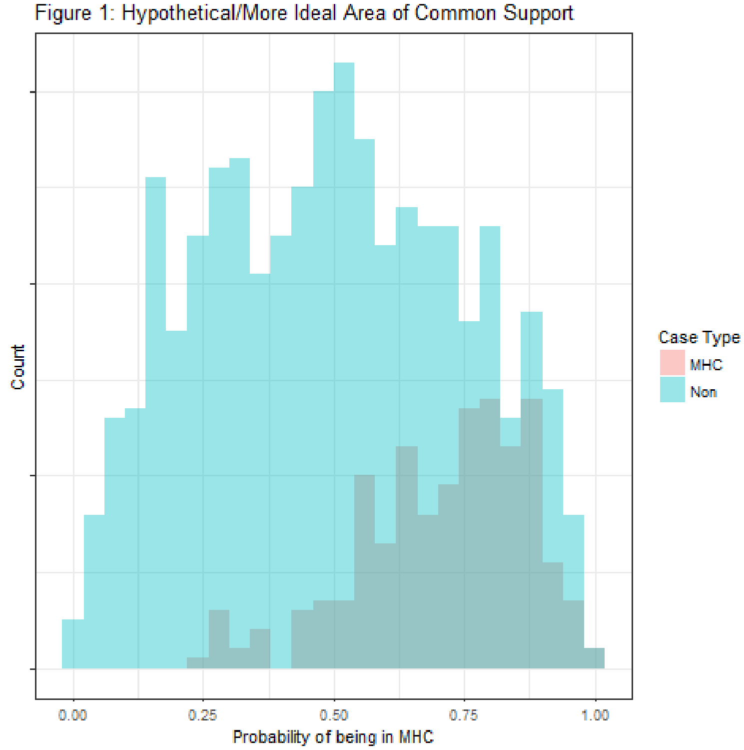


Figure 1 can be contrasted with Figure 2. Figure 2 is an actual example from the present MHC study. Here, one can see notable departures from the pattern found for Figure 1. While the ideal case showed overlap throughout all of the MHC cases, Figure 2 shows that, in the actual data, comparison cases do not cover the full range of MHC cases. Comparison cases in Figure 2 are positively skewed, meaning they are bunched up on the left-hand side, over low propensities of being in MHC. Also, MHC cases are fairly evenly dispersed across the range of probabilities, meaning several MHC cases have a low probability of being in MHC. While this can serve as an indicator that variables that identify MHC cases have not been identified, it can also indicate MHC cases lack defining characteristics.

Figure 2: Example Area of Common Support from MHC

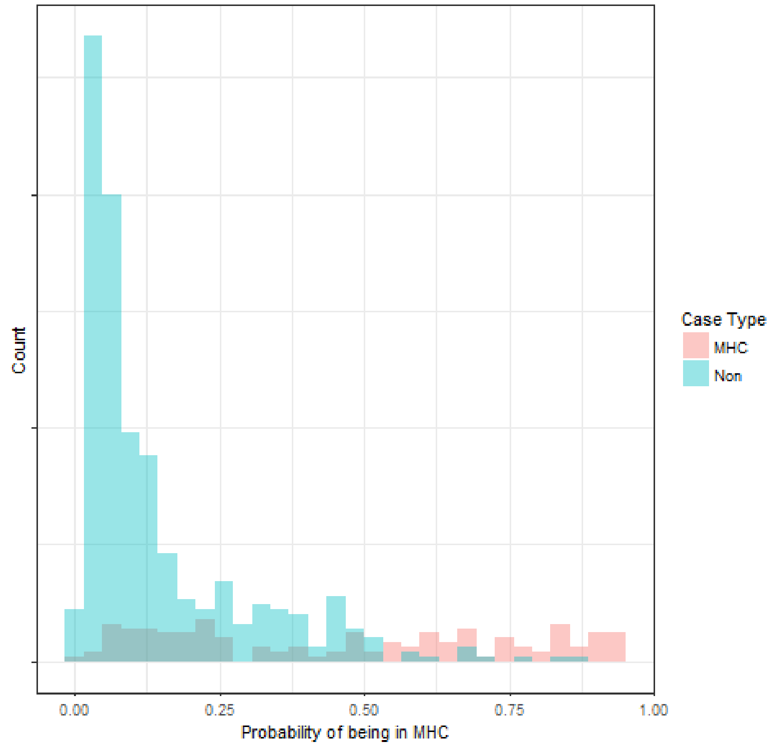


Figure 2 provided one indication that the PSM matching process would have a difficult time achieving balance, but whether or not balance was obtained depends on the algorithm as well as features of the data. That is, each matching algorithm has to be implemented and balance checked for each one before a determination of balance can be made. Several post-matching diagnostics are available in Appendix III. These diagnostics show in more detail the nature of the problems encountered in using PSM with MHC data.

Because of these issues, including those outlined in Appendix III, pursuing an outcome analysis based on matched cases was deemed inappropriate. The goal of the outcome analyses was changed to instead focus on the effect of MHC over time. This analysis is not ideal given that any changes in, for example, the rate of offending over time could be due to MHC or other factors (e.g., jurisdictional changes in enforcement or handling of offenses). Methodological remedies are considered that partially attenuate the limitations created by these issues, but, ultimately, the methodological criticisms cannot be eliminated without a matched comparison design.

Mental Health and Offending Trajectories of MHC Cases over Time

The analyses that follow examine the probability of committing an offense as well as scores on an indicator of mental health – the OQ-45 (as explained above) – over time. Trajectories for both outcomes were created by dividing the three years pre- and post-MHC start into discrete time blocks and modeling the trajectories. Trajectories were modeled using multilevel models for longitudinal data. These models have two desirable attributes required to model the outcomes for MHC cases over time. First, they are able to model the dependencies in the data created by following the same individuals over time. Second, they do not require complete data. In the case of the OQ-45, for example, an MHC client does not have to have an OQ-45 score in all discrete time blocks in order to provide useful information to the model. A detailed explanation of the advantageous properties of multilevel models is beyond the scope of this report, but interested readers should see Raudenbush and Bryk (2002) for further detail.

When modeling events over time, it is important to consider non-linearity in the trajectories. For example, the probability of offending may decline for a period of time, but then may increase again. In that case, a model that allows for a curved trend in offending trajectories is needed. One of the most accurate and flexible methods of modeling non-linear trajectories is semi-parametric regression using smoothing splines (also known as Generalized Additive Models, or GAMs; Wood, 2017). One nice feature of these models is that, when non-linear terms are not needed, the parameters of the model simply reduce to linear; that is, they become identical to typical linear models. However, when non-linear relationships are present, these models are generally more accurate than other models that accommodate non-linearity (e.g., polynomials). The utility of these models can most easily be understood visually. All models below provide a visual interpretation that is complimented by numeric values and numeric tests of significance. Understanding the numeric tests, however, is not required in order to understand the models.

GAMs provide a statistic known as the effective degrees of freedom (EDF). An EDF can be regarded as an indicator of the degree to which the slope of a line changes over time. An EDF of 1.0 means the model can be described by a simple linear effect (i.e., a perfectly or nearly perfectly straight line). Higher values indicate greater deviations from linearity; a value of 2.0, for example, corresponds approximately to a quadratic trend (Wood, 2017); a quadratic trend indicates that a line changes direction by either accelerating or reversing its course (e.g., a trend that was rising begins to rise more quickly, or, in the latter case, begins to decline) [11].

It is important to note that the curves do not simply follow the means of the data over time. If they did, they would overfit the sample and would not be generalizable to other samples or the population. GAMs have a built in penalty term to prevent overfitting of this type. In this sense, they provide accurate explanations of the data while, simultaneously, not being oversensitive to sample-specific anomalies.

