Findings from the Alternative Response Pilot: 2-Year Outcomes

Evaluation Report

REPORT HIGHLIGHTS:

- Senate Bill (SB) 21-118 **authorized the Alternative Response (AR) pilot** within Adult Protective Services (APS).
- SB21-118 **requires an independent evaluation** to build evidence for the AR practice and inform the pilot's future.
- This evaluation report summarizes **pilot reach, implementation,** and **impact data.**
- A **quasi-experimental design** was used to generate causal evidence on the **effectiveness of AR** for a 2-year pilot period (January 4, 2023 through December 31, 2024; n = 9,790 unique AR Pilot cases).
- Findings illustrate the **AR Pilot is having a positive impact on at-risk adults** in Colorado by reducing repeat involvement and case length through collaborative engagement.
- Based on outcome findings, the AR practice **is recommended for statewide scaling** with adequate resourcing for strong implementation.

AUTHORS:

Courtney L. Everson, PhD
Sr. Project Director/Principal Investigator
Colorado Evaluation and Action Lab

Ernest Boffy-Ramirez, PhD
Sr. Researcher/Project Director
Colorado Evaluation and Action Lab

Erin Wickerham, MPH
Sr. Researcher
Colorado Evaluation and Action Lab

For inquiries contact: Courtney L. Everson | admin@coloradolab.org | ColoradoLab.org

Report Number: 21-09E. Date: June 2025. Revised November 2025.

Abstract

Building Evidence for the Alternative Response Pilot

Senate Bill (SB) 21-118 authorized an Alternative Response (AR) pilot within the Colorado Department of Human Services (CDHS), Adult Protective Services (APS). This pilot creates a dual-track model for tailoring APS response to risk level. Allegations of low-risk mistreatment and self-neglect are tracked to AR and higher-risk allegations are tracked to Traditional Response. The Colorado Evaluation and Action Lab at the University of Denver served as the independent evaluator for the legislatively required outcomes study. The goal of the study was to understand the effectiveness of AR and inform the future of this practice model.

This evaluation report highlights reach, implementation, and impact outcomes from a 2-year period (January 4, 2023, through December 31, 2024). Results show the AR practice can improve collaborative engagement between APS staff and clients, which can help stabilize the client and improve well-being. Compared to equivalent cases in the pre-pilot period, repeat involvement in APS was significantly decreased by 2.5%. Case length was also reduced by 5.63 days. Descriptive data show the AR Pilot is having a strong reach in pilot counties, is especially needed for clients experiencing self-neglect, and is helping to support individuals with higher levels of social isolation and vulnerable conditions. Results also show the critical importance of building support networks to improve long-term well-being and maximize effectiveness of the AR practice.

Taken together, findings indicate the AR practice is a person-centered approach that can inform best practices for supporting at-risk adults, including a growing aging population. Based on findings, the dual-track model created by the pilot should be considered by CDHS for statewide scaling. Recommendations for statewide expansion include: a) prioritize rule changes that are responsive to data-informed learnings, such as revisiting the timeline for initial response; b) enable a phased rollout statewide over a period of time to ensure county readiness and to provide the state preparation time; c) provide adequate resourcing at state and county levels to ensure fidelity of implementation; d) advance partnerships within CDHS and across systems in caring for the aging population; and e) apply Colorado's Evidence-Based Decision Making approach in state government to activate results of the 2-year rigorous evaluation—*in commitment to achieving positive outcomes and smart state investments among the APS program and clients served.*

Table of Contents

Acknowledgements

This research was supported by the Colorado Department of Human Services (CDHS), Division of Aging and Adult Protective Services. The opinions expressed are those of the authors and do not represent the views of the State of Colorado, CDHS, or the University of Denver. Policy and budget recommendations do not represent the budget or legislative agendas of state agencies, the Governor's Office, or other partners. Any requests for funding or statutory changes will be developed in collaboration with the Governor's Office and communicated to the legislature through the regular budget and legislative processes.

Thank you to our partners who provided subject matter expertise and guidance on this project:

- CDHS leadership and staff: Mindy Gates, Stefanie Woodard, Bettina Morrow, Elena Romero, Rose Green, Erica Felder, and the Administrative Review Division team;
- County representatives from the 15 pilot counties;
- Caseworkers, supervisors, and leaders of pilot counties that worked tirelessly to design, implement, and strengthen the Alternative Response innovation for at-risk adults.

Data Sources

The study uses data from three sources:

- Colorado Adult Protective Services (CAPS) administrative data system.
- Fidelity of Implementation measures, using CAPS data.
- Qualitative data collected through focus groups, surveys, and pilot county feedback.

Suggested Citation

Everson, C.L., Boffy-Ramirez, E., & Wickerham, E. (June 2025). *Findings from the Alternative Responsive Pilot: 2-year outcomes* (Evaluation Report). (Report No. 21-09E). Denver, CO: Colorado Evaluation and Action Lab at the University of Denver.

Introduction

State Bill (SB) 21-118 requires a 2-year outcomes evaluation to assess effectiveness of the Alternative Response (AR) Pilot and to inform the future of the dual-track model.

This evaluation report summarizes 2-year data on pilot reach and implementation, communicates causal evidence on effectiveness of AR, and makes policy and practice recommendations.

SB21-118 (Alternative Response Mistreatment At-risk Adults) passed in the 2021 legislative session, authorizing a pilot of the Alternative Response (AR) practice for responding to reports of low-risk mistreatment or self-neglect of an at-risk adult. Current law allows for only one type of response for a county department of human services, regardless of the risk level reported.

"With AR, I noticed families are more willing to talk to you and they'll actually call now, saying 'Hey, this is what's going on. Do you have any ideas?'"

- Pilot County Caseworker

The AR Pilot enables a dual-track model to better tailor the response approach to the unique circumstances of the case and allegations. Track One is called Traditional Response (TR) and is reserved for higher-risk allegations of mistreatment; Track Two is called Alternative Response (AR) and is applied to all self-neglect allegations and lower-risk allegations of mistreatment. The AR practice opens the door for more collaborative engagement by establishing a strong partnership from case start to case end, as illustrated in the AR theory of change (Figure 1).

As an innovative practice in Colorado and nationally, SB21-118 requires a 2-year outcomes evaluation to assess effectiveness of the AR Pilot and inform the future of the dual-track model. The Colorado Department of Human Services (CDHS), Adult Protective Services (APS), partnered with the Colorado Evaluation and Action Lab (Colorado Lab) to fulfill this legislative opportunity.

This report summarizes 2-year data on pilot reach and implementation and provides causal evidence on effectiveness of AR at the person and system levels.

Participating Counties

SB21-118 authorized 15 counties to participate in the AR Pilot, with a requirement for a balance of rural/frontier and urban/suburban counties. Counties applied as interested in participating. The Colorado Lab used a method called random stratified sampling with weighting to select the 15 counties in a fair and balanced manner (Table 1).

Table 1. Participating Pilot Counties and Geographic Designation

Rural/Frontier Counties	Urban/Suburban Counties
Eagle	Adams
Garfield	Arapahoe
La Plata	Denver
Otero	El Paso
Prowers	Jefferson
Pitkin	Larimer
Routt	Mesa
---	Weld

Program Description

The Colorado APS program, located within CDHS, was established in statute in 1983 to provide protective services for vulnerable persons age 65 and older. The program was expanded in 1991 to the current statute, which establishes protective services for at-risk adults age 18 and older (Colorado Revised Statutes, Title 26, Article 3.1). The purpose is to intervene on behalf of an at-risk adult to correct or alleviate situations in which actual or imminent danger of abuse, caretaker neglect, exploitation, or harmful acts (collectively referred to as "mistreatment"), or self-neglect, exist.

Colorado APS is state-supervised and county-administered, such that county departments of human services are responsible for implementing the APS program, while the state is responsible for rule-making and oversight. APS is charged in statute with accepting reports of mistreatment and self-neglect of at-risk adults, investigating the allegations, assessing the client for other health and safety needs, and working with the client to implement protective services when appropriate. The APS program collaborates with law enforcement and/or district attorneys for criminal investigation and possible prosecution, as well as partners with community-based services, health care, family and friends, and other supports to promote safety and well-being. The APS guiding principles are consent, self-determination, and least restrictive intervention.

Colorado's Opportunity

The AR Pilot provides Colorado an opportunity to innovative approaches to adult protective services and inform best practices for a rapidly growing aging population. This includes contributing to Colorado's Multi-Sector Plan on Aging through learnings obtained though the 2-year evaluation.

Figure 1. Alternative Response Practice Theory of Change

Description of the Study

The evidence-building approach prioritizes data-informed learning alongside rigorous evaluation methods.

The study period is January 4, 2023, through December 31, 2024, inclusive of 2 years of implementing the dual-track model across 15 pilot counties.

- **Fidelity was measured to ensure pilot counties delivered the practice as intended to drive outcomes.**
- **Descriptive analysis illustrates pilot reach and implementation.**
- **A quasi-experimental design (QED) generates initial causal evidence of effectiveness.**
- **Qualitative data provide further context and ensure the experiences of APS staff are also elevated.**

Below we summarize the study approach. Additional details are included in the Methods section of this report.

Evidence-Building Approach

Colorado is committed to data-informed state investments and strategic decision making. As a pilot, it is imperative that research evidence is generated on AR to inform practice and policy development, scalability, and sustainability. To meet this goal, our evidence-building approach maximizes actionability with rigor. During the pilot, the focus was on data-informed learning and strengthening implementation. At the conclusion of a 2-year implementation period, the focus was on generating initial causal evidence of effectiveness. In this report, we present findings on reach, implementation, and impact evidence from 2 years of pilot implementation (January 4, 2023, through December 31, 2024). We centered authentic partnership with state and pilot county partners throughout the evidence-building process, from study design to interpretation of outcomes to moving results into action. As a result, the AR practice achieved Steps 1 through 4 of Colorado's Steps to Building Evidence and the practice is positioned to achieve a "promising" evidence designation according to House Bill 24-1428 (Evidence-Based Designations for Budget).

Fidelity of Implementation

Fidelity monitoring is an essential component of the AR Pilot evaluation. Fidelity monitoring helps to answer the question, "Is the pilot being implemented as intended?" Fidelity monitoring explores what activities actually occurred and contributed to outcomes and is essential to continuous quality improvement and to creating a cohesive, replicable version of the AR practice. Fidelity measurement was a collaborative process between the Colorado Lab, the AR Pilot Planning Specialist, and the CDHS Administrative Review Division (ARD).

Descriptive Analysis

Descriptive analysis is used to understand reach and implementation of the AR Pilot. These analyses are based on cases that were screened in, investigated, and closed between January 4, 2023, and December 31, 2024. In total, there were 9,790 cases, representing 14,991 allegations and 8,411 unique clients. To assess geographic variation, rural-urban county comparisons were done. To explore differences within the dual-track model, comparisons by track assignment (AR versus TR) were done. Tests of statistical significance were conducted throughout.

> *"I really appreciate being able to build some of those relationships with our clients prior to us going out and seeing them. I think it helps with some of that push back and just being like, 'Oh, my gosh! Someone's here at my door! What do you want?' versus, 'Hey, I would like to come and help you. Is it okay?' . . . I think that's [AR practice] helped build some of those relationships, especially with our self-neglect case clients."*
>
> **-Pilot County Caseworker**

Quasi-Experimental Design

The outcomes evaluation was a quasi-experimental design (QED); specifically, a matching approach using propensity scores called inverse probability weighting (IPW). IPW is a well-established and vetted procedure in the causal inference toolbox. Grounded in pilot design and the theory of change, this method identifies cases in the pre-pilot period with similar features to AR cases in the pilot period. This information is combined into a single propensity score representing the probability that a pre-pilot case would have an allegation tracked to AR had the dual-track model existed. Cases with similar scores are considered comparable and are weighted more heavily in the analysis. This method reduces subjectivity and improves the precision of causal estimates.

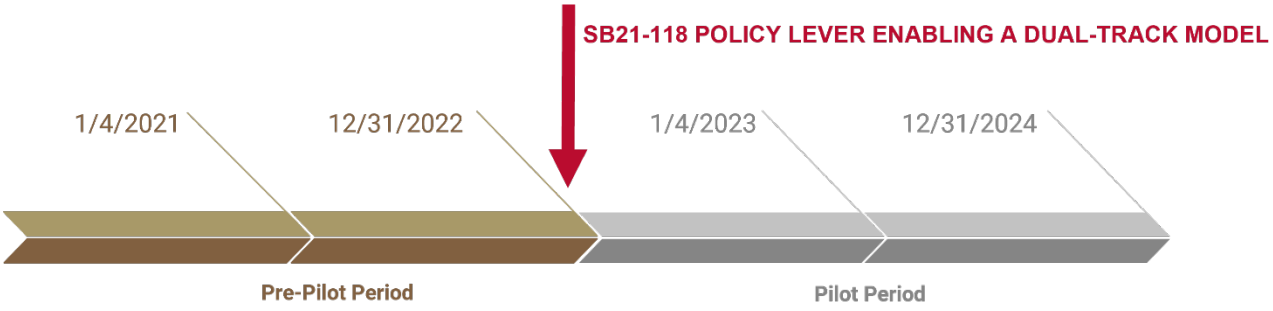
Defining the Sample

The analytic sample for the QED is defined as:

- **AR cases: Pilot period cases that contain one or more AR-tracked allegations (n = 5,547).**
- **AR-equivalent cases ("equivalent cases"): Pre-pilot period cases that would have at least one AR-tracked allegation had a dual-track model existed and are the strongest matches to AR cases in the pilot period (pool of n = 9,121)**

To understand the impact of the AR practice, outcomes of AR cases during the pilot (pilot period: January 4, 2023, through December 31, 2024) are compared with equivalent cases before the policy lever enabled a dual-track model (pre-pilot period: January 4, 2021, to December 31, 2022), as illustrated in Figure 2. We also present 6-month follow-up data through June 30, 2025.

Figure 2. Timing of Pre-Pilot and Pilot Periods



Why define AR cases as "one or more allegations?"

The decision to define "AR cases" as one or more AR allegations (versus cases with only AR allegations) reflects a more rigorous and inclusive approach to assessing the impact of AR. Including all cases with at least one AR allegation is a broader and more conservative approach. This allows us to detect whether even one AR-tracked allegation in a case makes a difference.

This approach also allows for more detailed analysis to inform implementation decisions and drive precision practice. For example, how do outcomes differ by cases with only AR allegations, versus those with mixed allegations? By defining the pilot sample as cases with at least one AR allegation, the dual-track model can be understood more holistically.

Practically speaking, the majority of AR cases (63.19%) have a single allegation.

