ASPE
ASSISTANT SECRETARY FOR PLANNING AND EVALUATION

OFFICE OF BEHAVIORAL HEALTH, DISABILITY, AND AGING POLICY

Feasibility of Calculating Measures to Monitor Quality Performance of Behavioral Health Programs

Prepared for
the Office of the Assistant Secretary for Planning and Evaluation (ASPE)
at the U.S. Department of Health & Human Services

by
Mathematica

May 2024

Office of the Assistant Secretary for Planning and Evaluation

The Assistant Secretary for Planning and Evaluation (ASPE) advises the Secretary of the U.S. Department of Health and Human Services (HHS) on policy development in health, disability, human services, data, and science; and provides advice and analysis on economic policy. ASPE leads special initiatives; coordinates the Department's evaluation, research, and demonstration activities; and manages cross-Department planning activities such as strategic planning, legislative planning, and review of regulations. Integral to this role, ASPE conducts research and evaluation studies; develops policy analyses; and estimates the cost and benefits of policy alternatives under consideration by the Department or Congress.

Office of Behavioral Health, Disability, and Aging Policy

The Office of Behavioral Health, Disability, and Aging Policy (BHDAP) focuses on policies and programs that support the independence, productivity, health and well-being, and long-term care needs of people with disabilities, older adults, and people with mental and substance use disorders. Visit BHDAP at https://aspe.hhs.gov/about/offices/bhdap for all their research activity.

NOTE: *BHDAP was previously known as the Office of Disability, Aging, and Long-Term Care Policy (DALTCP). Only our office name has changed, not our mission, portfolio, or policy focus.*

This research was funded by the U.S. Department of Health and Human Services Office of the Assistant Secretary for Planning and Evaluation under Contract and carried out by Mathematica Policy Research. Please visit https://aspe.hhs.gov/topics/behavioral-health for more information about ASPE research on behavioral health.

FEASIBILITY OF CALCULATING MEASURES TO MONITOR QUALITY PERFORMANCE OF BEHAVIORAL HEALTH PROGRAMS

Authors

Rachel Gringlas
Rachel Miller
Julia Baller
Mathematica

May 1, 2024

Prepared for

Office of Behavioral Health, Disability, and Aging Policy
Office of the Assistant Secretary for Planning and Evaluation
U.S. Department of Health and Human Services

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ACRONYMS

The following acronyms are mentioned in this report and/or appendices.

AMM-AD	Antidepressant Medication Management, Adult Version
AOD	Alcohol and Other Drug
CCBHC	Certified Community Behavioral Health Clinic
CCBHC-E	Certified Community Behavioral Health Clinic Expansion
CHIP	Children's Health Insurance Program
CMS	Centers for Medicare & Medicaid Services
DY	Demonstration Year
ED	Emergency Department
FFY	Federal Fiscal Year
FUA-AD	Follow-up after Emergency Department Visit for Alcohol or Other Drug Dependence, Adult Version
FUA-CH	Follow-up after Emergency Department Visit for Alcohol or Other Drug Dependence, Child Version
FUH-AD	Follow-up After Hospitalization for Mental Illness, Adult Version
FUH-CH	Follow-up After Hospitalization for Mental Illness, Child Version
FUM-AD	Follow-up After Emergency Department Visit for Mental Illness, Adult Version
FUM-CH	Follow-up After Emergency Department Visit for Mental Illness, Child Version
ID	Identifier
IET-AD	Initiation and Engagement of Alcohol and Other Drug Dependence Treatment, Adult Version
NPI	National Provider Identifier
NPPES	National Plan and Provider Enumeration System
NQF	National Quality Forum
SAMHSA	Substance Abuse and Mental Health Services Administration
SMI	Serious Mental Illness
SNR	Signal-to-Noise Ratio
SUD	Substance Use Disorder

T-MSIS	Transformed Medicaid Statistical Information System
TAF	T-MSIS Analytic Files

I. Introduction

The possibility of calculating behavioral health quality measures at the clinic level holds great promise for monitoring clinic performance over time, and for providing information for clinics to use to revise their processes and procedures to improve their performance. This report describes a novel process of testing the feasibility of using the Transformed Medicaid Statistical Information System (T-MSIS) Analytic Files (TAF) data to calculate behavioral health quality measures at the clinic level.

The report begins by explaining Mathematica's process of assessing the feasibility of calculating behavioral health quality measures at the clinic level. We describe how we obtained clinic identifiers (IDs), searched for clinics in the TAF data, attributed beneficiaries to clinics, and then calculated the measures. We then test the reliability and validity of the calculated quality measures. Finally, we explore potential applications for these findings in future work.

For this novel analysis, we used a set of Certified Community Behavioral Health Clinic Expansion (CCBHC-E) grantee clinics as a test case for determining the feasibility of calculating clinic-level performance on claims-based quality measures. In 2018, the Substance Abuse and Mental Health Services Administration (SAMHSA) launched the CCBHC-E grant program to expand the CCBHC model in states that received planning grants for the federal demonstration program. Through this program, clinics receive time-limited grant funding to provide services that meet the CCBHC certification criteria. These services include round-the-clock crisis intervention services for people with serious mental illness (SMI) or substance use disorders (SUDs), including opioid use disorder; children and adolescents with serious emotional disturbance; and people with co-occurring mental health and SUDs.

The following research questions guided the feasibility analysis:

1. Can national Medicaid claims and encounter data (from TAF) be used to calculate clinic-level performance on behavioral health quality measures?
    a. Can clinics be identified in claims and encounter data?
    b. Can the client population of each clinic be identified in the data?
    c. Can the client population be attributed to a clinic?
    d. Which behavioral health quality measures are feasible to calculate?
2. If the proposed methodology proves feasible, how does performance on the behavioral health quality measures change over time among the test case clinics?
3. If the proposed methodology proves feasible, how does performance on the behavioral health quality measures compare to national averages?

Our findings, based on a sample of five 2018 CCBHC-E grantee states, suggest that identifying CCBHC-E clinics (referred to simply as clinics in this report) in the TAF data and calculating behavioral health quality measures at the clinic level is challenging in most states. Although three of the five states in our sample had created IDs for CCBHC-E clinics, only one of the states' clinic-level IDs was present and complete in the TAF data. For the other states, we used state IDs or publicly available National Provider Identifiers (NPIs) to identify the health organizations that operate the CCBHC-E clinics, but could not identify the individual clinics themselves.

For the purposes of this analysis, we define clinics as individual locations that provide direct care and services, usually managed by a health organization that typically operates multiple clinic locations, which may provide a range of different or similar services. We define organization as a health care providing entity that operates one or more individual clinic locations where care is actually provided. In rare instances, organizations operate only one clinic location, so the organization and the clinic are the same entity. In the one state where we successfully identified individual clinics in the TAF data, we could attribute client populations to the clinics. However, the counts of beneficiaries attributed to most of the individual clinics in this state were very small ($n < 500$), resulting in high likelihood that there would be insufficient numbers of qualifying events or beneficiaries to calculate the measures *at the clinic level* even in this state. Therefore, it was feasible to calculate behavioral health quality measures at the clinic level for only two clinics in one state in our sample.

However, we could calculate the behavioral health quality measures *at the organization level* in all five states by either rolling up individual clinic IDs under a single organization or using organization-level IDs, depending on the state. Calculating the measures at the organization level did not enable us to look at an individual clinics' performance, and the connection between changes in practices and service provision at the clinic level and performance at the organization level is less clear. It is possible, however, that changed practices from the CCBHC-E clinics (or any clinics participating in a behavioral health demonstration or intervention) would filter throughout the organization as a whole, though this would likely take more time than we were able to include in this study. Nonetheless, this organization-level strategy enabled us to track the performance of the organizations under which the CCBHC-Es operate and offers strong potential for future monitoring of behavioral health programs that occur at the organization level.