OQ-45 Trajectories over Time

Only the MHCs of Salt Lake and Utah had a sufficient number of clients with OQ-45 scores to model changes in the questionnaire over time. For the OQ-45, analyses are limited to those individuals who received the assessment from a provider that reports to DHS. It is possible that the pattern found for changes in OQ-45 over time were different for those receiving DHS-reported

[11] Note that the value of the EDF is not tied to the significance value; instead, the significance value is derived from other components of the model (either the F-test or chi-square value and the reference degrees of freedom for models that follow). For that reason, an EDF of 1.0 can be as or more significant as an EDF of, for example, 5.2. It is best to think of these as independent terms where the EDF indicates the shape of the line (including a linear trend, or no curve, with a value of 1.0) and the significance value indicates whether that trend is meaningful (in a purely significance based sense).

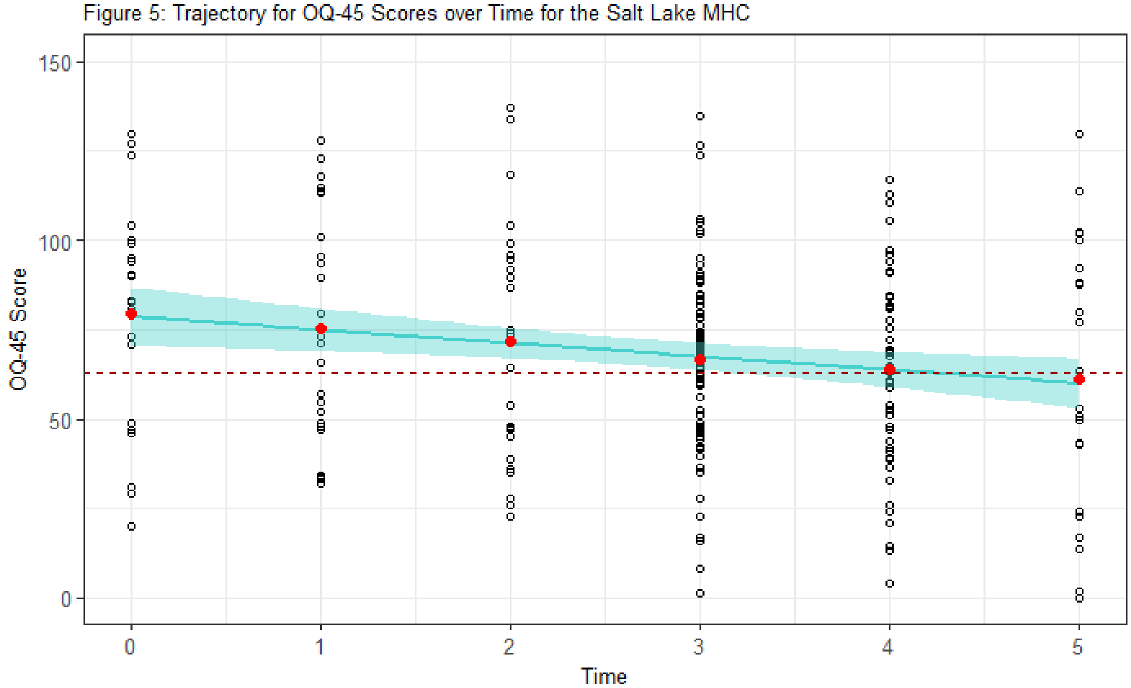
services relative to those who did not. On the OQ-45, lower scores are indicative of better functioning.

Salt Lake OQ-45 Scores

For the Salt Lake MHC, 103 of the 258 clients (39.9%) had at least one OQ-45 score in DHS data. Figure 5 below shows the results of the GAM model fit to the data. The model used discrete time as a predictor of OQ-45 score. The OQ-45 score is provided on the y-axis and time is provided on the x-axis. Times 0, 1, and 2 are pre-MHC, while times 3, 4, and 5 are post-MHC. The open dots represent the individual scores at each time point, while the solid red dots represent the mean score at each time. A dashed line on the figure is provided at the value 63; this is the cutoff on the OQ-45 above which scores are considered of clinical significance.

For these data, the GAM model reduced to a simple linear model; that is, no additional parameters were needed to account for a non-linear or curved relationship. A significant, negative linear trend was found (EDF = 1.00, $F(1.00, 158.21) = 10.69$, $p = .001$), indicating that OQ-45 scores in the Salt Lake MHC declined (i.e., mental health among participants improved) over the six year window. However, the interpretation of the finding is unclear, as the pattern of reduction in OQ-45 scores predates MHC participation. Scores were already declining before clients entered MHC and scores continued to decline at the same rate after MHC participation began. A more interpretable, and perhaps more expected pattern, would have revealed OQ-45 scores were increasing or staying relatively high and stable before MHC, but then began to decrease after MHC participation. Notably, by the final time point, scores had, on average, dropped below the clinical cutoff of 63 (i.e., the point at which a person's symptoms are not considered indicative of a clinical issue).

Figure 5: Trajectory for OQ-45 Scores over Time for the Salt Lake MHC

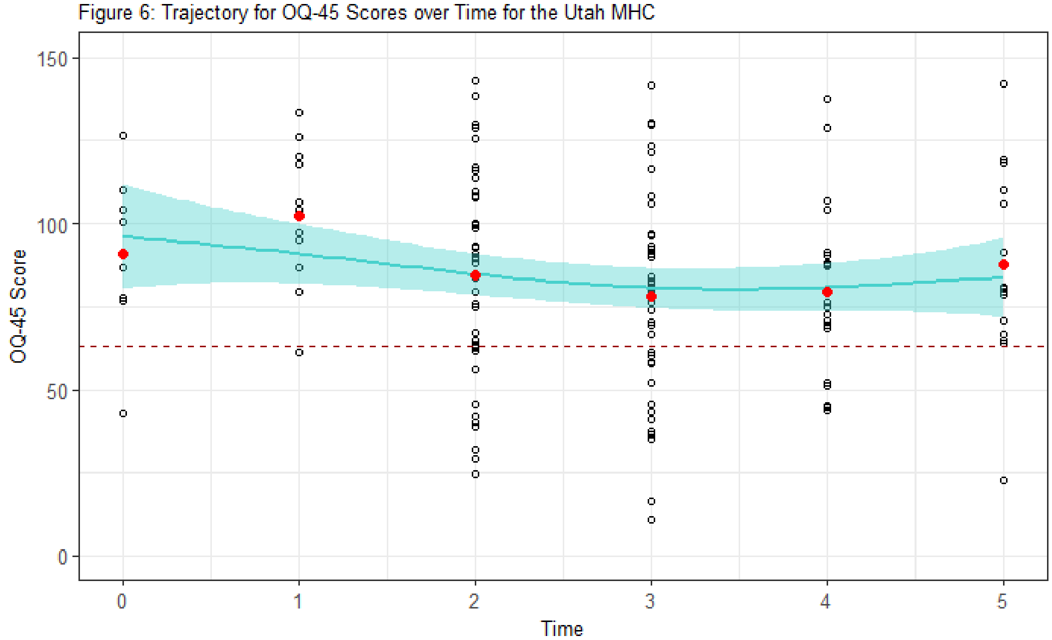


Utah OQ-45 Scores

In the Utah MHC, 45 of 51 clients (88.2%) had at least one OQ-45 score in DHS data. Figure 6 below shows the results of the GAM model fit to the data. Elements of the figure are the same as in Figure 5. For Utah, a curvilinear trend was necessary to model the data. The trend was significant (EDF = 2.29, $F(2.78, 99.12) = 7.29$ $p = .001$). The interpretation of curvilinear relationships in GAMs, or smoothing splines, are visual rather than statistical. Inspecting the figure, one can see that scores on the OQ-45 were trending downward prior to MHC participation and the negative trend continued until two years post-MHC start. At that point, a slight but non-significant increase in scores is observed.