Qualitative Data

Qualitative data was collected throughout the pilot, in commitment to implementation science principles. Implementation science grapples with issues such as effectiveness of practice for whom, in what setting, and under what conditions. Implementation science[1] can improve the relevancy of evidence generated and, in turn, increase the usefulness of results to state, county, and local decision makers. In the AR Pilot, qualitative data on acceptability, adaptability, and feasibility of the AR practice among state and county partners was collected. This included both formal focus groups[2] as well as iterative feedback from pilot counties. Qualitative data generated are used to provide context to results from fidelity, descriptive, and QED analyses. Further, qualitative data also speak to caseworker and client satisfaction with the AR practice—a valuable outcome in its own right. Narrative findings are integrated throughout fidelity, descriptive, and QED results sections.

Key Findings

Below we summarize key findings that speak to the impact of the AR practice and future potential of a dual-track model for Colorado APS. Only select findings are presented here, as aligned with APS priorities and leading results. Detailed findings are included in the Results section of this report.

1. Alternative Response pilot counties approached full fidelity by the end of the pilot.

By the end of the pilot, counties met or approached fidelity for all indicators except for Indicator 2 on scheduling the initial response. Counties dramatically improved their adherence to initial response during the pilot period thanks to their investment in continuous quality improvement. By the end of the pilot, the benchmark of 70% of AR-only cases receiving a scheduled initial visit was nearly reached (65% in final 6 months). Scheduling the initial response is hypothesized to be a driver in establishing collaborative engagement with clients and should continue to be closely monitored and adherence strengthened.

2. The dual-track model has meaningful demand in Colorado, especially to innovate approaches to self-neglect.

Close to half (42.9%) of all APS cases have only AR-tracked allegations, showing demand for a dual-track model. Self-neglect makes up over half (53.4%) of all AR-tracked allegations, showing a significant use case for the AR practice. The AR practice reaches a high number of clients who live alone and have few support networks, as well as clients with conditions that reflect an aging population (dementia/Alzheimer's, frail elderly).

3. The Alternative Response practice reduces repeat involvement.

Compared to equivalent cases, AR cases were 2.5% less likely to have a second screened-in case within 6-months of case closure ($p < 0.01$). The estimated repeat involvement rate for AR-equivalent cases was 10.01%, and the estimated rate for AR cases was 7.6%. Of the subset of clients who had a second screened-in case for self-neglect, compared to equivalent cases, AR cases with a self-neglect allegation were 7.2% less likely to have a second screened in case with a self-neglect allegation ($p < 0.01$). The estimated probability of repeat self-neglect in AR-equivalent cases was 17.2% and the estimated probability in AR cases was 10.0%. Of the subset of clients who had a second screened-in case for mistreatment, compared to equivalent cases, AR cases with a mistreatment allegation were 11.8% less likely to have a second case with a repeat mistreatment allegation ($p < 0.01$). The estimated probability of repeat mistreatment in AR-equivalent cases was 24.1% and the estimated probability in AR cases was 12.4%.

4. The Alternative Response practice reduces case length.

On average, AR cases closed 5.63 days earlier compared to AR-equivalent cases ($p < 0.01$). AR-equivalent cases closed 59.53 days after initial report receipt, and AR cases closed 53.90 days after

initial report receipt. This reduction comes from quicker completion of client baseline assessments, fewer days in determining findings or conclusions, and a shorter window between the date of last finding or conclusion and case closure. This shows long-term system efficiencies introduced through a dual-track model.

5. Conclusions for AR-tracked allegations reflect the low-risk nature of the AR track and signal that the AR track is being used appropriately.

A hallmark feature of AR is that there is no finding for AR allegations, but rather, a conclusion. It is important to understand the extent of impact that occurs in AR allegations and to assess whether this pattern aligns with the intended low-risk nature of the AR practice. This is because the dual-track model created a decision point for caseworkers to tailor their response to the level of risk. For allegations of mistreatment tracked to AR, conclusion data show an appropriate use of the AR track across all mistreatment categories where a decision point is available. All cases of substantial impact are reviewed by the AR Specialist to confirm the track is appropriately used.

6. Differences in Alternative Response and Traditional Response allegations changed the distribution of case closure reasons.

Case closure reasons have changed between pre-pilot and pilot periods. The top three closure reasons for AR cases are "intervention complete" (27.5%; compared to 31.8% in AR-equivalent cases); "no identified needs" (34.1%; compared to 31.3% in AR-equivalent cases); and "client not at risk" (13.0%; compared to 11.8% in AR-equivalent cases). This means that the proportion of cases not requiring intervention has increased in the pilot sample, reducing the proportion of cases where intervention can be completed. This is consistent with the analysis of extent of harm, showing the AR track is used to handle cases that are less severe, and is aligned with the theory of change and rules behind the AR practice, involving low-risk allegations and self-neglect only.

7. Support networks increase engagement.

Additional supports decrease the probability of client refusal, regardless of track. Refusal goes down by 1.10% per *each* additional support ($p < 0.01$). If APS caseworkers can better involve support networks, the opportunity for collaborative engagement with clients increases, making APS intervention more effective. The earlier engagement with support networks is done in the case, the better. Additional supports do lengthen cases, adding 7.37 days to a case ($p < 0.01$). Engaging support networks requires caseworker to make additional efforts to establish contact and coordinate together. The trade-off is worth it if building the support network results in sustained safety and improved client health in the long term. Engaging support networks does not necessarily increase caseworker burden—even if overall case length is increased—and a result of this upfront investment is the cost savings that can be realized later via a lower likelihood of client reinvolvement in APS.

> *"What is benefiting our clients—what they're finding value in—Is the ability to identify who their natural supports are, and it's allowing them to feel more in control of the situation."*
>
> **- Pilot County Caseworker**

Recommendations

> **The effective repeal date of SB21-118 is July 1, 2027. Prior to this date, CDHS must make recommendations on the future of AR in Colorado.**
>
> **Based on favorable findings, the AR practice should be recommended for statewide scaling.**

Final reporting to the General Assembly is due in the January 2026 legislative session by CDHS. Based on favorable evaluation findings—alongside support by implementing partners—the AR practice should be recommended for statewide scaling.

1. Areas for Priority Rule Change

Rule promulgation will establish the regulations governing a statewide dual-track model, such as criteria for track assignment and timelines for APS. During rule-making, we recommend revisiting initial response timelines to reflect pilot learnings and strengthen implementation.

2. Phased Rollout

Innovations that become permanent practice are benefited by a phased rollout to ensure county readiness and provide the state adequate time to build the structural and cultural conditions necessary for success. CDHS should develop a clear plan for how counties can opt in to the dual-track model and a feasible 3- to 5-year timeline to achieve statewide implementation.

3. Adequate Resourcing

The AR practice represents collaboration at its core—and that collaboration is true for caseworkers to clients, as well as between counties and the state. Implementing a dual-track model with fidelity to drive outcomes requires adequate resourcing at the state level (e.g., maintain AR Specialist position) and at the county level (e.g., provide professional development opportunities).

4. Advancing Partnerships for the Aging Population

While evidence building for AR was focused firstly on APS response, evaluation results also have implications for the aging population across units at CDHS, including the State Unit on Aging. The state's first-ever Multi-Sector Plan on Aging provides a prime opportunity to advance partnerships and leverage results of the two-year outcomes evaluation toward statewide infrastructure.

5. Applying the Evidence-Based Decision Making Approach

Colorado's Evidence-Based Decision Making (EBDM) approach for state government exists at the intersection of the best available research evidence, decision-makers' expertise, and community needs and implementation context. The EBDM approach provides a leading-edge framework to activate results from the 2-year rigorous evaluation.

Colorado Evaluation & Action Lab
UNIVERSITY OF DENVER

Detailed Methods

Methods

Details on the methods used in this study are provided below for:

- **Fidelity of implementation**
- **Descriptive analysis**
- **Quasi-experimental design**
- **Qualitative narratives**

Fidelity of Implementation

Fidelity monitoring was an essential component of the AR Pilot evaluation. Fidelity monitoring helps to answer the question, "Is the pilot being implemented as intended?" Fidelity monitoring explores what activities actually occurred and contributed to outcomes. Fidelity monitoring is essential to continuous quality improvement and to creating a cohesive, replicable version of the AR practice. Fidelity measurement was a collaborative process between the Colorado Lab, the AR Pilot Planning Specialist, and the CDHS ARD.

The Colorado Lab assessed fidelity on seven indicators that represent essential elements of the AR practice. We analyzed full fidelity data twice during the pilot: once in fall 2023 (initial) for learning and once in late 2024/early 2025 (final) for evaluation reporting. Based on initial results, additional fidelity assessment for select indicators was done in January 2024. This interim reassessment is described in more detail below.

To measure fidelity, each county and indicator were assigned a rating of "met," "approaching," or "not met" based on pre-determined benchmarks. We used quantitative data from Colorado Adult Protective Services (CAPS) and ARD, and attendance data collected by the APS team. These quantitative metrics were supported by qualitative data examples collected by the AR Pilot Planning Specialist. Appendix A describes each of the fidelity indicators, including data sources, definitions, cut points for *met*, *approaching*, and *not met* categories, and qualitative data examples.

After initial fidelity assessment in fall 2023, the decision was made to re-assess select measures after a full year of data from CAPS and ARD became available. Accordingly, the Colorado Lab updated and reported on select measures in January 2024. The following measures were re-assessed:

1. Measure 2: Initial Response. Rationale for re-assessment: Updated guidance was provided to counties in August 2023 regarding scheduling initial visits and a new data entry option was added to CAPS.
2. Measure 3: Track Changes. Rationale for re-assessment: A new field was added to CAPS in September 2023 to reflect the track change "date of decision in field" rather than just the date the track change was entered into CAPS. The underlying data was used to assess how this fidelity measure changed.

3. Measure 4: Investigation and Conclusion; and Measure 5: Matching Needs to Services. Rationale for re-assessment: ARD data for all counties became available and the ARD data were able to be split out for specific case types.

In this final evaluation report, we summarize county-level fidelity assessments for all 15 pilot counties, combined and separately for urban and rural counties. We also spotlight how the initial response fidelity measure changed over the course of the pilot as counties improved their approaches and strengthened implementation.

Descriptive Analysis

Descriptive analysis is used to understand reach and implementation of the AR Pilot. These analyses are based on cases that were screened in, investigated, and closed between January 4, 2023, and December 31, 2024. In total, there were 9,790 cases, representing 14,991 allegations and 8,411 unique clients. To assess geographic variation, rural-urban county comparisons were done. To explore differences within the dual-track model, comparisons by track assignment (AR versus TR) were done. Tests of statistical significance were conducted throughout.

Quasi-Experimental Design

The outcomes evaluation employs a quasi-experimental design (QED) approach; specifically, a matching approach using propensity scores called inverse probability weighting. Matching is an umbrella term used to describe techniques by which a researcher constructs a control group that is similar to a treated group using available untreated units. There are different ways to accomplish matching, but all approaches seek out potential comparisons by looking for similarities along sets of pre-specified variables called "matching variables." Differences in matching approaches stem from how matches are found and what criteria are used to determine what constitutes a valid match.

To understand the intuition behind matching in the context of the current evaluation, consider a one-to-one matching example using a case with an AR-tracked allegation from the pilot period. Reviewing the case, we gathered information collected at intake (e.g., the number and type of allegations on the case, client demographics, county of reporting, etc.) and then looked for a case with the same or similar characteristics in the pool of cases from the pre-pilot period. Once we found the comparison case that best matched the AR-tracked allegation case, we compared client outcomes across the two. The difference between the two was assumed to be attributed to the availability of AR and AR-track assignment. Importantly, causal interpretation of the difference assumes that the matched case, all else equal, would have had the same outcome as the case with an AR-track allegation had a dual-track model been available in the pre-pilot period.

When the previous exercise was repeated for all pilot cases with an AR-track allegation, the average of all the differences was interpreted as the causal effect of having the AR track available and using it (the average treatment effect on the treated). A crucial requirement for matching is that the pre-pilot period pool of cases is large and varied enough to contain valid counterfactual cases. Though the above is simple, the intuition behind the one-to-one matching example carries over to the current and more rigorous analysis.

Propensity Score Analysis

The use of propensity scores in matching is a dominant matching paradigm and vetted method for aggregating information from multiple matching variables into a single value. An observation's propensity score is its estimated likelihood of being "treated," conditional on a set of observable characteristics, making scores bounded in the range [0,1].[3] Observations are considered similar, and thus good matches, when they have similar probabilities of being treated. In the present evaluation, our observation was at the level of a case and treatment was defined as having an allegation tracked to AR. Pre-pilot and pilot cases were considered good comparisons if they had similar *likelihoods of having at least one allegation assigned to the AR track*— regardless of whether they actually had an allegation assigned to the AR track.

Calculating propensity scores requires modeling the assignment of treatment using a flexible logit regression model, then estimating it using maximum likelihood estimation. Included in the logit are a set of logic-model informed determinants of AR-track assignment.[i]

Once estimated, the logit regression model was used to predict the likelihood that a case contained an AR-tracked allegation (bounded between 0 and 1) for all cases in the pre-pilot period and cases with an AR-tracked allegation in the pilot period. These predictions are the propensity scores.

QED Assumptions and Justification

1. Credible matching QEDs can answer why cases, observably similar, would have different treatment statuses. Hypothetically, if two cases match perfectly along a set of relevant variables (and thus have the same propensity scores), but one has an allegation assigned to the AR track and the other does not, we may be suspicious that the cases are not valid comparisons. If the reason for the difference is unobserved to the researcher, there is the potential for bias in our estimation.

   In observational settings, this concern poses a significant challenge for researchers. Fortunately, in the current setting, pre-pilot cases could not have an allegation assigned to the AR track because the dual-track model was not available at that time. Had the AR track been available, allegations in comparable cases would arguably have been assigned to the AR track in the pre-pilot period.

[i] The set of covariates should account for a comprehensive list of case characteristics predictive of AR-track designation within a case.

2. In planning the QED, the research team determined that comparing cases in counties pre- and post-AR was superior to comparing cases in pilot counties to cases in non-pilot counties. Two key features of the state environment informed this choice. First, counties have extensive control over implementation and there are county-specific idiosyncrasies that we cannot observe.[ii] Thus, comparing contemporaneous cases in pilot and non-pilot counties would result in poor matches, even if similar on paper. In the current design, we were able to match cases on county.

   Second, there is a concern that between the pre-pilot and pilot periods, other policies, rules, or practices were introduced that impacted the implementation of TR simultaneously. Comparing pre-pilot cases to pilot cases would result in the QED estimating not just the introduction and use of the dual-track model, but also concurrent changes in practice. Consulting with our partners in CDHS, we determined that there were no noteworthy changes in practice, unrelated to the introduction of AR, that impacted the implementation of the dual-track model.