Overall, calculating reliable, valid behavioral health quality measures at the organization level was feasible for most organizations and years in all states included in this analysis. Our findings highlight the potential utility of monitoring behavioral health *organizations'* performance on quality measures over time, and we encourage states to develop methods to identify behavioral health *clinics* in federal Medicaid data so they could extend these kinds of analyses to the clinic level as well.

II. Feasibility of Using the TAF to Calculate Clinic-Level Performance on Behavioral Health Quality Measures

This section describes our primary data source for the feasibility testing, the TAF, and outlines our processes for: (1) obtaining clinic IDs; (2) identifying clinics in the TAF data; (3) attributing beneficiaries to clinics in the data; and (4) calculating the behavioral health quality measures at the organization level.

The TAF Data

The TAF is a research version of state T-MSIS submissions. Through the Virtual Research Data Center Innovator program, Mathematica has a data use agreement with the Centers for Medicare & Medicaid Services (CMS) that allowed us to use the TAF for this project. Data covered calendar years 2017, 2018, and 2019, capturing a baseline of two years before SAMHSA awarded the 2018 CCBHC-E grants and the first year of grant implementation. It is important to note, however, that some of the CCBHC-Es were demonstration CCBHCs prior to the expansion grants, meaning that data from years 2017 and 2018 cannot completely serve as a baseline for these clinics.

The TAF are standardized across states and are, for the most part, clean and well populated. This makes measure calculations easier than using state-provided Medicaid data, which vary widely by state in terms of format and completeness. However, there are a limited number of provider identifier fields on TAF claims, which are crucial fields for this kind of analysis attempting to identify individual clinics in claims. There is also a lack of CMS guidance on which provider identifiers to include on claims, which can lead to variances in reporting across states and clinics. For instance, a state can report several different provider identifiers on a claim, but CMS does not require states to report all identifiers and does not provide specific guidance on which identifiers to use across all states. Therefore, a claim can have a Medicaid provider identifier, an NPI, or both, leading to potential undercounting of claims for this kind of analysis depending on which identifier field(s) we use to identify clinics in each state.

Step 1: Obtain Clinic IDs

First, we attempted to obtain clinic identifiers for each of the five CCBHC-E grantee states in our sample.[1] We reached out to Medicaid representatives from these five states to determine if they maintained a list of CCBHC-E clinic identifiers we could search for in the TAF data. Two of the five states--States B and C--tracked their CCBHC-E clinics using state-specific identifiers reported on Medicaid claims that they expected to be present in the national-level TAF data. One other state--State A--also tracked its clinics using state-specific IDs, but the IDs were tracked in a data field that was present in the state's Medicaid data but not present in the TAF. Instead, this state provided a list of current NPIs for the CCBHC-E organizations, knowing there were a few clinics under the organizations not actually using the CCBHC-E grants. The remaining two states--States D and E--did not track the CCBHC-Es in their data and could not provide a list of IDs at any level. Instead, we used grantees' names and addresses to search for NPIs on the CMS National Plan and Provider Enumeration System (NPPES) NPI Registry.[2] We summarize the results of the process of searching for clinic IDs in **Table 1**.

[1] To protect state and CCBHC-E confidentiality, we refer to the selected CCBHC-E grantee states throughout this report as State A, State B, State C, State D, and State E. Clinics and organizations are similarly deidentified as Clinic 1, Organization 1, and so on.

[2] Centers for Medicare & Medicaid Services. "Search NPI Records." Available at https://npiregistry.cms.hhs.gov/.

Table 1. Clinic ID Types and Formats, by State	
Grantee State	**Clinic ID Types and Formats Obtained for This Analysis**
State A	State-provided organization-level NPIs validated through NPPES
State B	Two separate state-provided ID values that when combined create a clinic-level ID; one or more of the first ID values capture an organization, and one of the second ID values capture a clinic under an organization when paired with the first ID values
State C	State-provided clinic-level IDs: one ID per clinic
State D	State could not provide IDs; obtained organization-level NPIs for grantees' names and addresses through NPPES
State E	State could not provide IDs; obtained organization-level NPIs for grantees' names and addresses through NPPES

Overall, the process of working with state Medicaid representatives to obtain clinics' IDs was time intensive. Searching NPPES for clinic-level NPIs was also time consuming because it required manually searching all possible variations of clinics' names and addresses and cross-referencing against the organizations' websites. The NPI search process was ultimately unsuccessful at the clinic level; NPIs could not be mapped to individual clinics, and we could only obtain organization-level NPIs for the organizations under which the CCBHC-Es operate.

Based on the results of this process of working with states and searching NPPES, we concluded it is not feasible to calculate the measures for States A, D, or E at the *clinic level*. We attempted to calculate the measures at the organization level for these states and to calculate the measures at the clinic level for States B and C.

Step 2: Identify Clinics

After obtaining clinic or organization IDs for the five states, we searched for the IDs on claims and managed care encounter records in the TAF data.[3]

For states with state-provided clinic-level IDs (States B and C), we had mixed results. State B had two separate ID values that must be combined to identify individual clinics, but only one of the two ID values was present on claims: the value identifying an organization or partial organization.[4] **Therefore, in State B we could identify only the organization or partial organization under which the CCBHC-Es operate, rather than the individual clinics. For State C that had one ID value per clinic, we successfully identified state-provided clinic IDs on TAF claims.**

For states with organization-level NPIs only (States A, D, and E), we found most of the NPIs on the claims. This search was particularly successful for State A because it had provided us with a list of active NPIs; we found all these state-provided NPIs in claims. For States D and E, where we had to rely solely on searching for NPIs in NPPES, we could not find some of the NPIs in claims; we presume these NPIs were inactive and/or out of date. **Given these results, we decided to drop States D and E from our analysis.**

[3] Since most CCBHC-Es bill as outpatient clinics, we limited our search to the TAF Other Services file, which contains records billed on professional claims.
[4] That is, one ID represents some, but usually not all, of the CCBHC-E and non-CCBHC-E billing for that organization.

Step 3: Attribute Beneficiaries

For the three states remaining in the analysis--States A, B, and C--we attributed beneficiaries to clinics or organizations for each analysis year (2017-2019) to enable us to calculate the annual measures. Our primary attribution method was to attribute beneficiaries to clinics or organizations if they had *at least one claim with the relevant provider ID* in the calendar year.[5]

In State C, the only state where we could identify individual clinics in the data, the annual beneficiary counts at the clinic level were very small for many of the CCBHC-E clinics ($n < 500$). Beneficiary counts at the two largest clinics were of adequate size to calculate the measures: Clinics 1 and 2 under State C in Table 2. The remaining clinics' attributed beneficiary counts were simply too small to reliably calculate the quality measures (as low as 63 attributed beneficiaries at one clinic). For these clinics, which fell under two different organizations, we rolled up the clinics into the two broader organizations. However, these clinics do not represent all clinics under the organizations, because not all clinics used the CCBHC-E grant funding at either organization.

Table 2 shows the final beneficiary attribution results. A small number of beneficiaries visited more than one CCBHC-E clinic or organization in a state in each year. In these instances, we attributed the beneficiaries to all clinics they visited in the year.

To assess the accuracy of our counts of attributed beneficiaries, we compared them to the Medicaid beneficiary case load characteristics counts from the CCBHC demonstration quality reports from Demonstration Years 1 and 2.[6] This was not a perfect comparison, as in most cases the demonstration CCBHCs captured in the quality reports did not line up exactly (or at all) with the CCBHC-Es we included in this analysis, a limitation we describe further in the state-specific sections. Our findings were as follows:

State A (*IDs were at the organization level*). All CCBHC-E grantees were also demonstration sites, so we could benchmark directly to the CCBHC quality measure beneficiary counts for Organizations 1, 2, and 3. Our attributed beneficiary counts were similar to the benchmark counts for Organizations 1 and 2. This is likely because all clinics under these two organizations operated as both CCBHCs and CCBHC-Es, making this a direct benchmark comparison. For Organization 3 in State A, our attributed beneficiary count was higher than the benchmark count. We expected this because we knew Organization A had several clinic locations, only some of which used CCBHC-E grant funding.