As with Salt Lake, the interpretation of the pattern is unclear given that a pattern of decline in the OQ-45 score predates MHC participation. A more interpretable, and perhaps more expected pattern, would have revealed OQ-45 scores were increasing or staying relatively high and stable before MHC, but then began to decrease after MHC participation.

Figure 6: Trajectory for OQ-45 Scores over Time for the Utah MHC



Offending Trajectories over Time

In the examination of offending trajectories, all MHC cases and each of the four MHCs were considered owing to nearly complete representation of MHC cases in BCI data. The models that follow are similar to those for the OQ-45 above; they model offending (defined as committing any offense) in each of six discrete time periods (3 years pre- and 3 years post-MHC start).

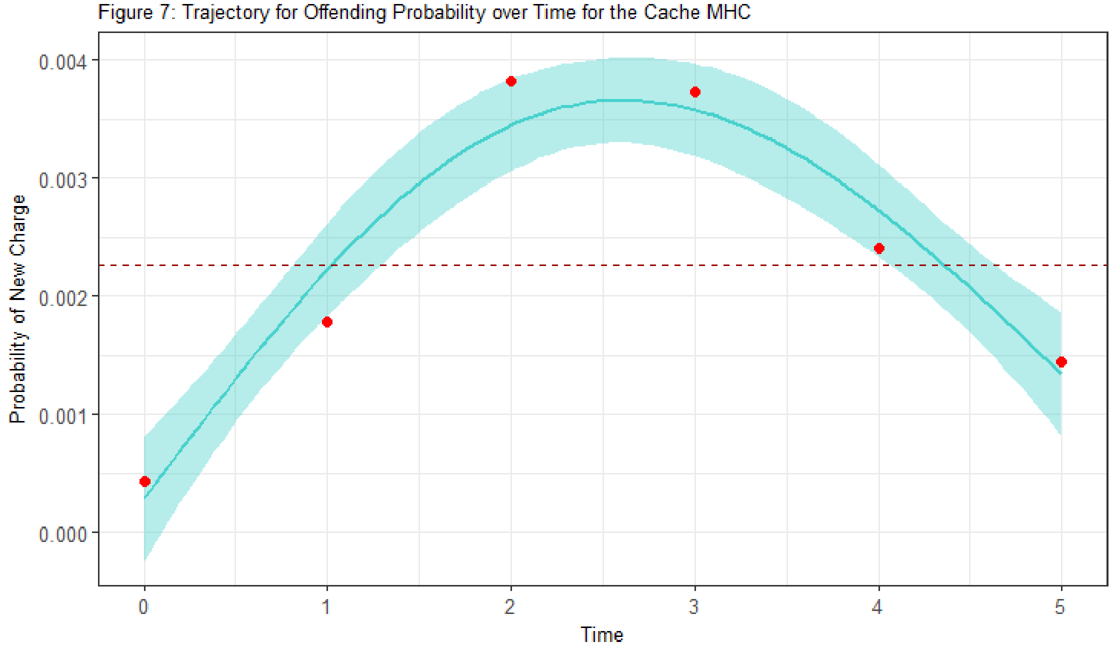
One of the limitations of examining only the within-group (within-MHC) offending trajectories is that it is not possible to fully account for time in the community. For example, perhaps periods in which offending is low are caused by incarceration or institutionalization. If MHC clients are incarcerated or institutionalized, they may not have the opportunity to reoffend. While DHS has an indicator of psychiatric hospitalization, no date is attached; therefore, psychiatric institutionalization could not be accounted for in the models that follow.

Incarceration was accounted for in full in Salt Lake through Salt Lake County Jail and UDC prison records, but only partially in the three other MHCs through prison records. For Salt Lake, the two sources were combined and the model of offending trajectories could account for any form of incarceration, but not psychiatric hospitalization. In the other three courts, UDC prison commitments were considered, but jail days could not be accounted for, as UCJC did not have access to jail data in jurisdictions other than Salt Lake.[12]

Cache Offending Trajectory

Figure 7 below shows the results of the GAM model fit to the Cache data. The model used discrete time as a predictor of the probability of offending. The probability of offending is provided on the y-axis; however, the probability has been rescaled as a function of time in the community. Under the rescaling, one can conceive of the y-axis as indicating the probability of reoffending per day of exposure (i.e., per day in the community). Time is provided on the x-axis. Times 0, 1, and 2 are pre-MHC, while times 3, 4, and 5 are post-MHC. The raw data are not provided as open dots as they were for OQ-45 scores because raw data on offending are simply 0s (did not offend) and 1s (did offend). The solid red dots represent the mean probability of offending, at each time, per unit of exposure. A dashed line on the figure indicates the average probability of offending across all time points.

Figure 7: Trajectory for Offending Probability over Time for the Cache MHC



[12] While jail data could possibly have been obtained in these jurisdictions, recall the original study design involved comparing MHC clients with a propensity matched sample of non-MHC cases. In that methodology, jail days were not needed. After changing the methodology, there was insufficient time to acquire jail data from additional jurisdictions.

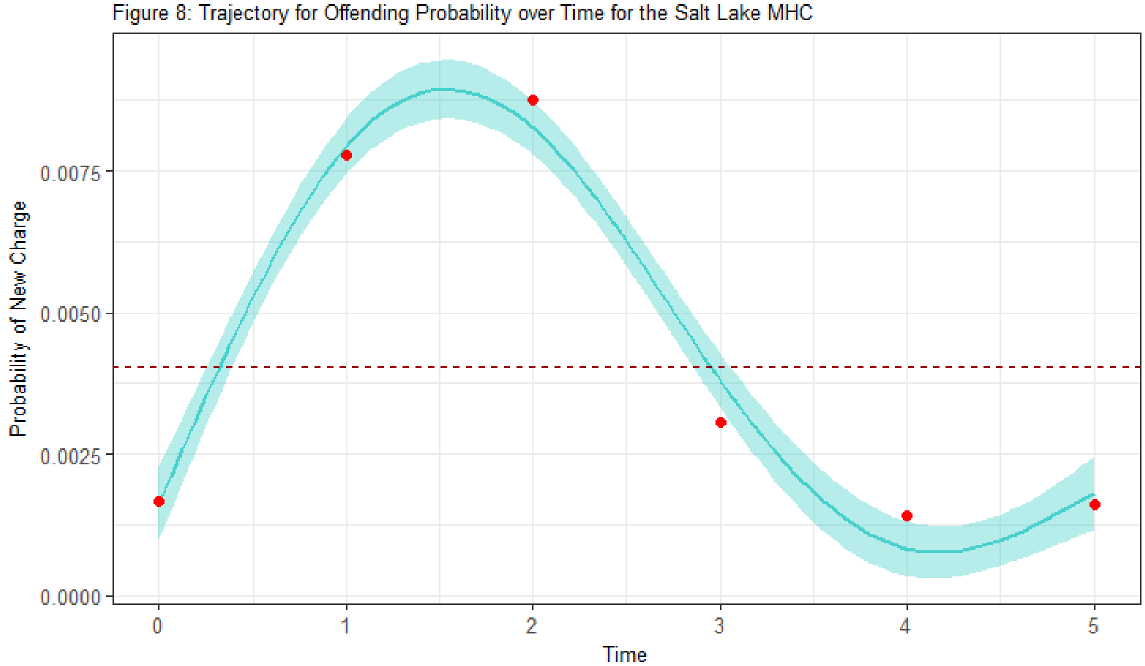
For Cache, a curvilinear trend was necessary to model the data. The trend was significant (EDF = 2.47, $X^2(2.80) = 10.78$, p = .020). Inspecting the figure, one can see that the probability of offending rose sharply until the year marking the start of MHC participation. The trend reverses in the year following starting MHC, and the trend continues downward over the course of the next three years, eventually dropping below the average probability of offending within the Cache MHC sample.