Inverse Probability Weighting

When sample sizes are large, it is possible to have more than one valid match. For example, a pilot case with an AR allegation could be similar to pre-pilot Case A along half of the observed measures and also similar to pre-pilot Case B along the other half of the observed measures, resulting in similar propensity scores. Both Cases A and B offer useful information as comparisons, but neither is superior; as such, instead of having to pick one case to use as a comparison, we construct a sample that weighs comparison cases.[iii]

A cases weight is the numerical approximation of that cases' importance as a comparison case. The present study uses inverse probability weighting (IPW)—pre-pilot cases with *high* propensity scores are given *greater* weight, while pre-pilot cases with *small* propensity scores are given *lower* weight. IPW is specifically designed for use with propensity scores and is the most precise way to estimate causal effects given a large enough sample and a flexible way to estimate the propensity score.[4] Both conditions were satisfied in the present analysis.

Pilot period cases with AR-tracked allegations that get the largest weight were the ones most like the *pre-pilot* cases—these cases had small propensities, meaning least likely to have an allegation assigned to the AR track, but who did, nonetheless. In other words, a pilot period AR-allegation case with a small propensity score will be weighted more (hence the word "inverse" in IPW)

[ii] We cannot control for these factors and an approach using county-level fixed effects would run into issues of perfect collinearity.

[iii] In exact matching procedures, researchers must make various decisions, including determining if the design will allow the replacement of comparison cases after matching, and if not, the order in which cases are matched. When exact matches are not possible, researchers often coarsen variables to enable better matching, but at the cost of reducing confidence in the causal estimate. In one-to-*k* nearest neighbor matching procedures, researchers must also decide *k* and set parameters on what is considered near.

because though it had an AR-tracked allegation, it had a low probability of having an allegation tracked to AR. At the other end, pre-pilot cases with the largest weights were the ones with large propensity scores. These are cases that, had the AR track been available, would have contained allegations tracked to AR and are thus the best comparisons. When comparing pre-pilot with pilot cases, IPW adjusts the two sides to prioritize the most comparable set of cases when calculating an average difference. Using the weights as a guide, IPW helps researchers avoid the subjectivity associated with other matching methods by putting the focus on specifying the regression model used to estimate propensity scores as opposed to the matching method itself.[iv]

For the present study, we estimated the average difference between cases that had at least one AR-tracked allegation and pre-pilot cases that did not but are comparable because they arguably would have had at least one AR-track allegation had the dual-track model been available. Cases in the pilot period with at least one AR-tracked allegation are from here on referred to as "AR cases," and comparable cases from the pre-pilot period which were weighted heavily are referred to as "AR-equivalent cases" or simply "equivalent cases."

> **AR cases:** Pilot period cases that contain one or more AR-tracked allegations.
>
> **AR-equivalent cases:** Pre-pilot period cases that would have at least one AR-tracked allegation had the dual-track model been in place. Equivalent cases provide the strongest comparisons to AR cases.

Case-Level Estimation Approach

Considering the potential for both AR and TR allegations within the same case, the estimation strategy approaches allegations as pieces of a single case rather than as a series of independent factors. This approach is in line with the study's aims and helps capture the complexity and potential interplay of multiple allegations in a single case. By defining AR cases as cases containing at least one AR-tracked allegation, we are transparent about the fact that cases could also contain TR-track allegations. This implies that the impacts of AR-track assignment at the case level could be muted when the case also contains TR-track allegations. Therefore, we interpret the estimated effects as floors, choosing to err on the side of caution.

Generating propensity scores at the case level requires defining treatment differently than if we were analyzing outcomes at the allegation level. Track assignment is determined at the allegation level, but outcomes are at case level; thus, treatment is not whether an allegation was tracked to AR or TR but instead must be a binary case-level analog.

[iv] IPW is not matching in the traditional sense; singular cases are not being directly matched to other singular cases. Instead, cases are weighted based on their similarities. Estimation of the average effect boils down to estimating a regression of the outcome on whether a case has at least one AR-tracked allegation after applying case weights.

Treatment is defined as whether a case had at least one AR-tracked allegation. In the propensity score estimation process, this means we are predicting the probability of a case having at least one AR-tracked allegation and searching for cases in the pre-pilot period with high probabilities of having at least one AR-tracked allegation, had the dual-track model been available. Pre-pilot cases that had a high probability of having at least one AR-tracked allegation are AR-equivalent cases.

Practically speaking, given that over two-thirds of cases were single allegation cases, the distinction between allegation and case becomes moot. The average treatment effect will strongly reflect a comparison to single AR-allegation cases and the language of "at least one AR-tracked allegation" will tend to reflect AR-allegation only cases.

Once the propensities and weights were calculated, we then focused on specific case variants and added covariates to improve statistical precision by reducing unexplained variance in the outcome. Generally, when modeling both the treatment in the first stage and then the outcome, we aim for a property called "double robustness" which gives an unbiased estimate as long as the treatment is fully modeled, *or* the outcome is fully modeled (or both). Combining regression and weighting can lead to additional robustness to misspecification by both removing the correlation between the omitted covariates, and by reducing the correlation between the omitted and included variables. This is the idea at the core of doubly-robust estimators. [5]

Analytic Sample

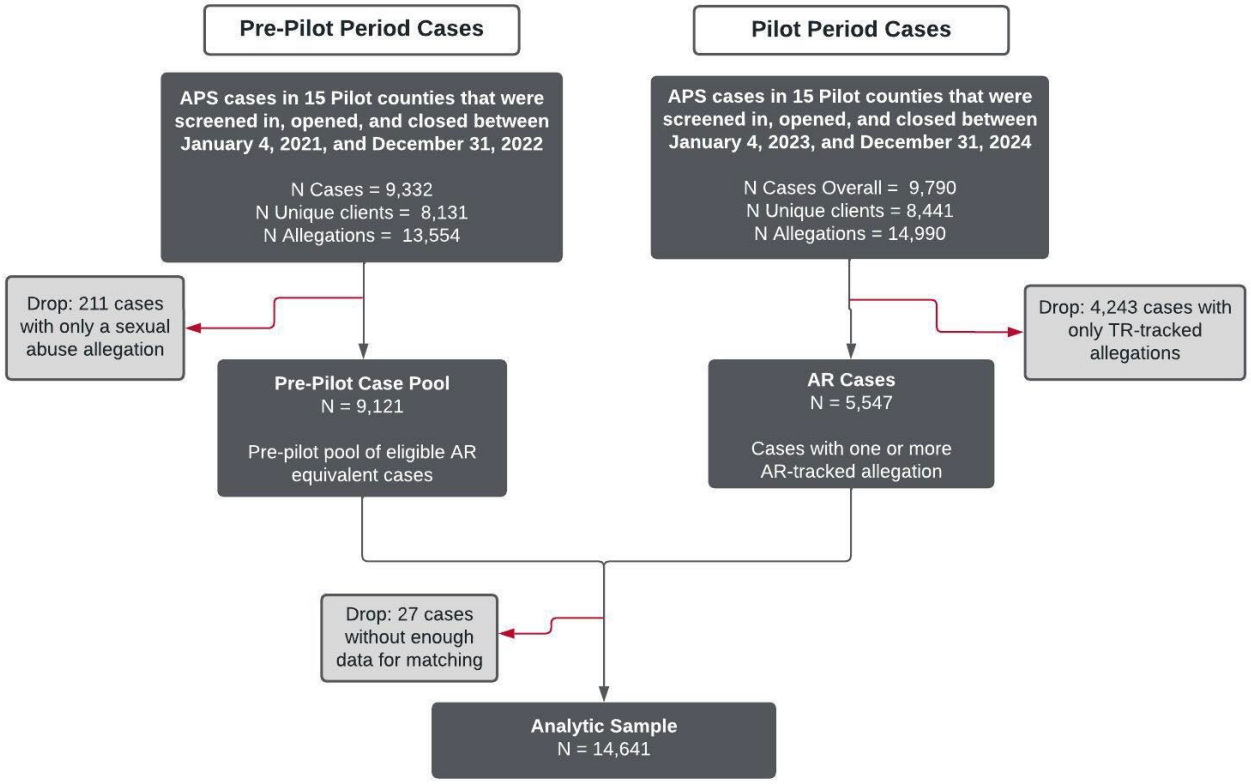
The analytic sample consists of all screened-in allegations from pilot counties that were opened, investigated, and closed between January 4, 2021 through December 31, 2022 (the pre-pilot period) and between January 4, 2023 through December 31, 2024 (the pilot period). A 6-month window (through June 30, 2025) is added at the end to observe the incidence of client repeat involvement in APS. Repeat involvement is an outcome and not considered separate cases; as such, it is not included in the analytic sample. All allegations in the sample met the definition of mistreatment or self-neglect. Allegations that did not meet the definition were not considered. Collapsing allegation-level data into case-level data, we began with 19,122 screened-in cases opened, investigated, and closed within either pre-pilot or pilot period.

Figure 3 illustrates how we arrived at the analytic sample for the QED. Of the 19,122 cases, 9,790 cases are from the pilot period. Of those, we dropped 4,243 TR-allegation-only cases. These should not be used to estimate the impact of AR as they violate the first identifying assumption discussed earlier and lower overall match quality.[v] This restriction leaves 5,547 AR cases.

[v] TR-allegation-only cases from the pilot period are not included and thus will not have propensity scores or weights. These cases contain poor matches, because had they been "equivalent" according to our definition, they should have an AR-track allegation since the option was available. To investigate this further, we performed a robustness check that left TR-allegation-only cases in the analytic sample. Doing so skewed the distribution of propensity scores substantially, created a pronounced spike in density very close to 0. Adding a lot of cases that have low propensities introduces thousands of objectively poor matches. Estimates are attenuated, though statistical significance holds. In addition, balance tests indicate weighting does a worse job at eliminating differences in the matching covariates.

On the other side, there are 9,332 cases in the pre-pilot period. By definition, these cases can only have TR-track allegations. We dropped 211 cases with only sexual abuse allegations, because corresponding cases in the pilot period were already dropped (even if estimated to have small weights, they are poor matches). Combining valid pre-pilot and pilot cases, the analytic sample contains 14,668 cases. During estimation, we lost an additional 27 cases because they did not have the full set of values for all matching variables, resulting in a final analytic sample of 14,641 cases.

Figure 3. Analytic Sample Flow Chart



Matching Variable Selection and Design Validation

When specifying the logit model to estimate the propensity scores at the case level, we aimed to include variables that determined whether a case had at least one AR allegation. The variables consider the available data and the baseline inputs from the logit model by including a variety of client characteristics and conditions. The number of variables and extent of the non-linearities of the relationship between the predictors and the outcomes can be modified to ensure, after weighting, baseline equivalence is achieved. More relevant variables increase the likelihood of achieving the best covariate balance between AR cases and equivalent cases and mitigates bias

from omitted variables. That said, the indiscriminate inclusion of too many variables increases the likelihood of including colliders and irrelevant variables.[vi]

The matching variables included were measured at intake, thus not impacted by the intervention itself.[6] We excluded variables where most values were missing (e.g., client income source and health insurance type) and "live" variables that could have been updated after intake (e.g., specific at-risk conditions). In addition, we created variables to capture dynamic interactions between age and some of the other included variables.

Case-Level Matching Variables

- Indicator variables specifying the presence of each mistreatment type (caretaker neglect, exploitation, harmful act, physical abuse, sexual abuse) and/or a self-neglect allegation.
- Indicator variables specifying the county of residence.
- Client age represented via a cubic function.
- Indicators for race/ethnicity, gender, and primary language.
- Indicators an existing condition, including difficulty with decision making, medical conditions, memory deficits, mental illness, substance abuse, physical conditions.
- Presence of any physical, mental, or behavioral condition.
- Whether the client was in a state of emergency or in immediate harm.
- Fourteen age interaction terms, including age and allegation-type indicators, age and initial condition indicators, and an age emergency or immediate harm indicator.

The logit model estimates the likelihood of a case having at least one AR-tracked allegation on all these matching variables. Importantly, including indicators for county captured idiosyncratic county-level determinants of AR-track assignment, including urbanicity, county size, and department norms. The cubic function in age and numerous age interaction terms aimed to account for various degrees of severity and urgency present as clients get older.

Importantly, the weights will mechanically generate an analytic sample where differences in the matching are as close to zero as possible. Thus, the lack of remaining variation across the two groups limits the ability to investigate the independent effect of these variables on outcomes.

[vi] Colliders can be covariates that are arguably affected by the use of the AR track itself and generate biased estimates. Irrelevant variables are not predictive of outcomes and can increase estimated variance, and thus imprecision. We also note that the "curse of dimensionality" is not applicable in the current analysis.

Validating the Design

Appendix B details the checks performed to ensure the validity of the IPW approach. Overall, validation tests indicated that the weighting scheme was appropriate (i.e., sufficient overlap in the propensity score distributions) and successful at generating strong balance across the two samples (i.e., equivalence at baseline across all matching variables). The validation checks reinforced the argument that the estimates produced are doubly robust and causal.

Confirmatory Outcomes

The confirmatory outcomes fall into six bins, for a total of 25 outcomes. Together, these groupings speak to client- and system-level outcomes. The bins are listed in Appendix C along with explanations of each client- and system-level outcome within them. All outcomes are reported for the full 2-year analytic sample.

Qualitative Narratives

Qualitative data was collected throughout the pilot, in commitment to implementation science principles. Implementation science grapples with issues such as effectiveness of practice for whom, in what setting, and under what conditions. Implementation science can improve the relevancy of evidence generated and, in turn, increase the usefulness of results to state, county, and local decision makers. In the AR pilot, qualitative data on acceptability, adaptability, and feasibility of the AR practice among state and county partners was collected. This included both formal focus groups as well as iterative feedback from pilot counties. Qualitative data generated are used to provide context to results from fidelity, descriptive, and QED analyses. Further, qualitative data also speak to caseworker and client satisfaction with the AR practice—a valuable outcome in its own right. Narrative findings are integrated throughout fidelity, descriptive, and QED results sections.

> *"I feel like people are able to be more honest when they don't have that kind of pressure of a substantiation. That creates rapport and uncovers more, faster."*
>
> **- Pilot County Caseworker**

Colorado Evaluation & Action Lab
UNIVERSITY OF DENVER

Detailed Results and Recommendations

Fidelity Results

By the end of the pilot, counties met or approached fidelity for all indicators except for Indicator 2: Scheduling the Initial Response.