State B (*IDs were at the organization or partial organization level*). A few CCBHC-E grantees were also demonstration CCBHCs, so we could directly compare benchmark beneficiary counts only for the organizations under which those CCBHC-Es operate (Organizations 2, 7, and 8). For two of the three CCBHC-E organizations, our attributed beneficiary counts were higher than the benchmark counts. We expected this because we know not all clinic locations under each organization were CCBHCs and/or CCBHC-Es, but we captured all clinic locations under each organization in our attributed beneficiary counts.

[5] As a sensitivity check, we also attributed beneficiaries to clinics or organizations at which they had *at least two claims with the relevant provider ID on separate days* in the calendar year. Both attribution methods yielded similar counts; we decided to use the primary attribution method of requiring one claim from a clinic or organization in the calendar year because this provided slightly larger sample sizes for calculating the measures.

[6] CCBHC Demonstration Year 1 (DY1) corresponds roughly with calendar year 2017 used in this analysis, and Demonstration Year 2 (DY2) with calendar year 2018.

Table 2. Number of Beneficiaries Who had at least One Claim from a Clinic or Organization in a Calendar Year			
	2017	2018	2019
State A (total)	**17,671**	**20,564**	**22,366**
Organization 1	10,791	12,832	14,652
Organization 2	5,941	6,605	6,546
Organization 3	1,123	1,346	1,416
2 or more organizations	184	219	246
State B (total)	**38,947**	**41,907**	**41,353**
Organization 1	2,447	2,588	2,799
Organization 2	12,289	14,405	15,226
Organization 3	5,482	4,998	4,130
Organization 4	9,849	10,477	10,403
Organization 5	2,668	2,832	2,959
Organization 6	850	1,044	1,122
Organization 7	6,149	6,858	6,740
Organization 8	406	413	444
2 or more organizations	1,191	1,699	1,466
State C (total)	**11,450**	**12,273**	**13,427**
Clinic 1	4,375	4,692	4,938
Clinic 2	2,249	2,070	1,830
Organization 1[a]	4,446	4,565	4,703
Organization 2[a]	436	1,023	2,048
2 or more clinics or organizations	56	75	88

Source: Mathematica's analysis of 2017-2019 TAF data.

[a] Organizations in State C represent only the CCBHC-E clinics operating under the organization. Organizations in States A and B usually include all clinics operating under the organization (i.e., both CCBHC-E and non-CCBHC-E clinics).

State C (*IDs were at the clinic level*). Several CCBHC-E grantees were also demonstration sites, so we could benchmark beneficiary counts for those organizations. Overall, the beneficiary attribution counts were not similar to the benchmark counts, which we expected based on guidance we received from the state. Therefore, our attributed beneficiary counts and the benchmark counts were not directly comparable.

Overall, despite the limited utility of the CCBHC demonstration quality measures benchmark for validating our beneficiary counts, we are confident in our attributed beneficiary counts at the clinic or organization level for the three states and analysis years.

Step 4: Calculate Behavioral Health Quality Measures

In our final step, we calculated behavioral health quality measures at the clinic or organization level for each state. As previously described, due to either: (1) our inability to obtain clinic-level IDs and/or find the IDs in the TAF (States A and B); or (2) small sample sizes at the clinic level (State C), it was largely

not feasible to calculate the behavioral health quality measures at the clinic level.[7] We therefore calculated the measures at the organization level.

We considered a range of nationally endorsed behavioral health quality measures to calculate for the feasibility testing. To select measures to include in this analysis, we considered whether the measure met three criteria:

1. Relevant to CCBHC-E grantees (that is, the measure assesses the delivery of services that CCBHC-E clinics typically provide).
2. Applies to a broad number of CCBHC-E clients, thus increasing the likelihood we will have a sufficient sample size for calculating the measure.
3. Has available benchmark data.

Table 3 lists the five behavioral health quality measures we ultimately decided to include in this analysis. They are all Medicaid and CHIP Adult and Child Core Set measures (Medicaid Core Set measures), which states report annually, and CMS uses to monitor the quality of health care received by Medicaid beneficiaries. We based our measures on the Medicaid Core Set federal fiscal year (FFY) 2020 technical specifications, with the following modifications:[8]

- To include beneficiaries in the eligible population, we did not require them to have continuous Medicaid eligibility. All the Medicaid Core Set measures required continuous eligibility during at least some of the time period covered by the measure.
- We did not exclude beneficiaries in hospice. All the Medicaid Core Set measures excluded beneficiaries in hospice from the eligible population.

Table 3. Behavioral Health Quality Measures included in the Analysis

Domain	Measure	Medicaid Core Set designation	NQF number
Care coordination	Follow-up After Hospitalization for Mental Illness, Adult Version[a]	FUH-AD	0576
	Follow-up After ED Visit for Mental Illness, Adult Version[a]	FUM-AD	3489
	Follow-up after ED Visit for AOD Dependence, Adult Version[a]	FUA-AD	3488
Medication management and treatment adherence	Antidepressant Medication Management, Adult Version	AMM-AD	0105
	Initiation and Engagement of AOD Dependence Treatment, Adult Version	IET-AD	0004

[a] We intended to include the child versions of the follow-up measures (FUH-CH, FUM-CH, and FUA-CH) in our analysis as well; however, the sample sizes for the child populations at the organizations in all three states were too small to calculate the measures.

[7] Clinics 1 and 2 in State C were the exception. Because all the other analyses were at the organization level, we mostly use that terminology throughout the rest of this report.

[8] We made these modifications with the goal of increasing sample size for the measures, which we expected to be potentially very small at the clinic or even organization level.

We calculated all five measures successfully for almost all organizations in States A, B, and C, except for a few organizations in one or more years that had numerator or denominator counts for a particular measure that were too small.[9] **Tables 4-6** show the final calculated rates at the state level by organization for each behavioral health measure for calendar years 2017 (when possible),[10] 2018, and 2019.

[9] Counts of less than or equal to 11 were suppressed by our data provider.

[10] The three follow-up measures (Follow-up After Hospitalization for Mental Illness [FUH-AD], Follow-up After Emergency Department (ED) Visit for Mental Illness [FUM-AD], and Follow-up After ED Visit for Alcohol and Other Drug (AOD) Dependence [FUA-AD]) do not require a look-back period into the previous calendar year, so we were able to calculate them for all three years. The Antidepressant Medication Management, Adult Version (AMM-AD) and Initiation and Engagement of AOD Dependence Treatment (IET-AD) measures require a look-back period into the previous calendar year, so we were able to calculate them only for 2018 and 2019 (because calculating the 2017 rate would have required a look-back into 2016 data, to which we did not have access).

Table 4. State A Measure Rates, by Organization and Year										
Calculated rates	FUH-AD		FUM-AD		FUA-AD		AMM-AD		IET-AD[a]	
	7-day	30-day	7-day	30-day	7-day	30-day	Acute phase	Continuation phase	Initiation of treatment	Engagement in treatment
2017										
Organization 1	44.3	66.0	48.6	68.9	26.9	37.3	NA	NA	NA	NA
Organization 2	25.9	43.7	27.6	39.3	32.2	50.2	NA	NA	NA	NA
Organization 3	51.1	78.8	42.1	59.2	25.5*	39.2*	NA	NA	NA	NA
2018										
Organization 1	49.3	72.4	51.9	71.1	27.5	41.2	71.5	49.0	60.5	30.9
Organization 2	28.0	44.7	31.3	43.0	35.1	46.7	55.9	31.7	73.7	53.0
Organization 3	55.0	78.5	**	25.0	**	**	61.7	42.2	64.8	22.8
2019										
Organization 1	51.3	71.4	47.8	68.3	30.4	40.9	70.2	48.0	56.7	27.3
Organization 2	27.3	47.6	31.6	40.6	36.0	50.8	55.4	33.2	71.4	50.9
Organization 3	56.8	79.6	**	16.5	22.6*	22.6	56.1*	41.5	63.6	23.6

Source: Mathematica's analysis of 2017-2019 TAF data.