The model for Cache accounts for time in prison, but not time in jail or hospitalization. While the trend appears to suggest that MHC participation may have helped stabilize clients after a period in which they were particularly likely to offend, other explanations are also possible, including: clients were not in the community to reoffend and systemic changes may have taken place in the enforcement of certain offenses. While the latter explanation cannot be ruled out, it is not a likely explanation for the pattern because participants started participation in the Cache MHC between 2009 and 2012. Therefore, the post MHC periods were not identical across participants, and the effect of systemic changes would likely present as noise in the data.

Salt Lake Offending Trajectory

As with Cache, a curvilinear trend was necessary to model the data in Salt Lake. The trend was significant (EDF = 3.00, $X^2(3.00) = 164.5$, p = .000). Figure 8 below shows the results of the GAM model fit to the Salt Lake data. Elements of the figure are the same as in Figure 7. One can see that the probability of offending rose sharply until the year marking the start of MHC participation. The trend reverses in the year following starting MHC, and the trend continues downward over the course of the next two years. At three years post MHC start, the trend levels out relative to two years post start, displaying a slight, but non-significant, upward trend. Notably, the probability of offending for all three years post MHC start is below the average probability of offending within the Salt Lake MHC sample.

Figure 8: Trajectory for Offending Probability over Time for the Salt Lake MHC



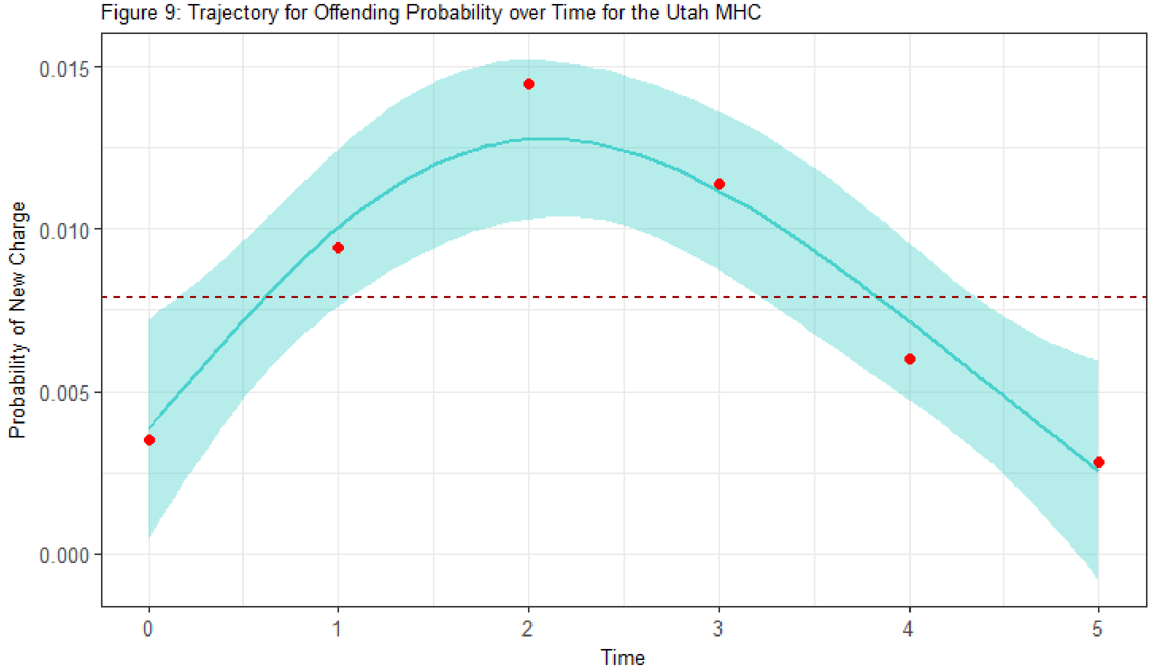
The model for Salt Lake accounts for time in prison and jail, but not hospitalization. While the trend appears to suggest that MHC participation may have helped stabilize clients after a period in which they were particularly likely to offend, other explanations are also possible, including: clients were not in the community to reoffend and systemic changes may have taken place in the enforcement of certain offenses. While the latter explanation cannot be ruled out, it is not a likely explanation for the pattern for the same reasons as outlined for Cache above.

Utah Offending Trajectory

A curvilinear trend was also necessary to model the data in Utah. The trend was significant (EDF = 2.60, $X^2(2.88) = 16.87$, p = .000). Figure 9 below shows the results of the GAM model fit to the Utah data. One can see that the probability of offending rose sharply until the year marking the start of MHC participation. The trend reverses in the year following starting MHC, and the trend continues downward over the course of the next three years, dropping below the average probability of offending within the Utah MHC sample by years two and three post start.

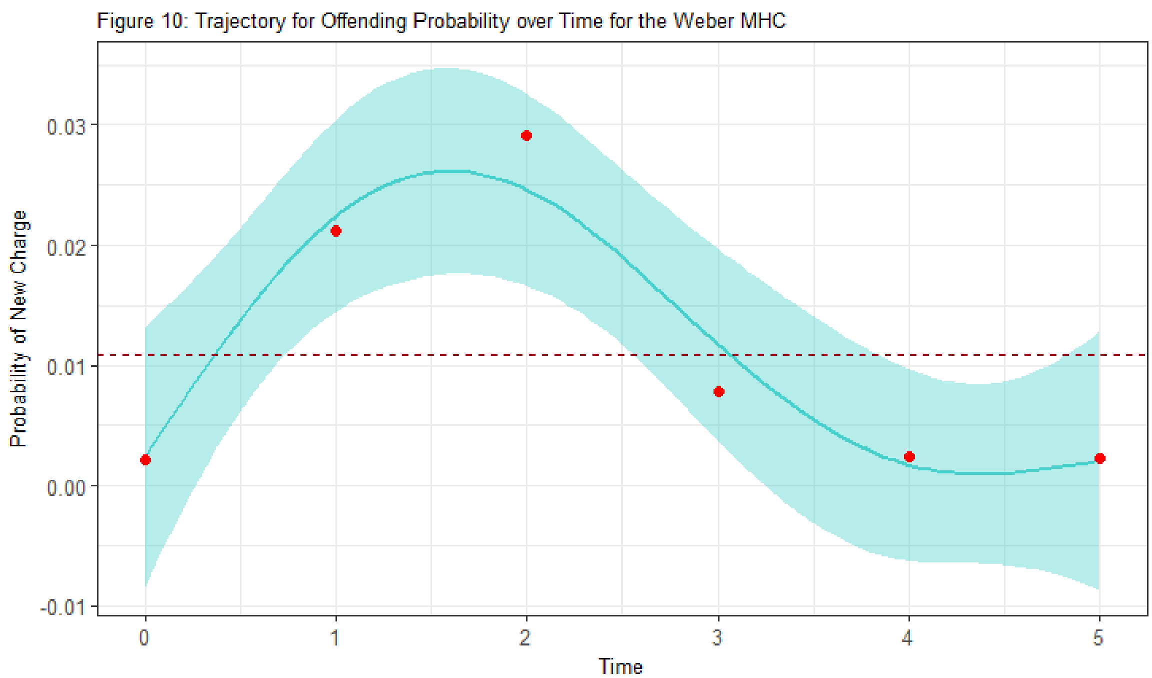
The model for Utah accounts for time in prison, but not jail or hospitalization. While the trend appears to suggest that MHC participation may have helped stabilize clients after a period in which they were particularly likely to offend, other explanations are also possible, including: clients were not in the community to reoffend and systemic changes may have taken place in the enforcement of certain offenses. While the latter explanation cannot be ruled out, it is not a likely explanation for the pattern for the same reasons as outlined above.