- **Counties dramatically improved their adherence to scheduling the initial response during the pilot period, thanks to their investment in continuous quality improvement.**
- **Scheduling the initial response is hypothesized to be a driver in establishing collaborative engagement with clients and should continue to be closely monitoring and implementation strengthened.**

Throughout the pilot period, counties delivered the AR Pilot as intended and, overall, met fidelity to the essential elements. Table 2 shows the percentage of pilot counties that met or approached fidelity for each of the seven indicators at the initial 8-month assessment, the 12-month reassessment, and the final 24-month assessment. By the end of the pilot, counties met or approached fidelity for all indicators except for Indicator 2: Initial Response, in which fewer than half of counties met or approached fidelity for scheduling initial visits on AR-only cases.

Table 2. Fidelity of Implementation Summary: Percent of Pilot Counties with a "Met" or "Approaching" Fidelity Assessment (N =15 Counties)

Fidelity Indicator	8 Months	12 Months	24 Months
1. Initial Track Assignment	100%	No reassessment	100%
2. Initial Response	0%	67%*	47%
3. Track Changes	100%	100%	100%
4. Investigation and Conclusion	100% (N = 10)	100% (N = 14)	92% (N = 13)
5. Matching Needs to Services	100% (N = 9)	100% (N = 12)	100% (N = 13)
6. Use of Data	93%	No reassessment	93%
7. Continuing Education/Professional Development	100%	No reassessment	100%

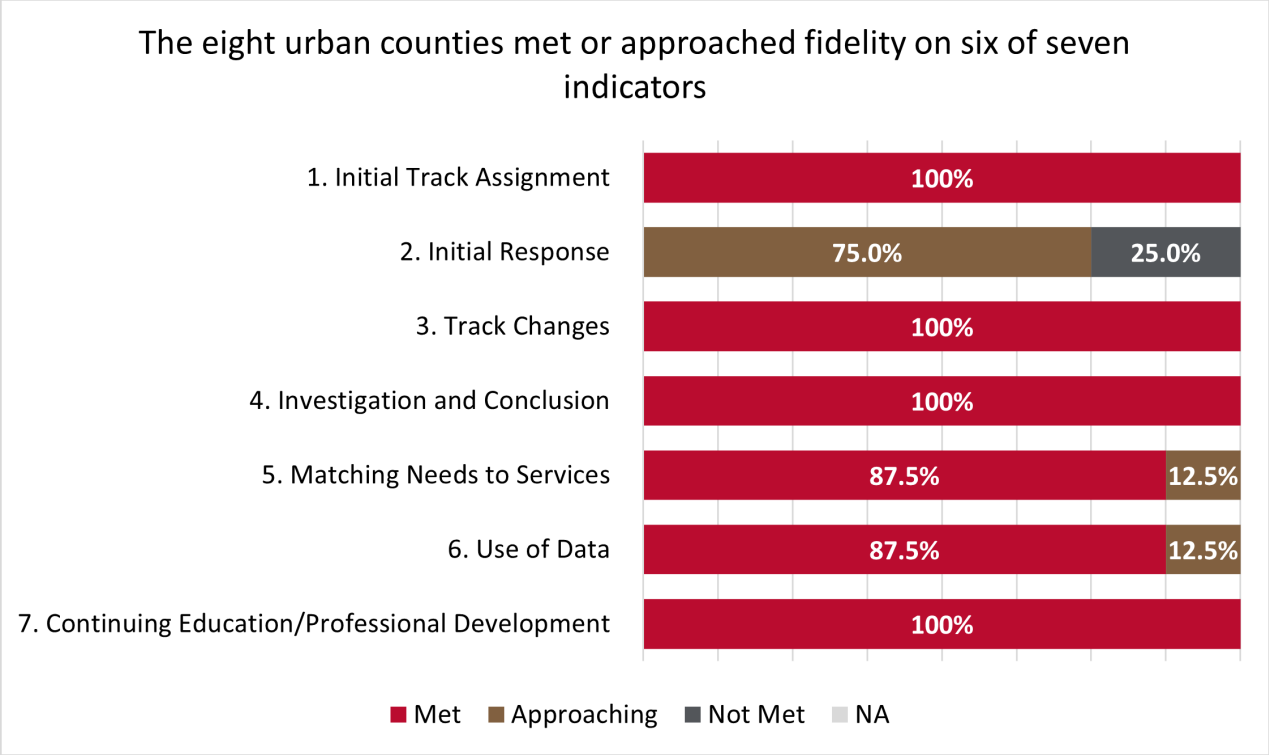
Note: For the first 8-month fidelity assessment, the number of counties is less than 15 because some counties had not undergone an ARD review in 2023 or there were no cases to review that met the criteria for the indicator. For the first 12-month fidelity assessment, some counties had no cases to review that met the criteria for the indicator. For the 24-month fidelity assessment, two counties had no cases to review that met the criteria for the indicator.

*12-month reassessment of Initial Response was based on cases that opened and closed between September and December 2023 only. In August 2023, counties received updated guidance on scheduling the initial response and CAPS data entry options were updated. The 24-month data reflect fidelity achievement across the full 2-year pilot period.

Fidelity Results for Urban and Rural Counties

Below, we report results of the 24-month fidelity assessment for the eight urban counties (Figure 4) and seven rural counties (Figure 5) separately. Urban counties met or approached fidelity on six of the seven fidelity indicators when data were available. For the initial response indicator, six urban counties were approaching fidelity, and two urban counties did not meet fidelity. Thus, in six urban counties, a scheduled initial visit was used in 50% to 70% of AR-only cases; and in two urban counties, fewer than 50% of AR-only cases received a scheduled initial visit.

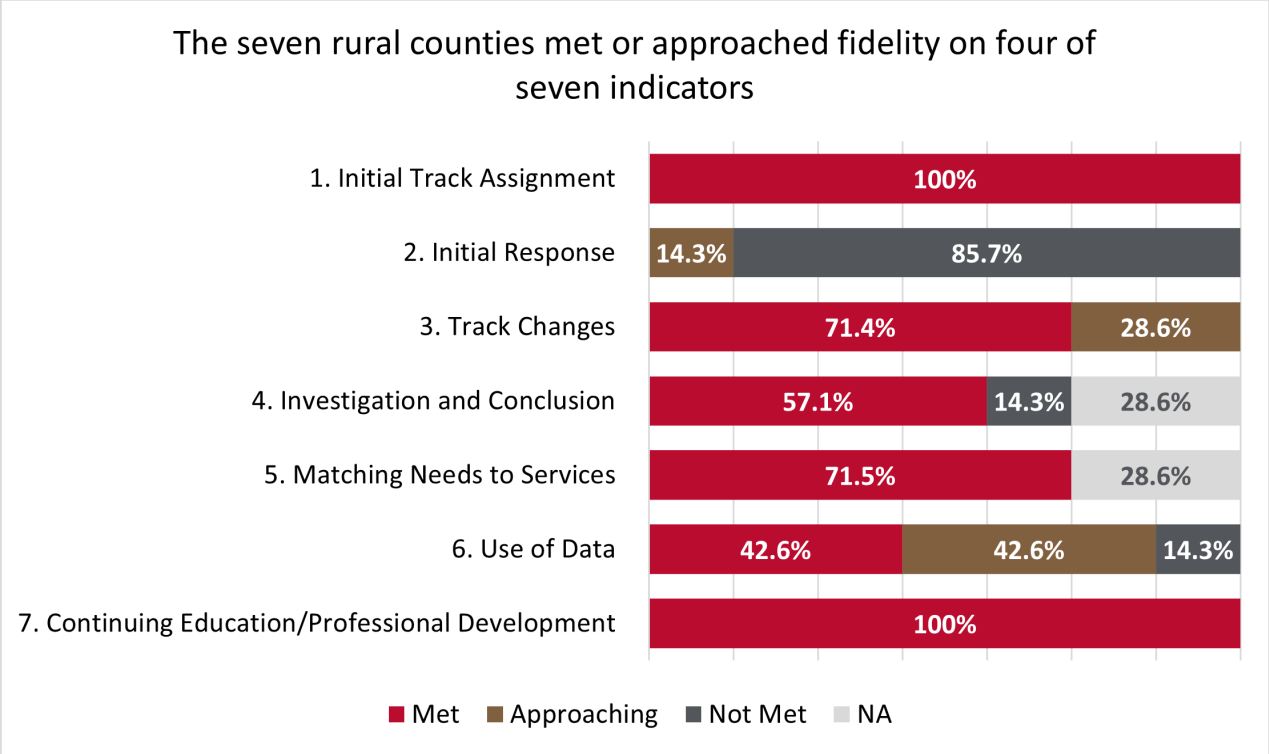
Figure 4. 24-Month Fidelity Assessment Among Urban Pilot Counties



Rural counties met or approached fidelity on four of the seven fidelity indicators when data were available. For the initial response indicator, only one rural county approached fidelity, and the remaining six rural counties did not meet fidelity. Thus, in one rural county, a scheduled initial visit was used in 50% to 70% of AR-only cases; and in six rural counties, fewer than 50% of AR-only cases received a scheduled initial visit.

Additionally, one rural county did not meet fidelity for Use of Data, such that a county representative attended fewer than 60% of learning sessions and county meetings over the course of the 2-year pilot. Counties were expected to regularly attend pilot county meetings and learning sessions so they can use all available opportunities to improve APS practice and drive outcomes for clients. All pilot counties committed to attending 80% of these meetings when they agreed to participate in the pilot or designate a proxy when the main representative is unable to attend.

Figure 5. 24-Month Fidelity Assessment Among Rural Pilot Counties



Note: Data for Indicator 4 (Investigation and Conclusion) is based on a very small sample size from an ARD review for the one county that did not meet fidelity. These data should be interpreted with caution. Earlier data from this county and data from other counties indicate that fidelity is typically met for this indicator.

Scheduled Initial Visits

With a focus on using data for learning, the initial fidelity assessment at 8 months provided information about areas for improvement. At first measurement, the rate of scheduled initial visits for cases with only AR-tracked allegations was very low (<50%). This implies that the option to schedule an initial visit may not have been exercised consistently or appropriately for AR-only cases.

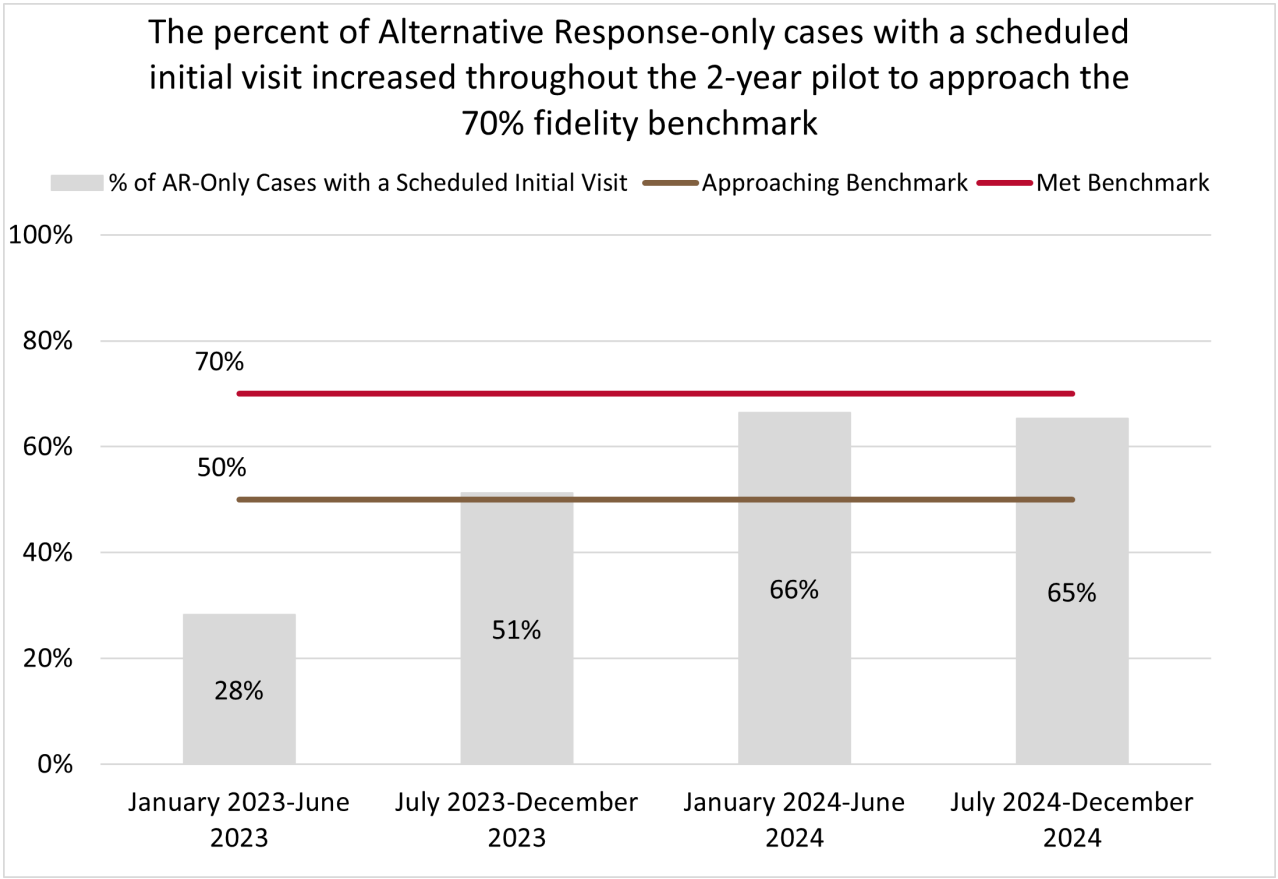
The option to schedule the initial visit is a hallmark of the AR track and is essential to the theory of change. This is because encouraging positive rapport at case start can help strengthen the relationship between caseworker and client and, in turn, improve service delivery and uptake. At first measurement, no counties met fidelity on this measure (goal: >70% scheduled initial visits for AR-only

> *"I think that being able to call ahead and schedule a home visit has been of value to clients and families. I often think how stressful it would be for me if somebody just showed up at my house and started asking me a bunch of questions, very intrusive questions...I think being able to call ahead and schedule a visit is more client centered."*
>
> **- Pilot County Caseworker**

cases) and no counties were approaching fidelity (goal: 50% to 70%). Qualitative data indicate that the low rates of scheduled initial visits reflect a combination of counties learning to schedule an initial visit, alongside a period of time when caseworkers were struggling to complete scheduled visits in the 3-day timeframe set by APS rule. The state issued a memo on August 2, 2023 to provide clarification and interpretive practice guidance for non-emergent initial contacts within the AR track. The guidance in this memo did not negate or contradict existing rules; rather, it provided practice guidance related to existing APS rules for AR-pilot counties. The state and pilot counties also heavily invested in learning exchanges to identify best practices in initial response.