Notes: Rate = NA (not applicable) for AMM-AD and IET-AD in 2017 because the measure calculation required a look back into the previous year, and we did not have access to 2016 data. Red font and * indicates the calculated measure rate had a SNR less than 0.7, suggesting the rate does not meet the conventional reliability threshold of 0.7. Please reference Section III and Appendix B for a more detailed explanation of SNRs and measure reliability.

[a] Total AOD abuse or dependence cohort rate.

** Indicates the rate could not be calculated because the numerator or denominator was less than or equal to 11, and therefore suppressed by our data provider.

Table 5. State B Measure Rates, by Organization and Year										
Calculated rates	FUH-AD		FUM-AD		FUA-AD		AMM-AD		IET-AD[a]	
	7-day	30-day	7-day	30-day	7-day	30-day	Acute phase	Continuation phase	Initiation of treatment	Engagement in treatment
2017										
Organization 1	52.7	78.0	73.5	81.5	35.2*	40.7*	NA	NA	NA	NA
Organization 2	53.8	83.1	48.7	62.0	33.0	47.0	NA	NA	NA	NA
Organization 3	67.0	86.1	54.3	75.0	23.3	31.7	NA	NA	NA	NA
Organization 4	53.1	73.1	41.9	55.6	41.6	55.8	NA	NA	NA	NA
Organization 5	51.2	79.5	42.5	63.8	27.3	36.0	NA	NA	NA	NA
Organization 6	62.6	84.8	44.3	68.4	37.9*	48.3*	NA	NA	NA	NA
Organization 7	32.2	57.9	44.2	58.8	48.2	61.8	NA	NA	NA	NA
Organization 8	**	**	**	**	41.8*	54.5*	NA	NA	NA	NA
2018										
Organization 1	54.7	80.0	61.9	80.4	31.9*	41.7	61.2	38.8	65.0	38.7
Organization 2	55.4	80.5	44.6	62.3	33.6	48.8	66.5	44.8	60.8	38.1
Organization 3	70.4	91.4	50.5	74.0	26.3*	36.3	64.0	39.3	71.4	26.9
Organization 4	54.2	77.1	45.4	58.4	42.0	58.8	67.3	43.8	75.1	48.2
Organization 5	60.8	84.2	40.8	62.6	27.9	36.3	64.2	41.3	62.0	37.8
Organization 6	59.4	91.0	54.1*	67.3	**	**	71.8	48.7	73.3	38.4
Organization 7	43.1	63.6	38.8	52.4	43.7	59.8	64.1	45.0	76.2	59.6
Organization 8	**	**	36.7*	40.0*	35.2*	55.6	67.1*	34.2	84.1	61.0
2019										
Organization 1	48.7	74.3	69.7	80.0	34.3	47.1	62.4	38.3	69.0	41.9
Organization 2	57.1	81.6	45.3	63.0	39.8	54.5	68.6	44.9	60.3	38.6
Organization 3	69.0	90.1	56.1	78.2	20.3	29.0	60.6	34.6	66.7	26.7
Organization 4	56.3	71.6	44.4	57.9	42.9	56.9	69.5	46.1	73.0	46.8
Organization 5	59.3	77.4	39.5	59.9	23.2	33.0	64.3	40.2	61.3	38.0
Organization 6	54.2	86.5	42.1	68.4	39.4*	39.4*	67.0	43.6	70.9	42.6
Organization 7	33.2	56.5	34.4	51.4	46.9	58.0	65.9	44.6	76.5	58.4
Organization 8	31.1	48.9	31.4*	45.7	43.6*	55.1	70.5*	36.8	83.6	64.4

Source: Mathematica's analysis of 2017-2019 TAF data.

Notes: Rate = NA (not applicable) for AMM-AD and IET-AD in 2017 because the measure calculation required a look back into the previous year, and we did not have access to 2016 data. Red font and * indicates the calculated measure rate had a SNR less than 0.7, suggesting the rate is not reliable. Please reference Section III and Appendix B for a more detailed explanation of SNRs and measure reliability.

[a] Total AOD abuse or dependence cohort rate.

** Indicates the rate could not be calculated because the numerator or denominator was less than or equal to 11, and therefore suppressed by our data provider.

Table 6. State C Measure Rates, by Organization and Year										
Calculated rates	FUH-AD		FUM-AD		FUA-AD		AMM-AD		IET-AD[a]	
	7-day	30-day	7-day	30-day	7-day	30-day	Acute phase	Continuation phase	Initiation of treatment	Engagement in treatment
2017										
Clinic 1	53.8	83.3	55.0	73.8	23.9*	26.1	NA	NA	NA	NA
Clinic 2	30.8	62.3	46.2	60.0	**	**	NA	NA	NA	NA
Organization 1	31.6	57.3	52.5	72.9	**	**	NA	NA	NA	NA
Organization 2	**	**	**	**	**	**	NA	NA	NA	NA
2018										
Clinic 1	51.3	77.8	48.8*	70.9	**	16.9	47.3	24.6	52.0	13.2
Clinic 2	43.4	71.7	38.0*	58.0	**	**	55.9	26.7	51.0	16.3
Organization 1	37.5	64.2	54.3*	67.4	**	**	58.8	29.1	56.4	22.7
Organization 2	**	**	**	**	**	**	62.1*	34.5	36.5	NA
2019										
Clinic 1	51.6	72.7	55.0	73.0	15.8	23.7	48.2	24.0	53.0	16.7
Clinic 2	30.0	51.4	50.0*	66.0	**	**	56.5	25.3	56.9	17.8
Organization 1	33.3	61.3	61.5*	78.8	**	**	58.5	32.0	51.9	18.6
Organization 2	54.5*	72.7*	52.4*	81.0	**	**	56.5	29.8	41.1	18.9

Source: Mathematica's analysis of 2017-2019 TAF data.

Notes: Rate = NA (not applicable) for AMM-AD and IET-AD in 2017 because the measure calculation required a look back into the previous year, and we did not have access to 2016 data. Red font and * indicate the calculated measure rate had a SNR less than 0.7, suggesting the rate is not reliable. Please reference Section III and Appendix B for a more detailed explanation of SNRs and measure reliability.