Figure 9: Trajectory for Offending Probability over Time for the Utah MHC



Weber Offending Trajectory

A curvilinear trend was necessary to model the data in Weber. The trend was significant (EDF = 3.00, $X^2(3.00) = 21.00$, p = .000). Figure 10 below shows the results of the GAM model fit to the Weber data. One can see that the probability of offending rose sharply until the year marking the start of MHC participation. The trend reverses in the year following starting MHC, and the trend continues downward over the course of the next two years. At three years post MHC start, the trend levels out relative to two years post start. Notably, the probability of offending for all three years post MHC start is below the average probability of offending within the Weber MHC sample.



The model for Weber accounts for time in prison, but not jail or hospitalization. While the trend appears to suggest that MHC participation may have helped stabilize clients after a period in which they were particularly likely to offend, other explanations are also possible, including: clients were not in the community to reoffend and systemic changes may have taken place in the enforcement of certain offenses. While the latter explanation cannot be ruled out, it is not a likely explanation for the pattern for the same reasons as outlined above.

Summary of OQ-45 and Offending Trajectories

Trajectories for OQ-45 scores for both the Salt Lake and Utah MHCs revealed a downward trend in average scores, but that trend began prior to MHC participation and merely continued at the same rate after MHC start. In Salt Lake, the average score three years post start was below the level of clinical significance, but that same pattern was not observed in the Utah MHC. The exact meaning of these findings is unclear. It may be the case that a matched comparison sample is required in order to disambiguate the result.

The analysis of within-group offending trajectories over time did reveal some favorable patterns for MHC participation. In all courts, offending trajectories were significantly reduced after MHC participation began. In Salt Lake, this was true even after accounting for time in jail or prison, while, for the other three courts, it was true after accounting for time in prison. None of the models could account for possible psychiatric hospitalization. While the downward trend in the probability of offending post MHC start is encouraging, it remains possible that a matched comparison group would have revealed the same result; that is, perhaps, for the population, a reduced probability of reoffending is the natural consequence of being caught offending rather than the result of MHC participation itself.

Discussion

Potential Causes of Matching Failures

The failure to identify an appropriate comparison group warrants a discussion of why such a failure likely occurred. As discussed above, and in Appendix III, the failure can be due to a number of reasons, including: 1) incorrect covariates, 2) incorrect or incomplete identification of an appropriate comparison group, or 3) the lack of a true counterfactual group.

While the concern over use of incorrect covariates cannot be completely eliminated, it is unlikely for two reasons. First, the study was designed, and variables selected, after a considerable review of the literature on MHCs and propensity score matching. Though less important, the variables are also face valid, meaning that it makes intuitive sense that MHC cases should be matched to non-MHC cases based on both criminal history and mental health factors. Second, variables were also removed from models for certain MHCs when they caused model stability issues or were not important predictors of being in that specific MHC. Overall, it is difficult to imagine variables related to MHC participation and subsequent outcomes that were not considered.

The second and third explanations are related to one another. If an adequate comparison population does not exist in Utah (explanation three), an appropriate comparison group could never be identified. The explanation that no comparison group exists in Utah seems unlikely given that, clearly, MHCs cannot serve all potential clients, and not all clients in need of services are referred to an MHC. However, in Utah, the comparison population may be a difficult population to identify given the relatively broad definition of client characteristics, what qualifies as mental health treatment within MHCs (at least at the time of this report), and corresponding limitations of behavioral health data in the state.

The second explanation (inability to identify an appropriate comparison group because of incomplete data) is most likely. The definition of treatment is not a singular concept under MHCs, even within jurisdictions. Considering only the cases where providers were named in court records, three providers provided post-MHC start services in Cache, 21 in Salt Lake[13], six in Utah, and eight in Weber. Eight clients in Cache had no provider listed, while this occurred for 53 clients in Salt Lake, 25 in Utah, and one in Weber.

Among the cases for whom providers were specified, when these diverse providers reported to DHS, they did not do so for all clients. Table 12 shows a summary of the number and percentage of MHC clients' cases reported to DHS by provider. Only providers with five or more MHC clients are shown, but it should be noted that no providers with more than two MHC clients reported data for all clients to DHS. As seen in Table 12, the percentage of MHC clients' data reported to DHS varied widely by provider, ranging from 22.2% to 90.0%.

Although these data reflect services after mental health court start (rather than pre-start, as would be used in matching), they highlight the fact that some providers appear not to report to DHS regularly if at all. The purpose of providing these statistics is not to call out individual providers or even to place the responsibility for reporting MHC clients' outcomes on the providers. Rather, the purpose is to highlight the data-related challenges an evaluation of MHCs in Utah would face. Given the large number of providers (33 named providers in all), it is likely not possible to obtain mental health data (particularly reliable, quality data) from all of them individually; this may be especially true of private providers. If accurate mental health histories cannot be obtained owing to a lack of data, it follows that accurate matching might not be possible. A post-MHC outcome analysis would face the same problem.

Table 12: DHS reporting frequency for providers with at least *five* MHC clients

Provider	Number of MHC Clients	Number of Clients with Data Reported to DHS	Percentage Reported to DHS
Bear River Drug & Alcohol	5	2	40.0%
Bear River Mental Health	9	2	22.2%
O.U.T. (Utah Cnty Jail SUD program)	6	4	66.7%
Utah State Hospital	12	7	58.3%
Veteran Affairs (VA)	10	2	20.0%
Valley Behavioral Health (VBH) - combined	184	102	55.4%
CORE	*22*	*9*	40.9%
CTP	*16*	*9*	56.3%
Forensics	*112*	*65*	58.0%
JDOT	*17*	*13*	76.5%
Wasatch Mental Health	20	18	90.0%
Weber Human Services	17	11	64.7%

One solution to the problem of obtaining data from all providers is to systematically obtain data only for cases reporting to DHS. As discussed above, this approach was used in the present study, but this too is problematic for several reasons. First, as already mentioned, individuals who obtain services from entities reporting to DHS may differ in systematic ways from those who obtain services from other sources. Second, simply because DHS has a mental health history for a person

[13] Private providers not otherwise named were grouped together as one provider in Salt Lake.

does not indicate it has the entire history; the available history may, in fact, be highly misleading. For example, consider an individual who has obtained services from two providers. Provider 1 reported to DHS, but the individual saw the provider only a small number of times. Provider 2 is the individual's primary mental health provider and does not report to DHS. In this scenario, it is likely that only a partial history and corresponding diagnoses were obtained; matching would, therefore, occur as if the absence of a diagnosis indicated a lack of that particular mental health issue when, in reality, it merely indicates data were not available.

Because of a lack of full reporting to DHS, matching on partial histories is likely the cause of matching failures observed in the present study. This falls under explanation two for reasons matching might fail: incorrect or incomplete identification of an appropriate comparison group. The matching process depends on complete and accurate covariates; quite simply, the algorithms cannot match on what they do not know.

Continuing Challenges

The review of the state of MHC evaluation research highlighted some of the challenges that will make an evaluation difficult in Utah as in other jurisdictions. The state of the literature on MHC evaluations is replete with methodological caveats. As discussed above, studies often did not match on mental health variables owing to a lack of access, and studies that were able to consider mental health variables did so because they had access to a centralized database to which all community providers reported (Lowder et al., 2016).

Other studies obtained a comparison group by placing strict limitations on how MHCs qualified for study participation. One multi-site study established inclusion criteria for participating MHCs that included having large caseloads for the creation of a treatment group, stable long-term operation, and large county jails from which to sample (Steadman, Redlich, Callahan, Robbins, & Vesselinov, 2011). Such limitations, if adopted in Utah, would limit the study to an examination of the Salt Lake MHC, but adopting these limitations would leave other issues unresolved. For example, the Salt Lake jail does not have detailed mental health service and treatment records (beyond services occurring in jail) and MHCs cannot provide these records either; any matching that occurred would be based only on criminal justice variables, or it would again be based on partial matching, which failed in this study.

In the case of the current study, it was anticipated by UCJC and stakeholders alike that DHS was, in fact, a centralized source of mental health treatment data, but, at least for the population of MHC cases and those considered as comparison cases, that was not the case. Any future evaluation of MHCs in the state of Utah will have to address this lack of centralized, reliable data.