At 12 months, the Colorado Lab re-assessed whether rates of scheduled initial visits increased in the last four months of 2023. As shown in Table 2, two out of three counties used scheduled initial visits on at least half of their AR-only cases that opened and closed between September and December 2023. Examining the initial response metric in 6-month intervals, we see the use of scheduled initial visits improved dramatically, from a low of 28% of AR-only cases in the first 6 months, to 51% in the second six months, to 66% and then 65% in the last two 6-month periods of the pilot, respectively (Figure 6). While we assessed fidelity at the county level, if fidelity for the initial response indicator was assessed for all cases in the pilot regardless of county, we see that the benchmark of 70% of AR-only cases receiving a scheduled initial visit was nearly reached in the last year of the pilot. Improving this rate is essential to drive consistency in other positive outcomes.

Figure 6. Percent of AR-Only Cases with a Scheduled Initial Visit by Case Open Date, All Counties



Descriptive Analysis

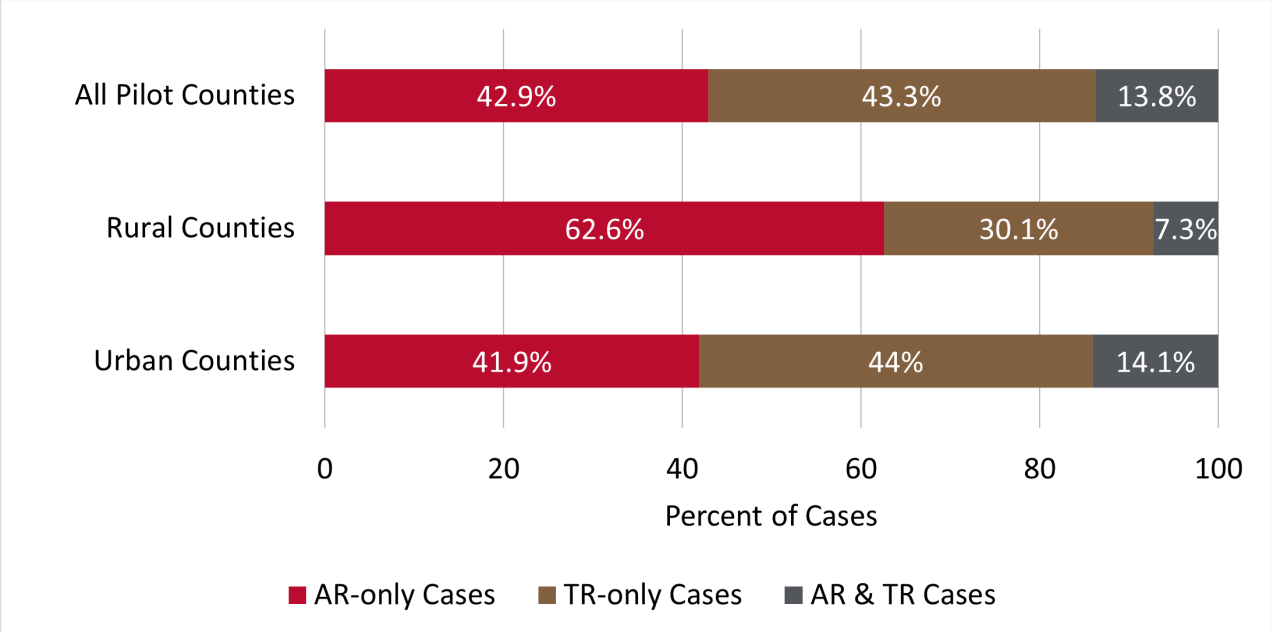
A total of seven indicators were descriptively analyzed, covering reach and implementation of the AR practice in pilot counties. Findings show:

- **Close to half (42.9%) of all APS cases have only-AR tracked allegations, showing demand for a dual-track model.**
- **Self-neglect makes up over half (53.4%) of all AR-tracked allegations, showing a significant use case for the AR practice.**
- **The AR practice reaches a high number of clients who live alone and have few support networks, as well as clients with conditions that reflect an aging population (dementia/Alzheimer's, frail elderly).**

Insight 1: Close to half of APS cases have only Alternative Response-tracked allegations. The percentage is higher in rural counties.

The AR track is being robustly used by pilot counties, signaling the need for a dual-track model that can tailor response to level of risk (Figure 7). Rural counties have a significantly higher percentage of cases with only AR allegations ($p < 0.01$). Qualitative narratives indicate that—especially in small-knit communities—the AR track can improve collaboration, particularly for older adults who are strongly independent and desire to age in place. The average age of clients served by APS in pilot counties is 67.8 years old.

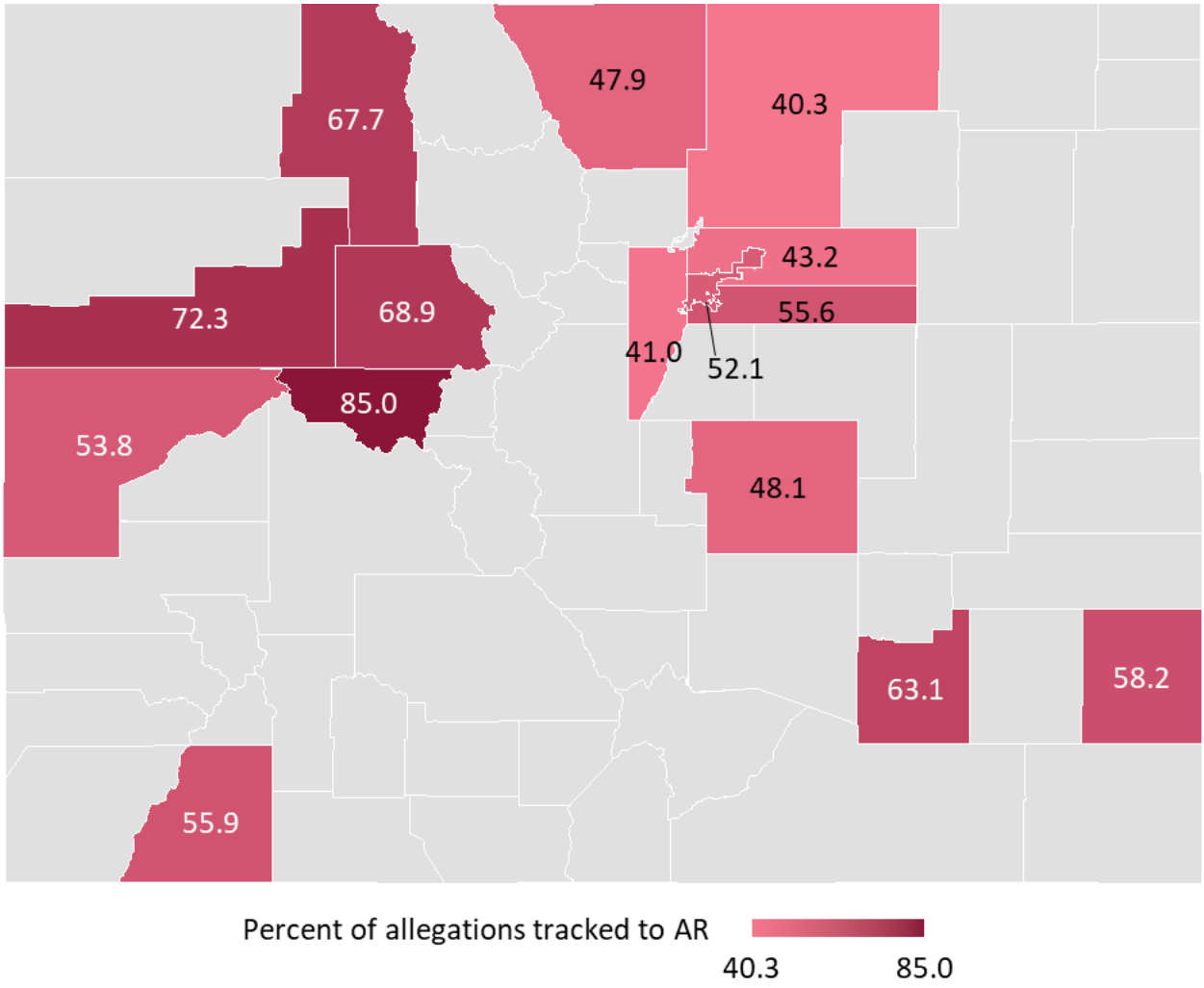
Figure 7. Allegation Breakdown on Cases



Insight 2: Alternative Response Pilot counties show variation in their use of the Alternative Response track.

It is important to explore how the dual-track system is unfolding in different contexts, such as rural versus urban (Figure 8). Variation may reflect caseworker discretion, different case complexities, different staffing structures (e.g., a generalist county where caseworkers play multiple roles, versus a "specialist" county where caseworkers play a specific role like intake caseworker versus ongoing caseworker), different community-based service availability, and different caseloads. Importantly, higher use of the AR track is correlated with higher self-neglect allegations (correlation coefficient: 0.6251). This is an anticipated result since; By rule, all allegations of self-neglect are automatically assigned to the AR track.

Figure 8. Use of the Alternative Response Track by County

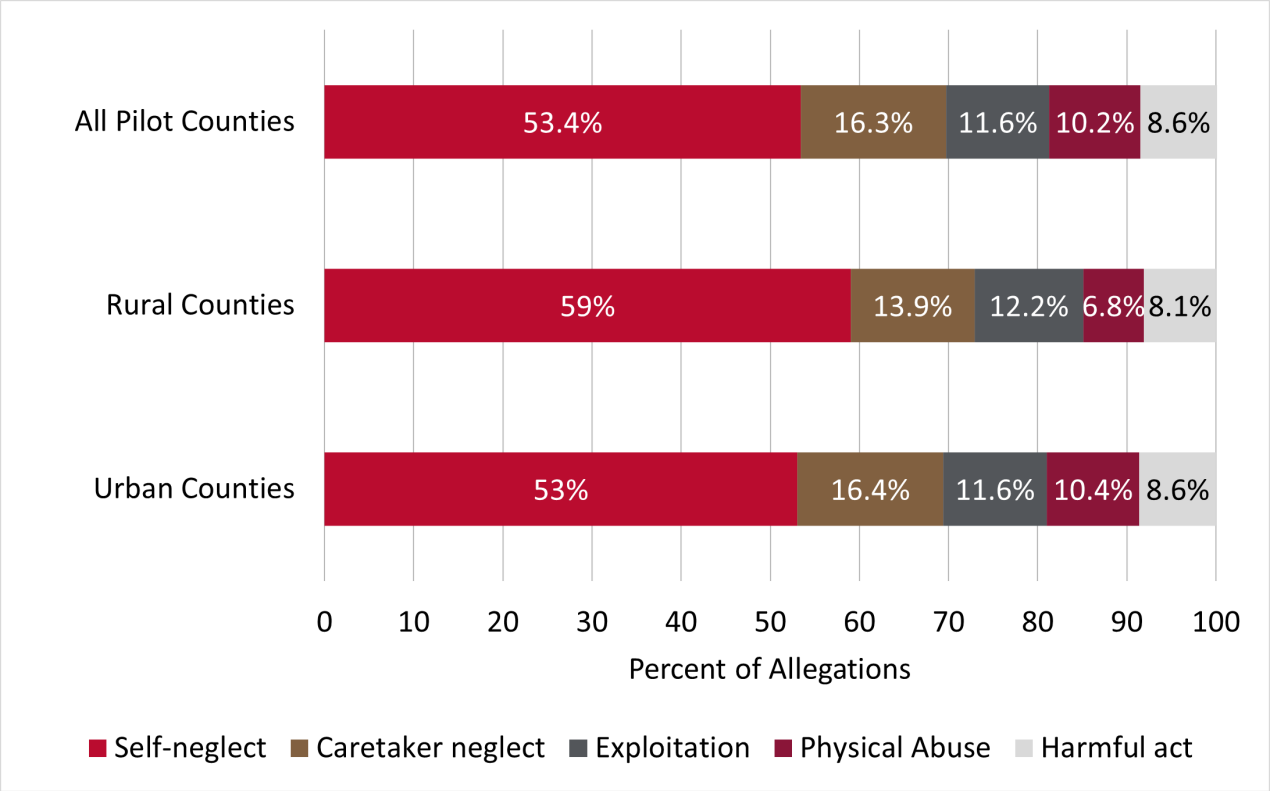


Insight 3: Self-neglect makes up over half of all Alternative Response-tracked allegations. The percentage is slightly higher in rural counties.

Understanding what is driving track assignment can help inform understanding of the model. Data show that self-neglect makes up over half of all AR-tracked allegations (Figure 9). This aligns with the theory of change and the underlying philosophy of the dual-track system to match response approach to risk level.

> *"I really like the AR track for self-neglect so I can focus on helping and support…which feels more client-centered [than making a substantiation]."*
>
> **- Pilot County Caseworker**

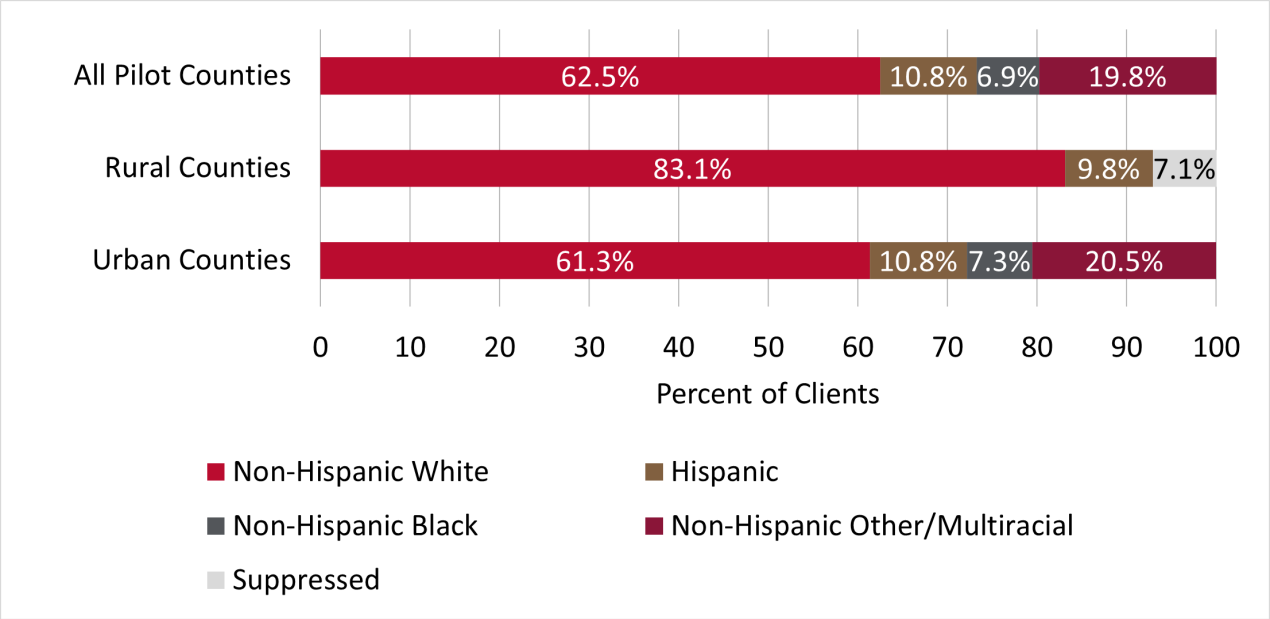
Figure 9. Breakdown of Alternative Response Allegations by Allegation Type



Insight 4: Understanding equitable reach in Alternative Response.