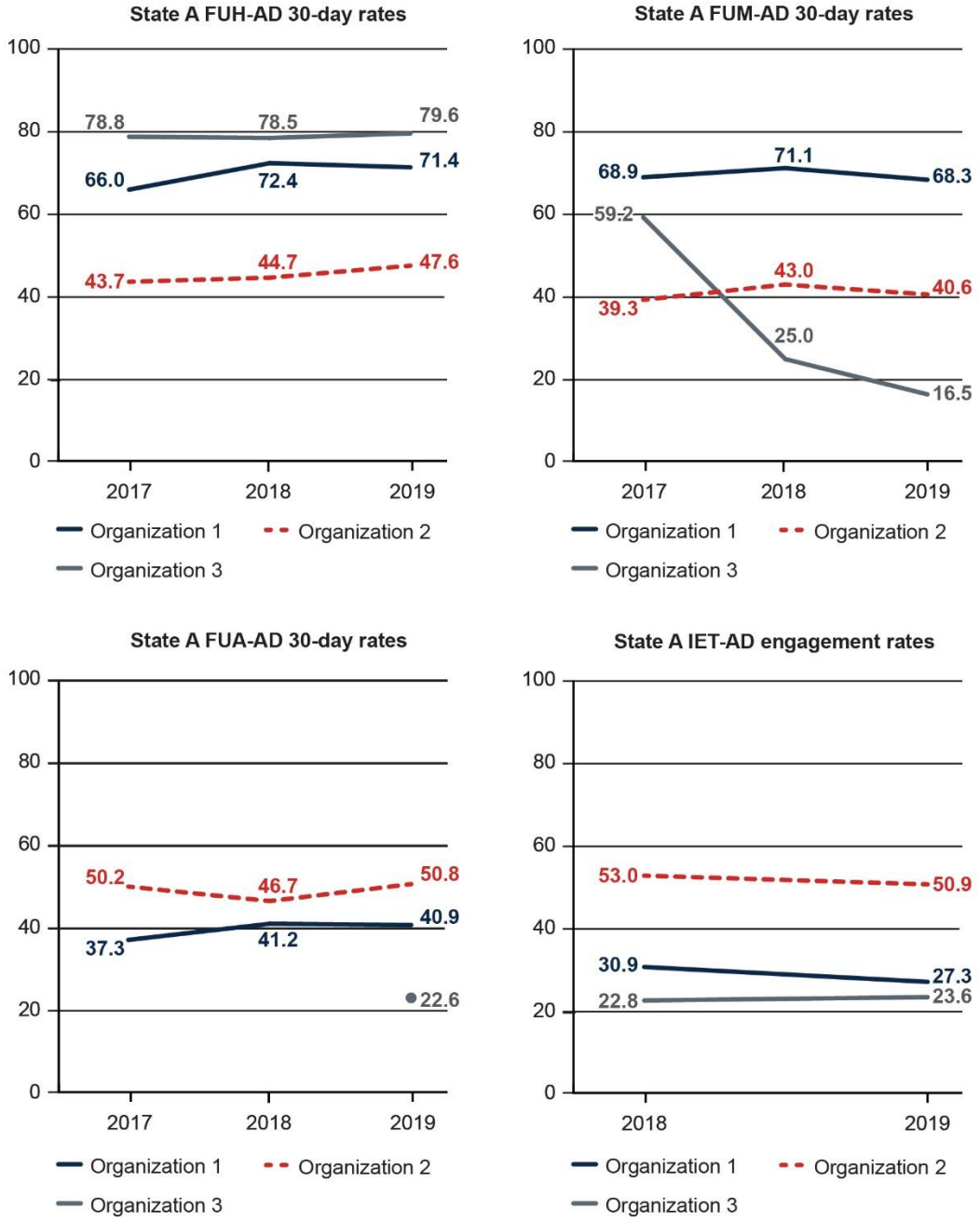
[a] Total AOD abuse or dependence cohort rate.

** Indicates the rate could not be calculated because the numerator or denominator was less than or equal to 11, and therefore suppressed by our data provider.

To explore the possibility of changes in performance over time around the implementation of a behavioral health program, we compared organization performance on the measures during the period before SAMHSA awarded the CCBHC-E grants (that is, 2017 [when available, depending on the measure] and 2018) to the period after CCBHC-E grant implementation (2019). Although we might expect small variations in the calculated rates across the pre-period years (due to real changes in care or patient populations), large changes in the results from one year to the next, especially when combined with a small denominator size, might reflect statistical noise or data quality changes that could influence performance on the measure and threaten its reliability at the clinic level. The latter is largely what we found; the data are mostly inconclusive and do not tell a clear story.

However, we did observe that trends in measure performance suggested some organizations performed consistently better on SMI measures than SUD measures, and vice versa. For example, in State A, Organization 2 was the lowest performer on the SMI-related measures (FUM-AD and FUH-AD) but was the highest performer on the SUD-related measures (FUA-AD and IET-AD). **Figure 1**, which compares Organization 2's performance on these measures to the other organizations in State A, demonstrates this pattern. We observed the same pattern for one organization in State B as well. We were unable to observe any similar comparisons for State C because small sample sizes for most clinics and organizations in this state made several of these measures incalculable at the clinic or organization level.

Figure 1. Comparison of Organization 2's performance on SMI-Related Measures (FUH-AD, FUM-AD) and SUD-Related Measures (FUA-AD, IET-AD) Against Other organizations for State A



Source: Mathematica's analysis of 2017-2019 TAF data.

Notes: Measure rates are not shown (e.g., in the bottom right panel Organization 3) if they had a SNR less than 0.7, or if they could not be calculated due to small numerators or denominators.

III. Reliability and Validity of Behavioral Health Quality Measures

After calculating the behavioral health quality measures at the organization level, we assessed the reliability (in terms of precision and stability of the measure scores) and validity of the calculated results. We framed this analysis according to two different perspectives. From a state perspective, we considered whether there was potential utility in monitoring CCBHC-E organization performance using these behavioral health measures. If so, we looked at whether some measures yielded more reliable and valid results than others. From a clinic perspective, we considered if, and how, clinics could use the results from these particular measures to manage their behavioral health service programs more effectively.

Measure Reliability

Measure reliability includes both precision and stability. To determine whether the calculated measure rates were **precise**, we computed signal-to-noise ratios (SNRs) for each calculated measure rate for each organization and year. SNR analysis is a method of reliability testing based on calculating variability within and among providers, thereby determining differences in performance across them. The *signal* is the proportion of variability in measured performance explained by real differences in performance; the *noise* relates to the total variability in measured performance due to chance or measurement errors (American Society of Clinical Oncology 2021). Noise can be introduced by beneficiary-level variability, which can include unmeasured beneficiary characteristics, or by the lack of precision in the measure estimates due to lack of sufficient beneficiary sample size within providers (Deutsch et al. 2012).

Although there is no clear cutoff for minimum reliability level, researchers often consider a reliability rate of 0.4 to be the lower limit of moderate reliability sufficient for public reporting (Schone et al. 2011), reliability above 0.7 is considered sufficient to see differences between providers and the mean, and reliability above 0.9 is considered sufficient to see differences between any provider pair (NQF 2013). For the purposes of this analysis, if a calculated SNR was greater than or equal to 0.7 for a measure-organization-year, we considered the measure rate reliable. If a calculated SNR was less than 0.7, we considered the measure rate unreliable and do not recommend using that rate for any performance monitoring purposes.[11]

Tables 4-6 present organization-level SNR results**.** Most measures for each organization and year met or exceeded the minimum SNR of 0.7. In particular, for the FUH-AD, AMM-AD, and IET-AD measures, almost all calculated rates for each organization and year had an SNR of at least 0.7; this was the case even for State C, which had smaller clinic or organization sample sizes than the other states. The FUM-AD and FUA-AD measures had a greater number of unreliable calculated rates, which we discuss later. **Table 7** presents mean SNRs by state for each measure and year, and largely confirms these findings. Mean SNRs for each state are consistently above 0.7 for all measures and years except for FUM-AD 7-day for State C, and FUA-AD 7-day for States B and C.

[11] The National Quality Forum (NQF) recently lowered its minimum reliability cutoff from 0.7 to 0.5 and has historically considered 0.5-0.69 as borderline acceptable (for example, Glance et al. 2021). We continue to use 0.7 as our reliability cutoff for this analysis because it is the well-established practice. As shown in **Table 7**, all of the mean SNRs in our analysis met or exceeded NQF's new 0.5 minimum reliability cutoff.