The characteristics of clients served by MHCs in Utah also present a continuing challenge. It is likely not accurate to assume MHCs are equally efficacious with all clients. This suggests any evaluation of MHCs should consider variables that differentiate clients in terms of need and symptomology. However, with the exception of Salt Lake's MHC, most MHCs do not have a sufficient number of cases to further split an analysis by these variables and still hope to find an effect. This presents a notable difficulty because, while it is likely the case that the effect of MHC and corresponding services differs by characteristics of the clients, there are not enough cases to interpret how that difference manifests in terms of outcomes.

Though not discussed in detail because of a failure to identify a matched comparison sample, the outcome analysis from the present study would have also faced difficulty interpreting the efficacy of Utah's MHCs owing to the diverse range of services one receives while in MHC. Table 12 highlights the fact that what qualifies as mental health treatment in Utah for those served by Utah's MHCs is quite diverse, ranging from "light touch" and outpatient services to more intensive inpatient and residential services. This diverse range of services is partly a consequence of serving a diverse population, even within the same MHC, but it makes evaluating MHCs difficult. Some service providers will, naturally, be more effective than others; some clients may receive either insufficient or overly intensive services. These differences in the quality of mental health providers will, inevitably, manifest as differences in the efficacy of MHCs. To some extent then, the determination of whether an MHC is efficacious is inexorably connected to the providers the court allows to provide services.

Recommendations

While there are myriad challenges posed to researchers studying MHCs, access to data will likely prove the most persistent challenge when studying them in Utah specifically. Most research on evaluations of MHCs has focused attention on tracking outcomes from the point of study participation. This amounts to conducting the study prospectively rather than retrospectively (as in the current study), which could be helpful for tracking mental health outcomes, particularly among comparison group members. However, this approach also limits the available follow up time during which outcomes are tracked unless funding for a multi-year study can be secured.

Using a prospective approach, other recommended procedures include having researchers track participants over the course of the MHC or treatment as usual processes by checking in with clinical, court, and corrections staff on a weekly basis and having mental health professionals working in the jails in order to record diagnoses, services, and other outcomes (Cosden et al., 2003; McNiel et al., 2015; McNiel and Binder, 2007; Steadman, Redlich, Callahan, Robbins, & Vesselinov, 2011). The feasibility of adopting these suggestions is largely a matter of resources. Prospective studies involving frequent contact with agency personnel, are expensive, and are particularly challenging when a study involves multiple courts. If a prospective study were adopted in Utah, it would be wise to pilot the effort in one MHC before expanding it to all MHCs.

However, these recommendations neglect the greater issue of how to identify and match treatment and comparison cases based on mental health histories. While MHCs would benefit from requiring that service providers report mental health information either to the court itself or to DHS this too would only benefit a post-MHC outcome analysis; it would not resolve the issue of identifying an accurate comparison group and it would not improve the overall matching process.

Given the existing structure of MHCs, treatment services in Utah, and data limitations, only tentative recommendations can be made to guide future research, and these recommendations necessitate methodological caveats. Among retrospective options, perhaps the best available recommendation for future research would be to use DHS as the starting point for both MHC and comparison cases. Benefits and limitations of this approach are discussed.

While DHS does not have all mental health service records, it is a large repository of such information, and that information is maintained in an organized and reliable fashion. For that reason, future research might decide to start with DHS as the primary source of both MHC and potential comparison cases. This would require access to most, if not all, DHS records for a specified time period. From that large dataset, one could identify cases that later entered MHC and could then use the histories of those MHC participants to match to all other DHS data where a client did not enter MHC (at least in a specific time period).

Several limitations follow, however. First, treatment service data are, rightfully, difficult to obtain owing to client protections associated with protected health information. DHS typically requires a target sample be identified before sharing data; access to all DHS data would likely require the matching process be performed by a DHS employee rather than risk violations of confidentiality and anonymity that would occur if access to all data was granted. Depending on resources, it may be necessary to fund this position in order to make the request feasible.

Second, and as was already mentioned, clients whose mental health providers report to DHS may differ in fundamental ways from those who do not. Relatedly, this means DHS would only have records for some of a particular MHCs clients rather than all. Also, DHS may be an incomplete source of mental health histories for individuals who received services from multiple providers, some who reported to DHS and some who did not. This creates the same problem as was observed in the current study: matching on partial treatment histories is likely to lead to a failure to obtain an accurate match. Some of these concerns may be attenuated by both partnering with DHS at early stage of study design and by providing funding for an internal position at DHS that could offer assistance in the matching process.

Because of the limitations of any retrospective design, a prospective randomized control trial (RCT) is likely the best methodology considering the data limitations inherent in the study of MHCs. Using any other method, the lack of access to data on mental health histories will remain difficult to overcome given dispersed treatment providers. In an RCT design, eligible participants are randomly assigned to treatment or control (treatment as usual) conditions. Thus, an RCT presents a strong option in terms of drawing valid inferences about whether Utah's MHCs "work" (however defined) because it removes the bias created by differing mental health treatment and service histories, as well as different criminal histories, through the random assignment mechanism. In the case of an RCT, one only needs to know that cases qualify for MHC; the specifics of their criminal and treatment histories are not required.

The question remains, however, whether an RCT is an ethical approach in the evaluation of Utah's MHCs. There are arguments, frequently cited when implementing RCTs in social research, that would suggest the approach is warranted. Whenever the effect of a program is not yet known, as is the case of Utah's MHCs, an RCT is often justified by the fact that a service is not being withheld that is known to be more efficacious than treatment as usual. On the grounds that the effect of Utah's MHCs in not yet known, an RCT is justifiable. However, it is also important to consider that RCTs are prospective, and, because of that, they are both time-consuming and resources intensive.

References

Anestis, J. C., & Carbonell, J. L. (2014). Stopping the revolving door: Effectiveness of mental health court in reducing recidivism by mentally ill offenders. *Psychiatric Services, 65*(9), 1105-1112.

Andrews, D. A., & Bonta, J. (1995*). The level of service inventory–revised.* Toronto, Ontario, Canada: Multi-Health Systems.

Andrews, D. A., Bonta, J. L., & Wormith, J. S. (2004). *Level of service/case management inventory (LS/CMI): An offender assessment system, user's manual.* Toronto, Ontario, Canada: Multi-Health Systems.

Beckstead, D. J., Hatch, A. L., Lambert, M. J., Eggett, D. L., Goates, M. K., & Vermeersch, D. A. (2003). Clinical significance of the Outcome Questionnaire (OQ-45.2). *The Behavior Analyst Today, 4*(1), 86-97. doi:10.1037/h0100015

Berman, G., & Feinblatt, J. (2001). Problem-solving courts: A brief primer. *Law & Policy*, *23*(2), 125-140.

Berman, G., & Fox, A. (2010). The future of problem-solving justice: An international perspective. *U. Md. LJ Race, Religion, Gender & Class*, *10*(1), 1-24.

Blevins, C., & Mullen, L. (2015). Jane, John…Leslie? A historical method for algorithmic gender prediction. *Digital Humanities Quarterly, 9*(3).

Campbell, M. A., Canales, D. D., Ran, W., Totten, A. E., Macaulay, W. A. C., & Wershler, J. L. (2015). Multidimensional evaluation of a mental health court: Adherence to the risk-need-responsivity model. *Law & Human Behavior, 39*(5), 489-502.

Cosden, M., Ellens, J. K., Schnell, J. L., Yamini-Diouf, Y., & Wolfe, M. M. (2003). Evaluation of a mental health treatment court with assertive community treatment. *Behavioral Science Law, 21*(4), 415-427.

Fiduccia, C. E., & Rogers, R. (2012). Final-stage diversion. *Criminal Justice and Behavior, 39*(4), 571-583.