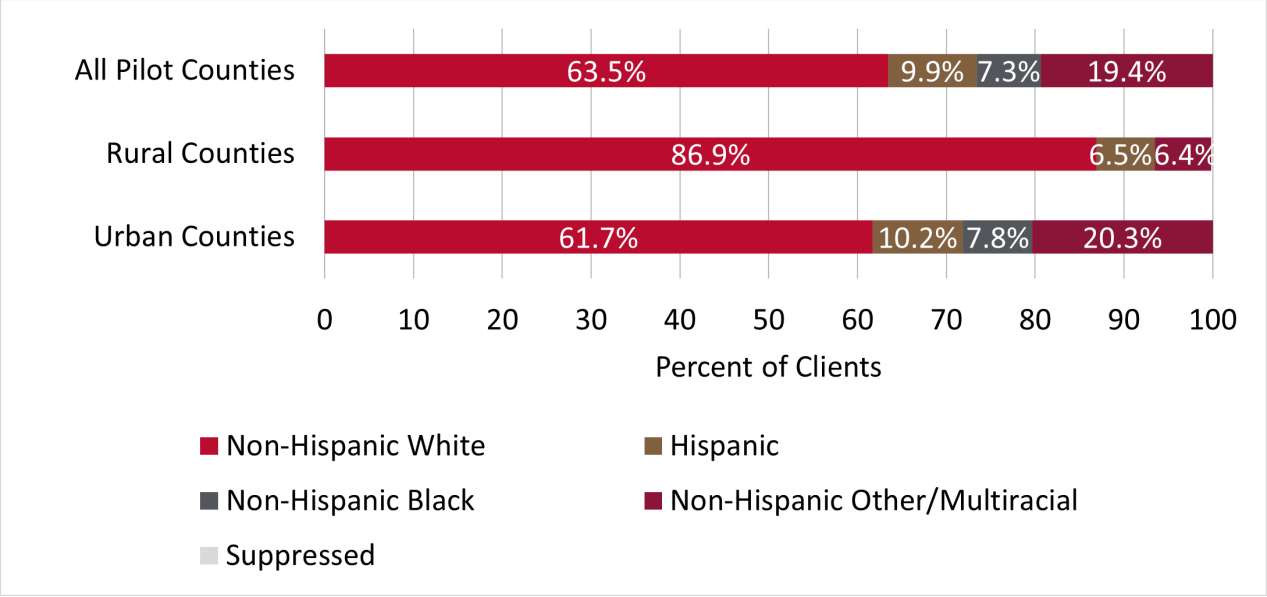
As part of the CDHS commitment to equity, diversity, and inclusion, it is important to understand whether the AR Pilot and the dual-track model are equitably reaching clients. Demographic data in the CAPS system is a known limitation, but recent data show that only 16.6% of race/ethnicity data is missing (down from as much as 50%). For clients with available race/ethnicity data, no disparities are observed in who the AR practice is reaching (Figures 10 and 11). Urban counties are significantly more likely than rural counties to serve clients identifying as Black, Hispanic, or multiracial ($p < 0.01$).

Figure 10. Race and Ethnicity of Clients Served by Adult Protective Services in the Pilot Counties (All Allegations)



Note: In rural counties, Non-Hispanic Black and Non-Hispanic Other/Multiracial are combined and suppressed due to privacy concerns.

Figure 11. Race and Ethnicity of Clients Served by Adult Protective Services in the Pilot Counties (Alternative Response Allegations)

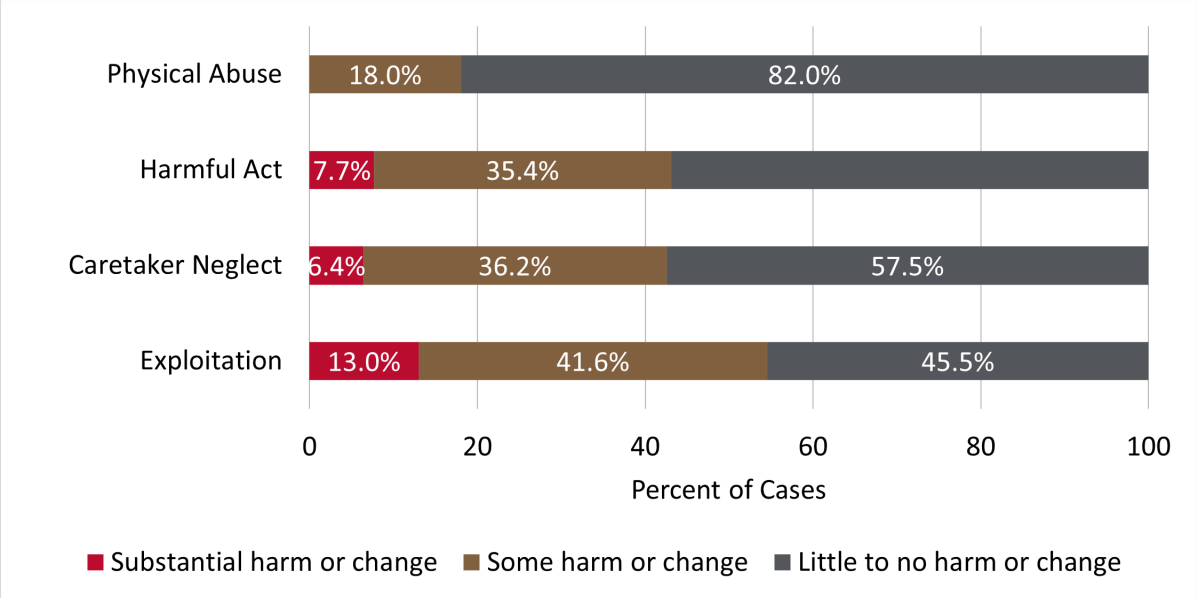


Insight 5: Conclusions for Alternative Response-tracked allegations reflect the low-risk nature of the Alternative Response track and signal that the Alternative Response track is being used appropriately.

> *"For me, it [AR] does feel better to not be doing substantiations on things that are educational opportunities, especially for caretaker neglect."*
>
> **- Pilot County Caseworker**

A hallmark feature of AR is that there is no finding for AR allegations, but rather, a conclusion. It is still important to understand the extent of impact that occurs in AR allegations and to assess whether this pattern aligns with the intended low-risk nature of the AR practice. This is because the dual-track model created a decision point for caseworkers to tailor their response to the level of risk. For allegations of mistreatment tracked to AR, conclusion data show an appropriate use of the AR track (Figure 12) across all mistreatment categories where a decision point is available. All cases of substantial impact are reviewed by the AR Specialist to confirm whether the track was appropriately used. Qualitative narratives show that in cases such as caretaker neglect involving a spouse, the AR track supports not only the client, but the relationship, and can stabilize the whole family by addressing root causes of involvement (e.g., husband having difficulty caring for wife with Alzheimer's; needs respite). Being able to support a client and their family through such best practices in social work in turn improves caseworker satisfaction—and satisfaction is a key driver of workforce retention in APS.

Figure 12. Extent of Harm for Alternative Response Allegations with a Conclusion



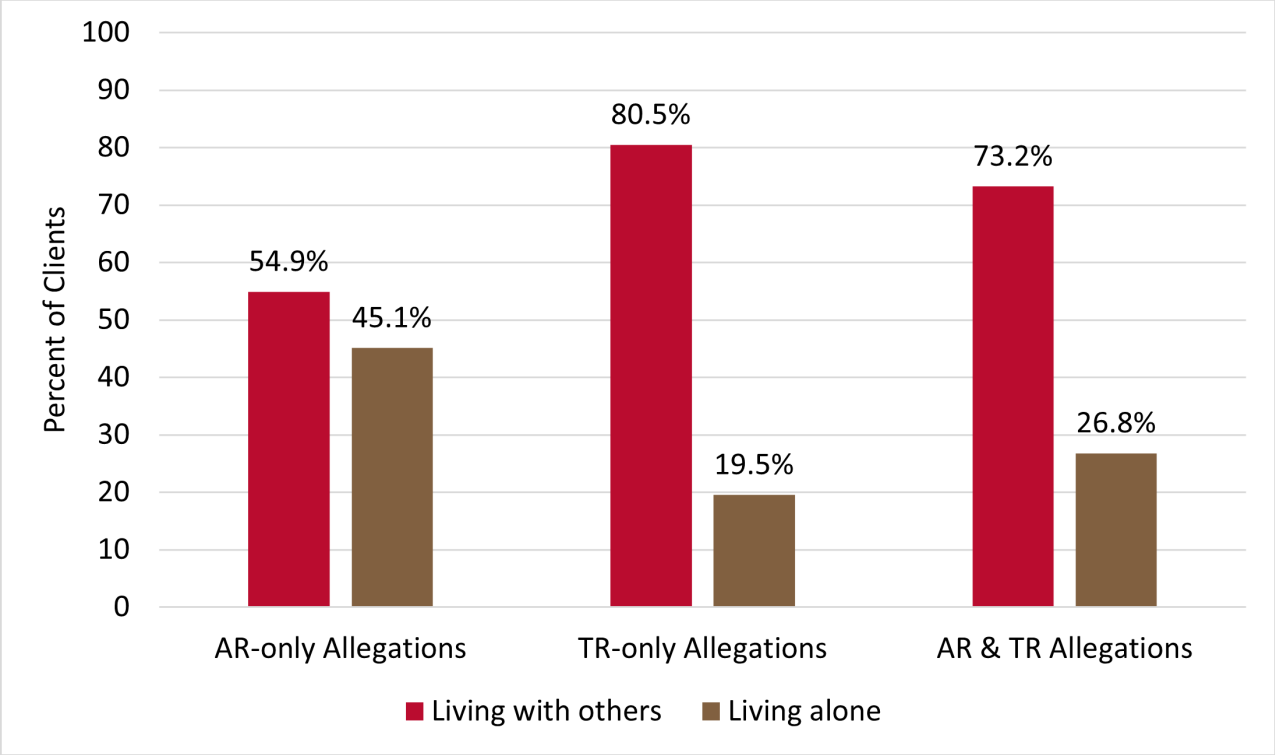
Note: There is a very small percentage of physical abuse allegations, with a substantial harm or change conclusion, tracked to AR; these are suppressed for privacy. To prevent identifying the exact number of cases, both remaining categories are rounded up.

Insight 6: Clients with only Alternative Response-tracked allegations are significantly more likely to live alone and have fewer support networks.

Caseworkers can collaborate with clients to strengthen their support networks and stabilize the client in the home to help prevent future involvement and escalation of mistreatment or self-neglect. Anticipatory practice guidance can be developed from a data-informed understanding of client situations at case start. Clients with only AR-tracked allegations are significantly ($p < 0.01$) more likely to live alone compared to clients with only TR-tracked allegations (Figure 13). Living alone status is a proxy for social isolation. Social isolation is heightened by a lack of support networks, which are significantly ($p < 0.01$) lower for clients with AR-only cases than for clients with TR-only cases (an average of 2.61 persons for AR vs. 2.91 persons for TR).

> *"I would say that [with AR], the family is more willing to engage with a care plan. Whereas in the past, if we had to say 'you're going to receive this letter, you're going to go on the 'registry' [CAPS Checks]. They were like, 'Get out of my house.' There was no engagement after that. I think Alternative Response has supported that willingness of clients and families to engage in care."*
>
> **- Pilot County Caseworker**

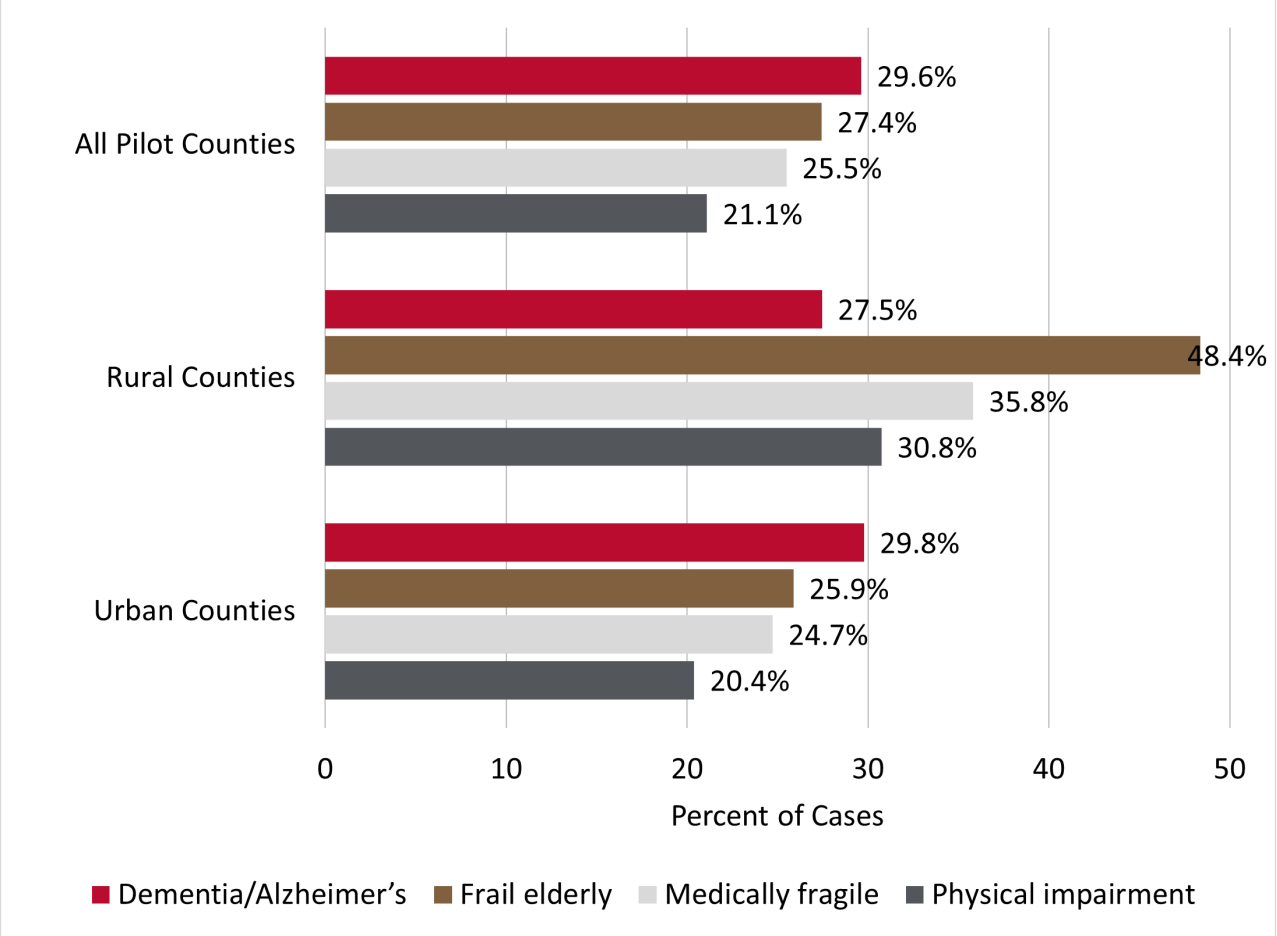
Figure 13. Clients Living Alone, by Track Type



Insight 7: Leading conditions vary with geography and reflect an aging population.