Table 7. Mean SNRs Across Organizations, by State and Year				
Measure	**Year**	**State A mean SNR**	**State B mean SNR**	**State C mean SNR**
FUH-AD 7-day	2017	0.94	0.91	0.87
	2018	0.93	0.90	0.79
	2019	0.93	0.90	0.76
FUH-AD 30-day	2017	0.95	0.94	0.87
	2018	0.95	0.95	0.85
	2019	0.95	0.92	0.78
FUM-AD 7-day	2017	0.81	0.85	0.68*
	2018	0.79	0.72	0.50*
	2019	0.88	0.83	0.61*
FUM-AD 30-day	2017	0.81	0.86	0.72
	2018	0.91	0.89	0.79
	2019	0.95	0.92	0.84
FUA-AD 7-day	2017	0.70	0.67*	0.65*
	2018	0.74	0.62*	NA
	2019	0.81	0.77	0.79
FUA-AD 30-day	2017	0.78	0.77	0.77
	2018	0.92	0.85	0.86
	2019	0.87	0.83	0.82
AMM-AD acute phase	2018	0.87	0.86	0.79
	2019	0.85	0.87	0.82
AMM-AD continuation phase	2018	0.91	0.89	0.85
	2019	0.89	0.90	0.89
IET-AD[a] initiation	2018	0.95	0.96	0.87
	2019	0.94	0.95	0.87
IET-AD[a] engagement	2018	0.97	0.97	0.97
	2019	0.97	0.97	0.96

Source: Mathematica's analysis of 2017–2019 TAF data.

Notes: Red font and * indicates the mean SNR is less than 0.7, suggesting the measure should not be considered reliable for that state and year.

Means are calculated using all available organization SNRs for each measure and year within each state. Not all organizations had adequate sample size to calculate a measure rate (and therefore an SNR) for each year, so not all organizations are included in each state mean SNR.

NA = not available because no measure rates were calculable in this state and year due to small sample size.

[a] Total AOD abuse or dependence cohort rate.

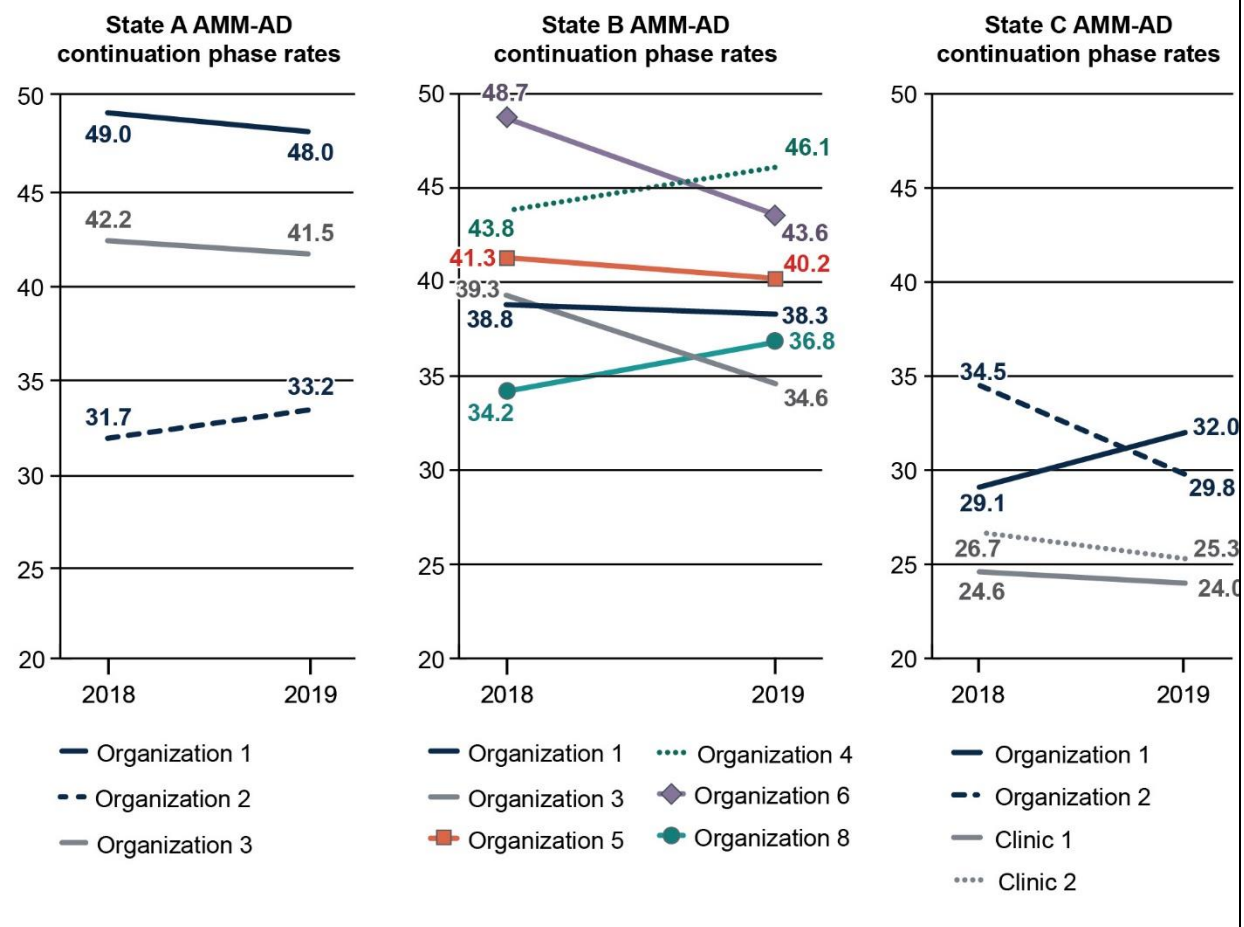
Although our calculated measure rates were mostly reliable at both the organization and state mean levels, there were a few general exceptions:

- State C was the only state where we captured only CCBHC-E beneficiaries in the measure calculations. Therefore, State C had smaller clinic and organization-level sample sizes than the other two states. This resulted in a greater number of low SNRs (and therefore unreliable

calculated rates) due to insufficient sample sizes for several measures compared to organizations in the other states.

- For all the follow-up measures (FUH-AD, FUM-AD, and FUA-AD), the 7-day follow-up rate had a greater number of unreliable calculated rates than the 30-day follow-up rate. We expected this because the 7-day rate has a smaller sample size by nature of the shorter time period, resulting in more calculated 7-day rates being found unreliable.
- Lastly, one or more organizations in all three states and in most years had FUA-AD calculated measure rates in which the SNR was less than 0.7, which was attributable across the board to small sample sizes for this measure.

Figure 2. Comparison of Stability in Organizations' AMM-AD Continuation Phase Rates Over Time, by State



Source: Mathematica's analysis of 2017–2019 TAF data.

Notes: We excluded Organizations 2 and 7 from the State B panel for readability. Measure rates for these organizations can be found in Table 5; they follow the same general pattern as the other State B organizations.

Overall, four of the five measures we calculated (all measures except FUA-AD) were for the most part reliable at the organization level for all states. For performance monitoring purposes, we therefore recommend excluding the FUA-AD measure completely and the 7-day rate for the other follow-up

measures, because these were the two areas with a higher frequency of low sample sizes, resulting in low SNRs and therefore more unreliable calculated rates. We would expect this to be the case for most behavioral health organizations and especially clinics across the nation.

To assess measure **stability**, we visually examined how organizations' performance on the measures changed over time. The calculated measure rates were relatively stable over time across organizations in States A and C (for the measures and years with adequate sample size to calculate the organization-level measure rates). Organizations in State B showed more variation in measure rates over time. For an illustrative example, **Figure 2** shows trends in the AMM-AD continuation phase rates for organizations in each state over time. Note the greater magnitude of change in rates from year to year among organizations in State B compared to organizations in the other two states.