Ford, M. (2015, June 8). America's largest mental hospital is a jail. *The Atlantic*. Retrieved from https://www.theatlantic.com/politics/archive/2015/06/americas-largest-mental-hospital-is-a-jail/395012/

Gendreau, P., Goggin, C., & Smith, P. (2002). Is the PCL-R really the "unparalleled" measure of offender risk? A lesson in knowledge cumulation. *Criminal Justice and Behavior, 29,* 397-426.

Goldkamp, J. S., & Irons-Guynn, C. (2000). *Emerging judicial strategies for the mentally ill in the criminal caseload: Mental health courts in Fort Lauderdale, Seattle, San Bernardino, and Anchorage.* Washington, D.C.: Bureau of Justice Assistance, U.S. Department of Justice.

Guo, S., & Fraser, M. W. (2011). *Propensity score analysis: Statistical methods and applications*. Los Angeles: SAGE.

Hiday, V. A., Wales, H. W., & Ray, B. (2013). Effectiveness of a short-term mental health court: Criminal recidivism one year postexit. *Law and Human Behavior, 37*(6), 401-411.

Honegger, L. N. (2015). Does the evidence support the case for mental health courts? A review of the literature. *Law & Human Behavior, 39*(5), 478-488.

Leite, W. L. (2016). *Practical propensity score methods using r*. Sage Publications Inc.

Lim, L., & Day, A. (2014). Mental health diversion courts: A two year recidivism study of a South Australian mental health court program. *Behavioral Sciences & the Law, 32*(4), 539-551.

Lowder, E. M., Desmarais, S. L., & Baucom, D. J. (2016). Recidivism following mental health court exit: Between and within-group comparisons. *Law & Human Behavior 40*(2), 118-127.

Luskin, M. L. (2013). More of the same? Treatment in mental health courts. *Law and Human Behavior, 37*(4), 255-266.

McGaha, A., Boothroyd, R. A., Poythress, N. G., Petrila, J., & Ort, R. G. (2002). Lessons from the Broward County mental health court evaluation. *Evaluation and Program Planning, 25*(2), 125-135.

McNiel, D. E., & Binder, R. L. (2007). Effectiveness of a mental health court in reducing criminal recidivism and violence. *American Journal of Psychiatry, 164*(9), 1395-1403.

McNiel, D. E., Sadeh, N., Delucchi, K. L., & Binder, R. L. (2015). Prospective study of violence risk reduction by a mental health court. *Psychiatric services, 66*(6), 598-603.

Prins, S. J. (2014). Prevalence of mental illnesses in US state prisons: A systematic review. *Psychiatric Services*, *65*(7), 862-872.

R Core Team. (2017). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models applications and data analysis methods*. Thousand Oaks: Sage.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika,* 70, 41-55.

Sarteschi, C. M., Vaughn, M. G., & Kim, K. (2011). Assessing the effectiveness of mental health courts: A quantitative review. *Journal of Criminal Justice, 39*(1), 12-20.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Steadman, H. J. (2005). *A guide to collecting mental health court outcome data.* New York, NY: Council of State Governments.

Steadman, H. J., Callahan, L., Robbins, P. C., Vesselinov, R., Mcguire, T. G., & Morrissey, J. P. (2014). Criminal justice and behavioral health care costs of mental health court participants: A six-year study. *Psychiatric Services, 65*(9), 1100-1104.

Steadman, H. J., Redlich, A., Callahan, L., Robbins, P. C., & Vesselinov, R. (2011). Effect of mental health courts on arrests and jail days: A multisite study. *Archives of General Psychiatry, 68*(2), 167-172.

Wolff, N., & Pogorzelski, W. (2005). Measuring the effectiveness of mental health courts: Challenges and recommendations. *Psychology, Public Policy, and Law*, *11*(4), 539-569.

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. Boca Raton: CRC Press.

Worwood, E. B., Sarver, C., Borgia, A. D., & Butters, R. P. (2015). *Statewide evaluation of Utah mental health courts: Phase I report.* Salt Lake City, UT: Utah Criminal Justice Center, University of Utah.

World Health Organization. (1992). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines*. Geneva: World Health Organization.

Appendix I: ICD-10 Categories

- Mental and behavioral disorders due to psychoactive substance use [F10-F19] – this category includes, as examples, alcohol and opioid-related disorders, including dependence and abuse.
- Mood [affective] disorders [F30-F39] – this category includes, as examples, bipolar disorder, major depressive disorder, and persistent mood disorders.
- Schizophrenia, schizotypal, delusional, and other non-mood psychotic disorders [F20-F29] – this category includes, as examples, schizophrenia, delusional disorders, and psychoses.
- Disorders of adult personality and behavior [F60-F69] – this category includes, as examples, antisocial, borderline, histrionic, and narcissistic personality disorders.
- Persons encountering health services for examinations [Z00-Z13] – Although this category contains a number of reasons for medical observation, in the current sample, it was dominated by code Z03.89, "Encounter for observation for other suspected diseases and conditions ruled out".
- Anxiety, dissociative, stress-related, somatoform and other nonpsychotic mental disorders [F40-F48] – this category includes, as examples, phobias, anxiety disorders, obsessive-compulsive disorder, and stress reactions.
- Behavioral syndromes associated with physiological disturbances and physical factors [F50-F59] – this category includes, as examples, eating disorders, sleep disorders, and sexual dysfunction.
- Behavioral and emotional disorders with onset usually occurring in childhood and adolescence [F90-F98] – this category includes, as examples, attention-deficit hyperactivity disorders and conduct disorders. Though other diagnoses are include in the category, these two diagnoses (F90 and F91) were the only codes found in the current sample under this classification.
- General symptoms and signs [R50-69] – this category is largely characterized by physiological rather than psychological symptomology. Examples in this category include headache, fever, and unspecified pain. In the current sample, code R69 (illness unspecified) dominated codes within this category.
- Mental disorders due to known physiological conditions [F01-F09] – this category includes, as examples, dementia, delirium, mood, and psychotic disorders specifically known to be caused by a physiological condition.
- Intellectual disabilities [F70-F79] – while this category includes intellectual disabilities of all levels, in the current sample, only code F70, mild intellectual disability, was found.
- Pervasive and specific developmental disorders [F80 – F89] – this category includes, as examples, autism, Asperger's, neurodevelopment disorder, reading, and coordination disorders.
- Persons with potential health hazards related to socioeconomic and psychosocial circumstances [Z55-Z65] – this category includes, as examples, problems related to the death or disappearance of a family member, divorce, stress due to return of a family member (e.g., from military service), stress due to incarceration, and being the victim of crime.
- Persons encountering health services in other circumstances [Z69-Z76] – this category includes, as examples, encounters for mental health services related to counseling for

victims and perpetrators of abuse and problems related to lifestyle (smoking, poor diet, lack of exercise).

- Other and unspecified effects of external causes [T66-T78] – Although this category contains a number of illnesses, such as radiation sickness and hypothermia, in the current sample, it was dominated by code T74, "Adult and child abuse, neglect and other maltreatment, confirmed".