Understanding client conditions can further inform anticipatory guidance and identify who the AR practice is most appropriate for. Leading conditions also directly interact with social isolation (Insight 6) and can influence both why a client is involved with APS and how best to serve them. In cases with only AR-track allegations, the leading conditions are dementia/Alzheimer's, frail elderly, medically fragile, and physically impaired (Figure 14). The latter three conditions are higher in rural communities ($p < 0.05$). Rural communities have nearly double the incidence of frail elderly allegations. Qualitative narratives show the AR practice enables caseworkers to better help the client build support networks and connect with community-based supports, which are essential to managing these conditions long term and for a growing older adult population.

Figure 14. Leading Conditions for Alternative Response-Only Cases



Quasi-Experimental Design

A total of five main outcomes were analyzed using a rigorous QED, covering both system- and person-centered outcomes. Leading results include:

- **AR cases were 2.5% less likely to have a second screened-in case, showing effects of the AR practice on reducing repeat involvement.**
- **AR cases closed 5.63 days earlier, showing long-term system efficiencies created by the dual-track model.**
- **AR cases are significantly less likely to include allegations with ratings of severe or substantial extent of impact, showing the AR track is being used appropriately and not increasing harm.**
- **For *each* additional support, client refusal goes down by 1.10%, regardless of track, showing an opportunity to reduce social isolation long term and improve effectiveness of APS intervention.**

Inferential: Confirmatory Results

A total of five outcome categories were analyzed covering repeat involvement, case length, client engagement, and client safety. In addition, we analyzed severity level and extent of impact, which is not an outcome of the AR intervention, but rather, speaks to whether the AR practice is being appropriately applied and not increasing risk of harm. Together, these results speak to client- and system-level outcomes of the AR practice and opportunity for a dual-track model in Colorado APS.

"AR allows [caseworkers] to feel like they're doing more of the work that they signed up to do: social work to engage with people. It feels less like law enforcement and less punitive…To have the ability to tailor the response helps."

- Pilot County Supervisor

Summary Table for Each Outcome

Throughout this section, we provide a summary table near the beginning of each outcome group. This allows a quick way to understand results, as well as provides helpful information for decision making. All QED estimates displayed compare AR cases to AR-equivalent cases.

- **Green** indicates high practical significance, **orange** indicates a moderate level of practical significance, and **yellow** signals a cautious interpretation of practical significance. Practical significance is important because it speaks to a magnitude of change that matters in practice for APS and the clients they serve (i.e., bigger percent change on a meaningful outcome like re-entry).
- A **bold P-value** indicates the estimated difference is statistically significant at the 5% level.
- Practical and statistical significance together provide a powerful frame for evidence-based decision making (EBDM).

Outcome 1. The Alternative Response practice reduces repeat involvement.

For all three measures of repeat involvement, having at least one AR-tracked allegation on the initial case reduced the likelihood of having a subsequent screened-in case. The estimated differences in repeat involvement are all statistically significant. Qualitative narratives indicate the AR practice leads to more collaborative case planning, which in turn helps address root causes and reduces the need for repeat involvement for the same issue. This holds true for cases with only AR allegations, as well as cases with mixed AR and TR allegations, showing how the AR practice has positive ripple effects across case types. Table 3 summarizes the repeat involvement findings.

Making Meaning of Results

Reducing repeat involvement in APS has implications for long-term cost savings. There is a cost for every open case, largely driven by personnel time. By reducing repeat system use for the same client and issue, cost-savings can be realized and caseworker time re-directed to new or more acute cases.

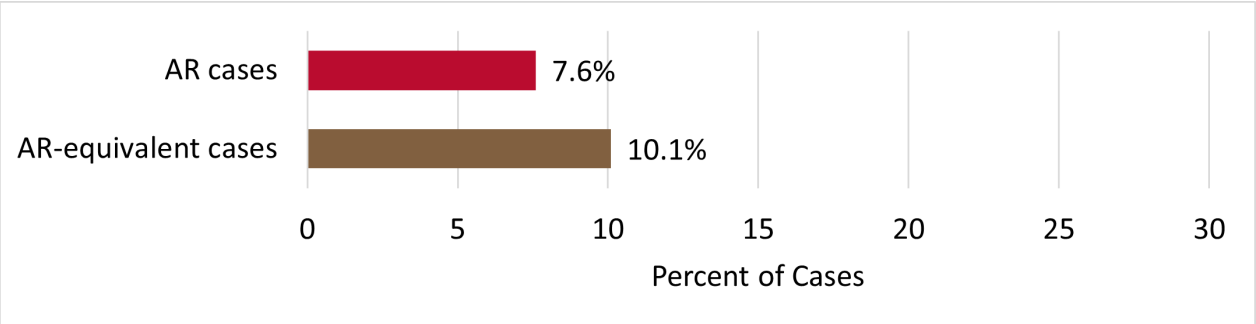
Table 3. Summary: Repeat Involvement Outcome Group

Measure Name	QED Estimate	P-Value	95% Confidence Interval	Sample Size
Overall repeat involvement	-2.50% points	**<0.01**	[-3.64, -1.35]	14,641
Repeat involvement: Self-neglect	-7.24% points	**<0.01**	[-9.07, -5.41]	6,095
Repeat involvement: Mistreatment	-11.79% points	**<0.01**	[-13.76, -9.83]	7,917

Outcome 1a. Repeat Involvement

Finding: Compared to equivalent cases, AR cases were 2.5 percentage points less likely to have another screened-in case within 6 months of closing ($p < 0.01$). The estimated repeat involvement rate for AR-equivalent cases was 10.1%, and the estimated repeat involvement rate for AR cases was 7.6% (Figure 15).

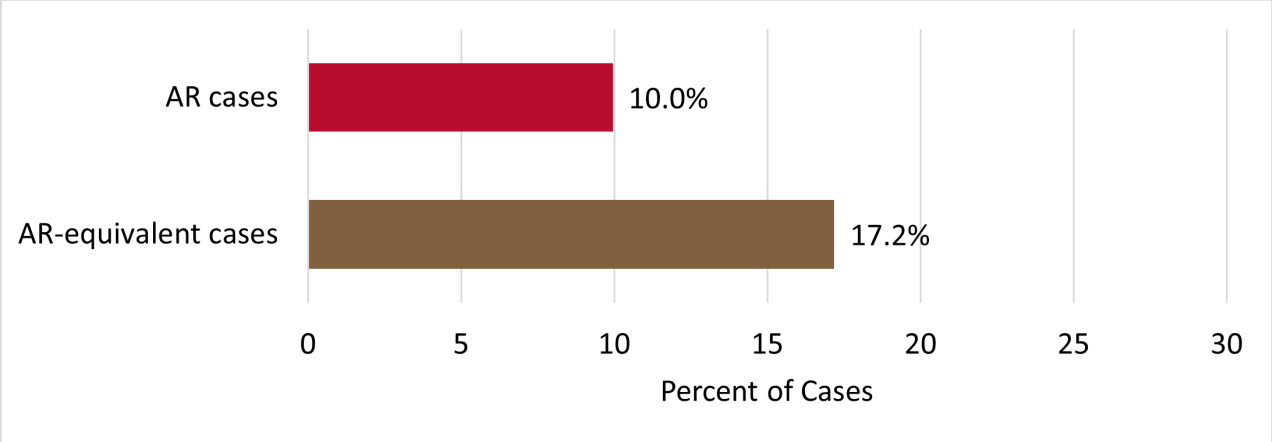
Figure 15. Comparison of Estimated Rates of Repeat Involvement for Alternative Response Cases and Alternative Response-Equivalent Cases



Outcome 1b. Repeat Self-Neglect

Finding: Compared to equivalent cases with a self-neglect allegation, AR cases with a self-neglect allegation were 7.2 percentage points less likely to have a second case with a self-neglect allegation ($p < 0.01$). The estimated probability of repeat self-neglect in AR-equivalent cases was 17.2%, and the estimated probability of repeat self-neglect in AR cases was 10.0%.

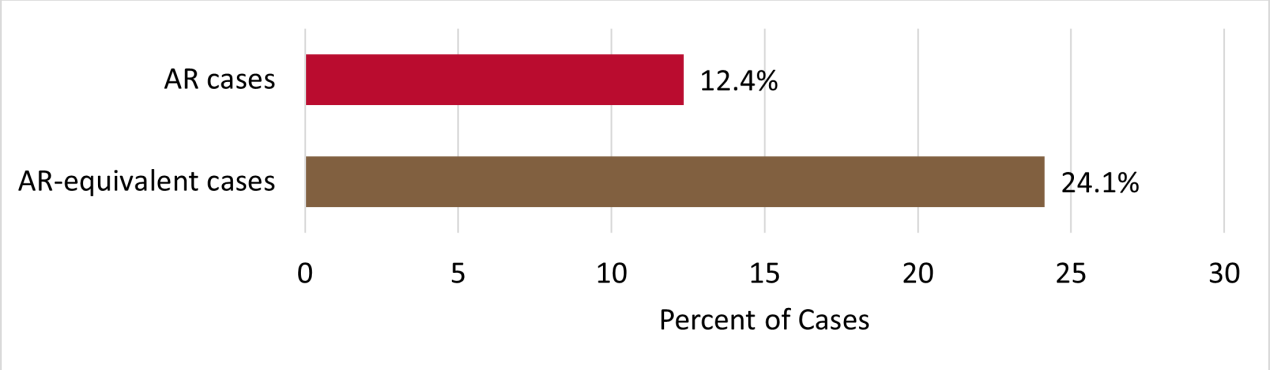
Figure 16. Comparison of Estimated Rates of Repeat Self-Neglect for Alternative Response Cases and Alternative Response-Equivalent Cases



Outcome 1c. Repeat Mistreatment

Finding: Compared to equivalent cases with a mistreatment allegation, AR cases with a mistreatment allegation were 11.8 percentage points less likely to have a second case with a mistreatment allegation ($p < 0.01$). The estimated probability of repeat mistreatment in AR-equivalent cases is 24.1%, and the estimated probability of repeat mistreatment in AR cases is 12.4% (Figure 17).

Figure 17. Comparison of Estimated Rates of Repeat Mistreatment for Alternative Response Cases and Alternative Response-Equivalent Cases



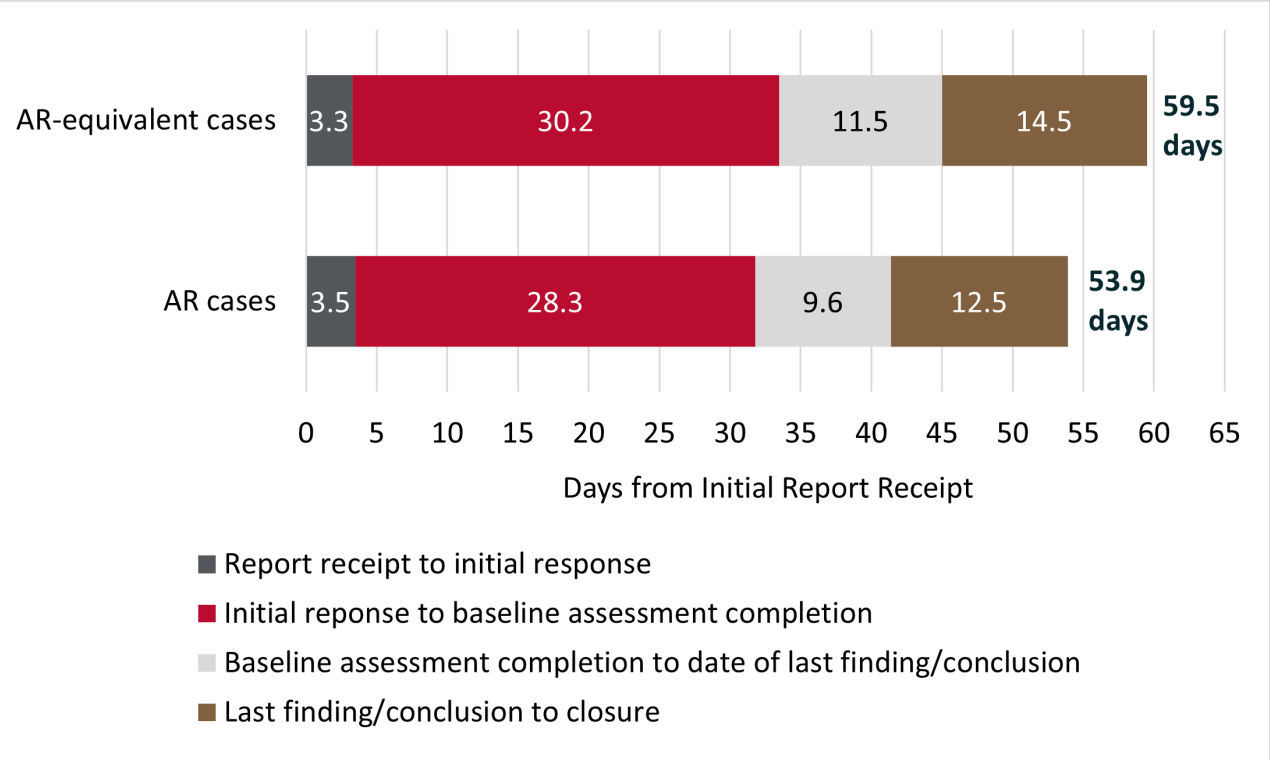
Outcome 2. The Alternative Response practice reduced case length.