Measure Validity

To assess measure validity, we compared our calculated behavioral health quality measure state rates[12] to two benchmarks: Medicaid Core Set rates[13] for the five measures from the comparable FFY and the CCBHC demonstration quality measure[14] rates for the five measures from the comparable demonstration year. **Appendix A** shows how time periods of the benchmark sources align with each calendar year of our calculated state rates. Our measures differed from the two benchmarks in the following ways:

- Medicaid Core Set measures:
  - We did not require beneficiaries to have continuous Medicaid eligibility to include them in the eligible population. All the Core Set measures required continuous eligibility during at least some of the time period covered by the measure.
  - We did not exclude beneficiaries in hospice. All the Core Set measures excluded beneficiaries in hospice from the eligible population.[15]
  - Our eligible beneficiary population for the measures, comprising beneficiaries who received services from CCBHC-E clinics or organizations, is quite different from the Core Set measures, which can include all Medicaid beneficiaries in the state.
- CCBHC demonstration quality measures:
  - Our eligible beneficiary population for the measures also differs from the CCBHC demonstration quality measures, which are limited to only beneficiaries who received services from CCBHCs.

[12] We calculated our state rates by aggregating the attributed beneficiaries from all included CCBHC-E organizations in a particular state into a single pool, and then calculating the measure from that CCBHC-E state-level beneficiary pool.
[13] Available at https://www.medicaid.gov/medicaid/quality-of-care/performance-measurement/adult-and-child-health-care-quality-measures/adult-health-care-quality-measures/index.html. States report Medicaid Core Set measures annually and CMS uses them to monitor the quality of health care received by Medicaid beneficiaries.
[14] The CCBHC quality measures are a set of measures that CCBHCs or their states must collect and report under the demonstration. Of the five measures (each with two different rates) that we calculated, the CCBHC quality measures include FUH-AD 30-day, FUM-AD 30-day, FUA-AD 30-day, AMM-AD acute and continuation phases, and IET-AD initiation and engagement. More information is available at https://www.samhsa.gov/section-223/quality-measures.
[15] We made these two modifications with the goal of increasing sample size for the measures, which we expected to be potentially very small at the clinic and even organization level.

- Although we know the CCBHC demonstration quality measures were based on the same Medicaid Core Set measure technical specifications we used to specify the measures (although probably from FFY 2017; we used FFY 2020), we do not know what decisions the states made when implementing the measure technical specifications that might differ from ours. For example, it is possible (though probably unlikely) that the CCBHC demonstration quality measures are limited to care *provided only by the CCBHCs*. For our measures, we attributed beneficiaries to the CCBHC-Es to use as the base eligible measure population, but then included in the measure *care provided anywhere*, not limited to only CCBHC-E-provided care.

Given the differences between our calculated measures and the two benchmarks, comparisons with the benchmarks for the purpose of testing measure validity were less useful than we had hoped.[16] Our calculated state rates were substantially higher than the Medicaid Core Set state rate benchmarks for all five measures. We expected this because the Medicaid Core Set rates include all Medicaid beneficiaries, including those with less severe behavioral health conditions than those typically served by CCBHC-Es.

On the other hand, no clear pattern emerged when comparing our calculated state rates with the CCBHC demonstration quality measure benchmark rates. Our calculated state rates were generally lower than or similar to the CCBHC demonstration quality measure rates for the follow-up measures (FUH-AD, FUM-AD, and FUA-AD). This is likely because of the newness of the CCBHC-E grants; unless the CCBHC-E recipient was also a demonstration CCBHC, it had less time to develop the systems and practices that could contribute to higher performance on these measures, potentially leading to slightly lower performance. The other two measures--AMM-AD and IET-AD--did not follow this pattern. For these two measures, our calculated state rates were higher than the CCBHC quality measures for the first-stage rates (acute phase for AMM-AD and initiation for IET-AD), but the comparisons varied by state for the second-stage rates (continuation phase for AMM-AD and engagement for IET-AD). We conclude that validity testing for these two measures is noisy but does not suggest any obvious issues with our measure calculations. AMM-AD and IET-AD are extremely complex measures; the calculated rates therefore likely vary significantly based on the time period and beneficiary population used to calculate them, as well as the analytic decisions made when implementing the measures from the technical specifications into the data source.[17]

Overall, we conclude from the measure reliability and validity testing that our calculated rates at the clinic or organization level are for the most part reliable and correspond roughly as expected with relevant benchmarks, although comparisons with the CCBHC quality measures benchmarks were noisier than expected. Therefore, we believe that states or other funders could potentially use these types of measure results for evaluation activities such as identifying high and low-performing organizations within particular clinical areas (for example, SMI medication management or SUD treatment and follow-up), or for monitoring performance over time. However, the technical complexity of specifying these measures

[16] Unfortunately, we cannot show the state rates and benchmark comparisons graphically in this report because the benchmark rates are publicly available and doing so would compromise the confidentiality of our three states (and therefore their associated 2018 CCBHC-E grantees).

[17] In addition, the patterns we see when comparing our AMM-AD and IET-AD rates to the Adult Core Set state rates and the CCBHC quality measure state rates are similar across both measures, perhaps suggesting convergent validity between these two measures. As an additional caveat, however, the FFY 2020 IET-AD measure specifications used for this analysis include telehealth-provided services, which were not included in the specifications used for the CCBHC quality measures. We do not have a good sense for the prevalence of these kind of minor differences between specifications, but they certainly could lead to some of the differences in rates we see here.

and the delay between clinics providing services and the eventual impact on performance on these annual measures likely makes them of limited use to clinics for managing their programs more effectively or for more immediate, real-time monitoring purposes.

IV. Conclusions and Future Applications

Overall, the measure calculation and analysis feasibility testing process provided noteworthy findings with useful applications for future work in the behavioral health space. We conclude, based on the challenges in obtaining clinic IDs, that identifying individual behavioral health clinics in the TAF data will likely require state-specific approaches, which is extremely time intensive. Identifying individual clinics using the methods from this analysis is largely not feasible, but we found *organization-level* analysis to be possible in all states in our sample, a finding that has broad, positive implications for monitoring and evaluating behavioral health programs that operate at the organization level.

In addition, our experiences attempting to identify CCBHC-Es in these five states sheds light on the unique challenge that CCBHC-Es represent: because SAMHSA provided the 2018 CCBHC-E grants directly to behavioral health organizations, rather than to states, as has been the case for many other behavioral health programs, some states had little to no involvement in tracking and monitoring the CCBHC-E grantees. Unsurprisingly, identifying individual CCBHC-E clinics was easier in the three states that had direct involvement with CCBHC-Es. Expanding our view beyond just the CCBHC-E grant program, the process of identifying individual clinics might be easier for behavioral health programs where states are directly involved with the programs' administration and/or monitoring.

We also have to consider the strengths and limitations of the data source when assessing the feasibility of identifying individual clinics in the data. The TAF data present unique strengths and challenges when attempting to identify CCBHC-Es. The TAF are standardized across states and are, for the most part, clean and well populated. This makes the measure calculation step easier than using state-provided Medicaid data, which varies widely by state in terms of format and completeness. On the other hand, because the TAF is a standardized federal data source, it has a limited set of ID fields available in which to find clinic IDs. For our analysis, several states used alternative identification methods that were present in their Medicaid data but did not filter up to the TAF. For this reason, we strongly encourage states to create simple, straightforward methods of identifying their behavioral health clinics, and requiring clinics to use those IDs when billing Medicaid. For the one in five states in our sample that already did this, identifying CCBHC-Es in that state was simple. Additionally, the variability in identification methods among the selected states in this study suggests that more specific guidance from CMS on which provider IDs to include on claims would improve the usability of the TAF for research and evaluation purposes.

This analysis showed that these five behavioral health quality measures, based largely on Medicaid Core Set measure specifications, are complex to construct. State leadership could, perhaps, use these behavioral health quality measures to identify low or high-performing clinics or organizations for evaluation activities, but the efficacy of doing so would likely vary by state. For states similar to States A and C, where the organizational trends were mostly steady over time, tracking performance on these quality measures might be useful; perhaps less so in states like State B, where clinic and organizational trends varied more over time. The complexity of calculating these measures and the delay between clinics providing services and the resulting change in performance on these annual measures likely makes them of little use to clinics for performance monitoring or timely clinic quality improvement efforts.