Appendix II: PSM Matching Algorithms

- Nearest neighbor: The nearest neighbor method attempts to match each treatment case with the closest matching comparison case where "closest" is defined in terms of the propensity score. Most implementations of this algorithm adopt what is known as a caliper. The caliper dictates how far apart (in terms of standard deviations) two cases can be. If the caliper is exceeded, cases will not be matched. This method also allows for one-to-one or one-to-many matching. In the latter instance, more than one comparison case can be matched to a single treatment case. Notably, while this method attempts to limit how different matched cases can be, it does not minimize imbalance across the entire sample (as other methods do).
- Optimal matching: Optimal matching was created to address the fact that nearest neighbor matching does not address the total difference across all cases. While optimal matching does account for the total difference across all cases, that does not guarantee it will be a superior method. This method also allows for one-to-one or one-to-many matching.
- Coarsened exact matching: Coarsened exact matching is a relatively new technique that uses binning to create matched groups. Here, binning refers to breaking a variable (e.g., age) and the matching only people who share the same bin across the various predictors. The problem with this method is that, across a large number of variables, a bin combination for a treatment case may not occur for a comparison case or vice versa. For that reason, this method tends to eliminate a large number of cases unless the bins are relatively coarse.
- Full matching: Full matching derives its name from the fact that cases are not discarded. Treatment cases are matched to comparison cases without replacement. This method tends to perform well when there are relatively large differences in the propensities of treated and comparison cases. Indeed, it generally did fairly well in the present study, where large differences were observed.
- Genetic matching is probably the most advanced matching algorithm. Though this alone does not guarantee it will produce the best matches, it often does. Though often treated as a PSM method, it is actually its own method, separate from PSM. It is treated as a PSM method because the purpose is the same. Genetic matching seeks matches that minimize covariate imbalance across the set of covariates. It does this by checking a large number of diagnostics during the matching process and iteratively recalculating matches based on results. The only drawback of the method is the time it takes to create matches, which can be on the order of minutes to hours rather than the seconds taken by other methods.

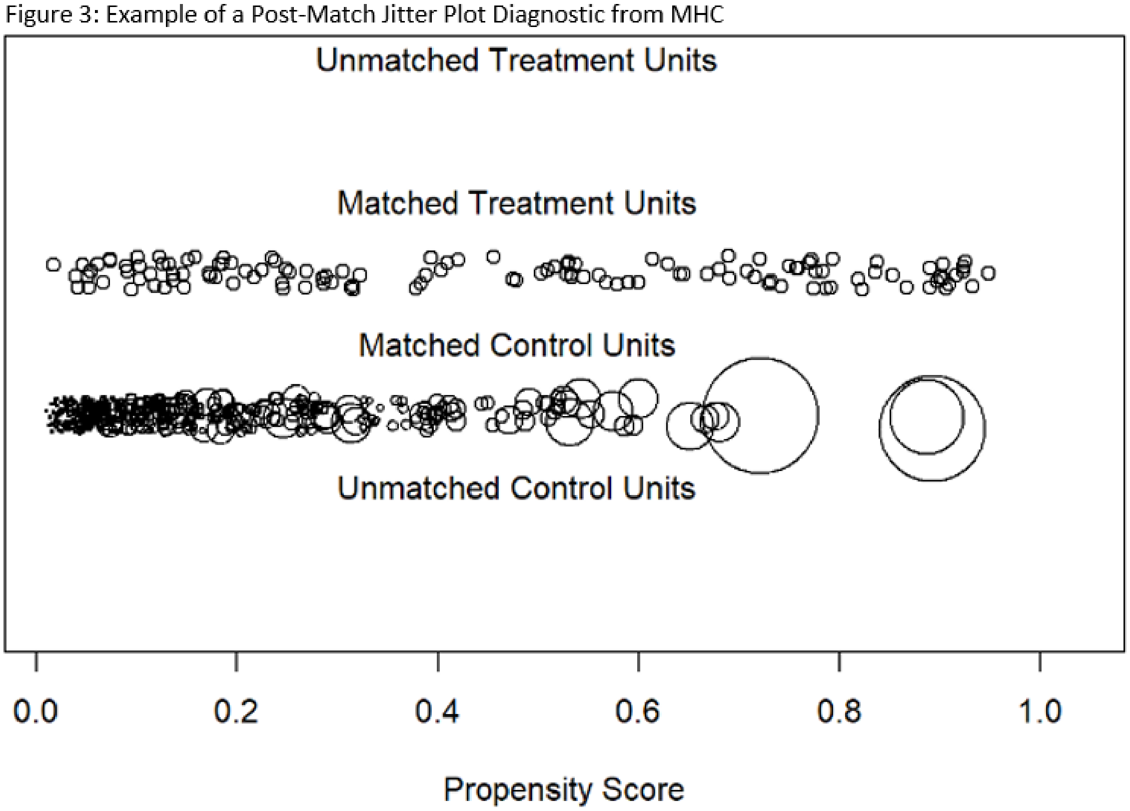
Appendix III: Further Results of PSM Diagnostics

In this appendix, additional diagnostics are presented to show further evidence of the failure of PSM matching with MHC and comparison data. One such diagnostic is a jitter-plot, which is really just a different display of the results provided in Figures 1 and 2, in the main text. However, whereas those figures represent an indication of imbalance pre-matching (i.e., is it likely to work?), jitter plots provide an indicator of balance post-matching (i.e., did it work?) that is a function of the particular algorithm's success.

Figure 3 provides an example from one particular algorithm, full matching, within one MHC. In reality, 36 such figures were evaluated across algorithms and courts. This particular example was selected for presentation because it most clearly displays the nature of the problem faced for the MHC study.

Full matching forces the use of all available cases, which can have positive or negative consequences, depending on features of the actual data. The figure displays the propensity of a case to be in MHC among the treated and untreated, comparison cases. For the row "Matched Control Units", the size of the circles is a visual display of the weight given to a case. A comparison case that is matched to several treatment cases will be represented in this row by a large circle. At the far right, one can see three comparison cases have extremely large weights because they are the only comparison cases with a high propensity to be in the treated, MHC group.

Figure 3: Example of a Post-Match Jitter Plot Diagnostic from MHC



Consider the largest circle, third from the far right. The algorithm was tasked with matching *all* cases the best it could given the available comparison cases. To accomplish that task, this case was matched to 16 treatment cases. While it is true that, within the confines of the algorithm, that single comparison case was the best match, one might reasonably wonder if it was actually a *good* match for the 16 cases, or merely the *best available*. That issue is addressed in more detail below.

Different algorithms could avoid the issue of one comparison case having such a large impact using different settings and matching criteria. Nearest neighbor matching, for example, will allow cases to be discarded if the cases become “too dissimilar”. However, as with all algorithms, the decision to discard cases with no acceptable match comes at a cost. In the case of the current study, discarding cases with no acceptable match often meant removing the MHC cases with the greatest propensity to be in MHC. Conclusions drawn from a study that excluded the cases most likely to be in MHC should, reasonably, be questioned in terms of generalizability.

The aforementioned discussion provides a couple of examples of some of the challenges that prevented adequate matching in the current study. It is certainly not exhaustive of all of the challenges encountered, but hopefully it highlights the fact that matching, by any method, is limited to the quality of the available cases. Matching algorithms will do as they are told; they will match cases, but they cannot remedy a lack of available comparison cases that are similar to MHC cases. In that sense, they are limited in the degree of balance they can achieve.

One objective criterion used to evaluate the quality of matched balance is the standardized mean difference, which is regarded as the most important metric used to indicate balance. One can conceive of the standardized mean difference as a measure of how far apart the treatment and comparison cases are on the matching variables after matching has occurred. Ideally, the standardized mean difference should not exceed .1 on any variable used in the matching process (Guo & Fraser, 2009; Leite, 2016).

Figure 4 shows a typical result from one MHC across several matching algorithms. The method of coarsened exact matching is not represented in the figure because, though it often came close to achieving an acceptable standardized mean difference, it did so at the cost of eliminating far too many cases (in some cases, over 90% of all MHC cases were removed by the method). Each of the remaining methods are provided along the y-axis and are ranked, in descending order, by how close they came to meeting the criterion of a .1 standardized mean difference. The maximum standardized mean difference is provided along the x-axis, and a vertical line has been drawn on the graph at the cutoff value of .1. Though some methods performed better than others, none came close to meeting the recommended criterion.

There are other diagnostics one could evaluate as well, such as percentage improvement in balance achieved, but the common finding across all metrics was that, not only was balanced not achieved, comparison cases were simply not available to represent the MHC cases with a high propensity of being in MHC.

Figure 4: Maximum Standardized Mean Difference by Matching Algorithm