Finally, our measure validity analysis highlighted the unexpected impacts of seemingly minor differences in measure specification, even among well-established behavioral health quality measures such as these. Translating these types of measures from high-level specifications into actual implementation is a complex process that requires highly specific decision making that could be challenging to implement in a standardized way across multiple states. Therefore, there is likely limited utility in comparing clinic or organization-level performance to benchmarks, in terms of both validating results and monitoring progress. This is particularly true if updates to the clinic or organization measure specifications do not occur at the same frequency as benchmark measure specifications updates. Instead, monitoring clinics' or organizations' progress by limiting comparisons to only measure rates over time *for that same clinic or organization* is probably the more effective strategy.

Overall, this analysis produced valuable information regarding the feasibility of calculating behavioral health quality measures at the clinic level using Medicaid data. Although we encountered significant challenges in identifying clinics and could calculate the measures at the clinic level for only two clinics in one state, we identified all the CCBHC-E organizations in all states, and we calculated all five measures largely successfully for those organizations. This analysis demonstrated the potential utility of monitoring behavioral health organizations using current federal Medicaid data, with some caveats around the complexity of the measures and the process of implementing the specifications to the data source. We suspect that as models like CCBHC continue to expand to more states, states will continue to develop new and better ways to identify their behavioral health clinics, and the value of monitoring and evaluating behavioral health clinics and organizations will only increase.

References

American Society of Clinical Oncology. "ASCO Measures Methodology Manual." American Society of Clinical Oncology, September 9, 2021. Available at https://www.asco.org/sites/new-www.asco.org/files/content-files/advocacy-and-policy/documents/2020-Measures-Methodology-Manual.pdf. Accessed April 25, 2022.

Breslau, J., B. Briscombe, M. Dunbar, C. Kase, J. Brown, A. Wishon Siegwarth, and R. Miller. "Interim Cost and Quality Findings from the National Evaluation of the Certified Community Behavioral Health Clinic Demonstration." Mathematica, October 16, 2020. Available at https://aspe.hhs.gov/reports/interim-ccbhc-cost-quality-findings. Accessed March 1, 2022.

Deutsch, A., L. Smith, B. Gage, C. Kelleher, and D. Garfinkel. "Patient-Reported Outcomes in Performance Measurement." National Quality Forum, 2012.

Glance, L.G., D.R. Nerenz, and K.E. Joynt Maddox. "Reproducibility of Hospital Rankings Based on Centers for Medicare & Medicaid Services Hospital Compare Measures as a Function of Measure Reliability." *JAMA Network Open*, vol. 4, no. 12, 2021, p. e2137647. doi:10.1001/jamanetworkopen.2021.37647. Accessed May 19, 2022.

National Quality Forum (NQF). "Review and Update of Guidance for Evaluating Evidence and Measure Testing- Technical Report." National Quality Forum, October 2013. Available at https://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=74076. Accessed April 25, 2022.

Schone, E., M. Hubbard, and D. Jones. "Reporting Period and Reliability of AHRQ, CMS 30-day and HAC Quality Measures--Revised." Mathematica Policy Research, November 18, 2011. Available at https://econpapers.repec.org/scripts/redir.pf?u=https%3A%2F%2Fwww.mathematica.org%2F-%2Fmedia%2Fpublications%2Fpdfs%2Fhealth%2Fhvbp_measure_reliability.pdf;h=repec:mpr:mprres:cab712bf5e324d0db15eca9c404f3eb2. Accessed April 25, 2022.

APPENDIX A
Mapping of Calculated State Rates and Benchmark Sources

To assess measure validity, Mathematica compared our calculated state rates for behavioral health quality measures to two benchmarks: Medicaid Core Set rates[18] for the five measures from the comparable FFY and the CCBHC demonstration quality measure[19] rates for the five measures from the comparable demonstration year. Demonstration years and FFYs do not align exactly with the calendar years we used for our calculated measures, so we matched the timeframes as seemed reasonable for the purpose of comparing performance on the five behavioral health quality measures as shown in **Table A.1**.

Table A.1. Mapping of Calculated State Rates and Benchmark Sources, by Year

Our calculated measures	Medicaid Core Set measures	CCBHC demonstration quality measures
Calendar Year 2017	FFY 2017	NA
Calendar Year 2018	FFY 2018	CCBHC DY1
Calendar Year 2019	FFY 2019	CCBHC DY2

[18] Available at https://www.medicaid.gov/medicaid/quality-of-care/performance-measurement/adult-and-child-health-care-quality-measures/adult-health-care-quality-measures/index.html. States report Medicaid Core Set measures annually and CMS uses them to monitor the quality of health care received by Medicaid beneficiaries.

[19] The CCBHC quality measures are a set of measures CCBHCs or their states must collect and report under the demonstration. Of the five measures (each with two different rates) we calculated, the CCBHC quality measures include FAH-AD 30-day, FUH-AD 30-day, FUA-AD 30-day, AMM-AD acute and continuation phases, and IET-AD initiation and engagement. More information is available at https://www.samhsa.gov/section-223/quality-measures.

APPENDIX B
Reliability Testing Technical Documentation

This appendix describes Mathematica's reliability testing methods and process. As outlined in Section 3: Reliability and Validity of Behavioral Health Quality Measures, to determine whether the calculated measure rates were reliable, we calculated SNRs for each calculated measure rate for each organization and year.

SNR analysis is a method of reliability testing based on calculating variability within and among providers, thereby determining differences in performance across them. The signal is the proportion of variability in measured performance that real differences in performance can explain; the noise relates to the total variability in measured performance usually due to chance or measurement errors (American Society of Clinical Oncology 2021). Although there is not a clear cutoff for minimum reliability level, researchers typically consider a reliability rate of 0.4 to be the lower limit of moderate reliability sufficient for public reporting (Schone et al. 2011), reliability above 0.7 is considered sufficient to see differences between providers and the mean, and reliability above 0.9 is considered sufficient to see differences between any provider pair (NQF 2013). For the purposes of this analysis, if a calculated SNR was greater than or equal to 0.7, we considered the measure rate reliable.[20]

We tested reliability by calculating the SNR at the clinic or organization level for each measure and year within each state to see if it was greater than or equal to 0.7. The equations to calculate reliability are as follows:

$$\sigma^2 \; within \; clinicians = \frac{p(1-p)}{n}$$

$$\sigma^2 \; between \; clinicians = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$$

$$Reliability = \frac{\sigma^2 \; between \; clinicians}{\sigma^2 \; between \; clinicians + \sigma^2 \; within \; clinicians}$$

Where σ^2 equals the variance between clinics or organizations within a state, *p* equals the measure rate for each clinic or organization and year, *n* equals the number of events or beneficiaries included in the measure denominator (that is, the sample size), and α and β are the parameters that describe the shape of the fitted beta distribution.

Reliability is therefore based on both the within-organization variance (the "noise"), and the between-organization variance (the "signal"), both of which depend on the measure rate and sample size for each organization within a state, by year. Reliability is calculated as the ratio of the variance between clinics or organizations and the total variance of the measure rates within a state.

[20] The NQF recently lowered its minimum reliability cutoff from 0.7 to 0.5 and has historically considered 0.5-0.69 as borderline acceptable (for example, Glance et al. 2021). We continue to use 0.7 as our reliability cutoff for this analysis because it is the well-established practice.

For the most part, larger sample sizes lead to more reliable measure rates. However, measure rates with larger variance between clinics or organizations within a state can also lead to reliable measure rates despite smaller sample sizes because the larger variation increases our ability to detect differences between clinics or organizations.