Having at least one AR-track allegation shortened the time it took to manage and progress through a case. This reduction comes from quicker completion of client baseline assessments, fewer days in determining findings or conclusions, and a shorter window between the date of last finding or conclusion and case closure. Figure 18 displays these changes by showing where cases are getting shorter.

Making Meaning of Results

Reduced case length illustrates system efficiencies that are introduced when response can be tailored to level of risk through a dual-track model.

Figure 18. Breakdown of Number of Days from Report Receipt to Case Closure



After a report is received, the initial response date varies very little (usually 2 to 5 days, but sometimes up to a week). The time between initial caseworker response and the baseline assessment varies more with assessment due dates set for multiple weeks out, but the actual date depends on the caseworker and client coordinating. The period after the baseline assessment is a period in which the caseworker has more capacity to shorten case length.

Qualitative narratives indicate several factors influencing the shorter case timelines. These include improved rapport building using the AR practice that increases trust and can help uncover strengths and needs quicker, as well as a reduced documentation burden that is alleviated through the no finding requirement. The pace at which AR cases progressed later in 2024 compared to

earlier in the year also suggests improved implementation and increased caseworker efficiencies. The median number of days from initial response to completion of the baseline assessment in AR cases in the first three quarters of 2024 was 34 days, while in the fourth quarter it was 26 days. Table 4 summarizes the case length findings.

Table 4. Case Length Outcome Group

Measure Name	QED Estimate	P-Value	95% Confidence Interval	Sample Size
Receipt to close	-5.63 days	**<0.01**	[-7.22, -4.03]	14,641
Receipt to initial response	0.15 days	**<0.01**	[0.08, 0.21]	14,629
Receipt to assessment complete	-1.70 days	**<0.01**	[-2.32, -1.08]	13,148
Initial response to assessment complete	-1.72 days	**<0.01**	[-2.33, -1.11]	13,148
Receipt to last finding/conclusion	-3.61 days	**<0.01**	[-4.59, -2.63]	14,641
Initial response to last finding/ conclusion	-3.66 days	**<0.01**	[-4.64, -2.68]	14,640

Outcome 2a. Total Case Length

Finding: On average, AR cases closed 5.63 days earlier compared to AR-equivalent cases ($p < 0.01$). AR-equivalent cases closed 59.53 days after initial report receipt, and AR cases closed 53.90 days after initial report receipt.

Outcome 2b. Days Between Report Receipt and Initial Response

Finding: On average, an initial response occurred 0.15 days later with AR cases ($p < 0.01$) compared to AR-equivalent cases. Initial responses occurred 3.34 days after report receipt for AR-equivalent cases and 3.49 days after report receipt for AR cases.[vii]

Outcome 2c. Days Between Report Receipt and Baseline Assessment Completed

Finding: On average, for cases with a completed baseline assessment, the number of days between report receipt and assessment completion was 1.70 days shorter for AR cases ($p < 0.01$) compared to AR-equivalent cases. Baseline assessments were completed 33.51 days after report receipt for AR-equivalent cases and 31.81 days after report receipt for AR cases.

[vii] The sample size does not match with total case length because 17 cases either had initial response days prior to report receipt or had initial response days that were clearly outliers in violation of APS policy (e.g., more than 50 days). Upon inquiry, it was confirmed these that cases should not be included in analysis.

Outcome 2d. Days Between Initial Response and Baseline Assessment Completed

Finding: On average, for cases with a baseline assessment, the number of days between initial response and when the baseline assessment was completed was 1.72 days shorter for AR cases ($p < 0.01$) compared to AR-equivalent cases. The number of days between an initial response and the baseline assessment complete was 30.06 days for AR-equivalent cases and 28.33 days for AR cases.

Outcome 2e. Days Between Report Receipt and Last Finding or Conclusion

Finding: On average, the number of days between report receipt and last finding or conclusion was 3.61 days shorter for AR cases ($p < 0.01$) compared to AR-equivalent cases. The last finding or conclusion occurred after 45.00 days for AR-equivalent cases, and after 41.39 days for AR cases.

Outcome 2f. Days Between Initial Response and Last Finding or Conclusion

Finding: On average, the number of days between initial response and last finding or conclusion was 3.66 days shorter ($p < 0.01$) compared to equivalent cases. The number of days from initial response to last finding or conclusion occurred after 41.56 days for AR-equivalent cases, and 37.90 days for AR cases.

Outcome 3. The Alternative Response practice has the potential to improve client engagement.

For five of the six client engagement measures, the estimated difference between AR cases and AR-equivalent cases was not statistically significant. Furthermore, estimates were small in magnitude. Overall, there is scant evidence—when using CAPS data alone—that the implementation of the dual-track model significantly altered client behavior. The only significant result showed that AR cases had a slightly smaller number of caseworker notes on file. This can be attributed in part to the shortened length of AR cases and implementation efficiencies introduced with the AR practice. Table 5 summarizes the client engagement findings.

> *"I think the collaboration with the client makes me feel like I am helping rather than investigating. That's what AR does."*
>
> **- Pilot County Caseworker**

Behavioral changes and relationship-based changes, however, are often best measured by qualitative methods, as they are not easy to quantify in administrative systems. Qualitative narratives and the theory of change thus help shed light on client engagement outcomes. Caseworkers consistently report that a major value of the AR practice is the ability to engage collaboratively from case start. Further, they report that for self-neglect and low-risk mistreatment allegations, the AR

Making Meaning of Results

Client engagement is best measured with a mixed methods approach. Promoting best practices for initial scheduled response and client collaboration—such as Motivational Interviewing—is likely to improve client engagement and maximize positive outcomes for clients.

practice is more person-centered. Empowering clients and communicating that AR-tracked allegations do not result in a finding can foster a more productive relationship between caseworker and client. Conversely, empowerment can also mean a client chooses not to engage APS and leans into their right to autonomy. Client engagement findings from CAPS—taken in conjunction with qualitative narratives and the fidelity results on initial response—show the need to advance best practices in client engagement to maximize value of the AR practice. To this end, the Colorado Lab recommends training in Motivational Interviewing as a best practice for AR.

Table 5. Client Engagement Outcome Group

Measure Name	QED Estimate	P-Value	95% Confidence Interval	Sample Size
Refused contact	0.61% points	0.14	[-0.19, 1.41]	14,641
Refused contact: self-neglect only	0.75% points	0.33	[-0.76, 2.26]	5,046
Refused contact: mistreatment only	0.50% points	0.38	[-0.62, 1.62]	7,244
Refused all services	-1.45% points	0.17	[-3.53, 0.64]	5,996
Percent of services refused	-1.27% points	0.24	[-3.40, 0.86]	5,996
Services ineffective	0.09% points	0.75	[-0.47, 0.65]	5,996
Total number of case notes	-0.22 notes	**<0.01**	[-0.30, -0.13]	13,339
Case notes per month	0.01 notes	0.64	[-0.03, 0.04]	13,339
Number of interviews	<0.01 interviews	0.86	[-0.03, 0.03]	13,339

Outcome 3a. Client Refusing Contact

Finding: On average, the probability a case was closed due to a client refusing contact in AR cases was 0.61 percentage points higher compared to AR-equivalent cases (p = 0.137). The contact refusal rate for AR cases was 4.41% percent and 3.80% for AR-equivalent cases. This difference is estimated to be approximately 1.36 more contact refusals for every 200 cases.

We further investigated whether a client refusing contact differs by the type of allegation. First, consider the subsample of cases with only a single self-neglect allegation. On average, the probability a case closed due to a client refusing contact in AR cases was 0.75 percentage points higher compared to equivalent cases (p = 0.330). If we consider the subsample of cases with only mistreatment allegations, the parallel difference is 0.50 percentage points (p = 0.381).

In general, refusal rates were higher for self-neglect-only cases (6.93%) than for mistreatment-only cases (3.01%), but there are no statistically significant differences in refusal rates by track type. Qualitative narratives show the right to autonomy is often invoked in self-neglect cases, giving

context to this observed trend and showing the importance of scheduling an initial visit for collaborative engagement at case start.

Outcome 3b. Client Refusing All Services

Finding: On average, of the clients who were offered services, the probability a case was closed due to a client refusing all services in AR cases was 1.45 percentage points lower compared to AR-equivalent cases ($p = 0.173$). The all-service refusal for AR cases was 12.65% and 14.10% for AR-equivalent cases. This difference is estimated to be approximately 2.90 fewer all-service refusals for every 200 cases.

While not statistically significant, the estimated difference signals the potential for greater positive engagement for clients with AR cases, and this observation is consistent with qualitative findings that show caseworkers and clients experience improved collaboration through the AR practice. When pairing these findings with fidelity findings on initial response, a clear opportunity to strengthen practice emerges. Specifically, caseworkers should adopt best practices for client engagement that could improve initial response rates and in turn further reduce client refusal. Motivational Interviewing was identified as the leading best practice to this end.

Outcome 3c. Percent Of Services Client Refused

Finding: On average, of the clients who were offered services, the percentage of services a client refused for AR cases was 1.27 percentage points lower compared to AR-equivalent cases ($p = 0.243$). The service refusal rate for AR cases was 19.72% and 20.99% for AR-equivalent cases.

Outcome 3d. All Services Determined Ineffective

Finding: On average, of clients who were offered services, the probability a case closed due to all services being deemed ineffective for AR cases was 0.09 percentage points higher compared to AR-equivalent cases ($p = 0.745$). The rate at which all services were deemed ineffective for AR cases was 1.01%, and 0.91% for AR-equivalent cases.

> *"That [collaboration] is an important aspect of AR. We are not judging, we are helping and supporting."*
>
> **- Pilot County Caseworker**

Outcome 3e. Total Number Of Client Communications

Finding: On average, for cases that had a communication on file, AR cases had 0.22 fewer communications compared to AR-equivalent cases ($p < 0.01$). AR cases had 2.42 communications and AR-equivalent cases had 2.64 communications over the life of the case.

The difference in number of communications came entirely from AR cases having 0.22 fewer monthly contacts ($p < 0.01$). There is no difference in the number of interviews (<0.01; $p = 0.858$). This aligns with AR cases being shorter overall, being of lower risk, and generating a positive rapport so fewer monthly contacts are required, while at the same time not sacrificing interviews and the integrity of the outreach. For example, an AR case with a caretaker neglect allegation may

not require forensic evidence, so it would be shorter in length and have fewer documented communications in the case file, but in-person time through interviews remains the same.

Outcome 3f. Average Number Of Client Communications

Finding: On average, for cases that had a communication on file, AR cases had 0.01 more communications compared to AR-equivalent cases ($p = 0.604$). AR cases had 1.40 communications per month, and AR-equivalent cases had 1.39 communications per month.

Total number of client communications and average number of client communications illustrate appropriate case planning by caseworkers.

Outcome 4. The Client Safety and Risk Assessment cannot be used to make conclusive findings about safety.

Making Meaning of Results

AR is restricted to allegations of self-neglect and low-risk mistreatment. By definition, that creates higher safety at baseline. This requires APS intervention to be tailored and precise; combined with the high baseline safety scores, change may not be detectable using the assessment alone

The client safety assessment is a previously validated tool designed to help guide service planning and monitor outcomes over time. In 2024, APS contracted with the Colorado Health Institute to review the assessment and make recommendations for improvement. The current evaluation aimed to use safety and risk scores as proxies for the effectiveness of the AR practice on improving client safety; however, results showed significant limitations to this analysis. For transparency purposes, results are still specified below. For all specifications, scores are estimated controlling for the initial baseline score.

By construction, only 19.91% of safety scores are below 90 out of a possible 100. In other words, only 19.91% of cases have the potential for a 10 or greater point improvement. The extreme compression of the safety score distribution meant very limited room for improvement could be detected.[viii] This is true for both AR and AR-equivalent cases, but more so for AR cases. The low-risk nature of AR inherently means safety will be high to begin with. Qualitative narratives backed this, showing how clients were more likely to need a few services or supports put in place, rather than large sweeping intervention.

> *"And with the lower risk nature [of AR], I think we are more service based. We're going into this low-risk case and saying, maybe they just need home health or Medicaid, that's it."*
>
> **- Pilot County Caseworker**

[ix] Alternative specifications were checked based on a 90 points cutoff in the safety score distribution. For example, we estimated separate specification for cases above and below 90 points, as well as a specification that interacted baseline score with the AR designation indicator. The results of these variations do not differ substantively for the results reported in the text.

Overall, the analysis suggests that, as currently measured, aggregate safety and risk scores may not be the most appropriate measures to understand changes in client safety and risk. Table 6 summarizes the safety and risk score findings.

Table 6. Client Safety and Risk Score Outcome Group

Measure Name	QED Estimate	P-Value	95% Confidence Interval	Sample Size
Safety score	0.01 points	0.95	[-0.27, 0.29]	4,555
Safety score: Self-neglect only	0.18 points	0.51	[-0.35, 0.70]	1,866
Safety score: Mistreatment only	0.16 points	0.32	[-0.13, 0.47]	1,598
Risk score	-0.14 points	0.47	[-0.52, 0.24]	4,555
Risk score: Self-neglect only	-0.54 points	0.07	[-1.12, 0.04]	1,866
Risk score: Mistreatment only	0.49 points	0.08	[-0.05, 1.03]	1,598

Outcome 4a. Client Safety Score

Finding: On average, for cases with a completed baseline assessment and controlling for score at baseline, the aggregate safety score of AR cases was 0.01 points higher compared to AR-equivalent cases (p = 0.946).

For self-neglect-only cases, the safety score of AR cases was 0.17 points higher compared to AR-equivalent cases (p = 0.511). For mistreatment-only cases, the safety score of AR cases was 0.16 points higher compared to AR-equivalent cases (p = 0.318).

Outcome 4b. Client Risk Score

Finding: On average, for cases with a completed baseline assessment and controlling for score at baseline, the aggregate risk score of AR cases was 0.14 points lower compared to AR-equivalent cases (p = 0.466).

For self-neglect-only cases, the risk score of AR cases was 0.54 points lower compared to AR-equivalent cases (p = 0.070). For mistreatment-only cases, the risk score of AR cases was 0.49 points higher compared to AR-equivalent cases (p = 0.078). Statistically significant at the 10% level, the estimates for both the self-neglect-only and mistreatment-only sample are very small in magnitude. Similar to what was found with safety scores, the lack of an effect is likely due to the compressed risk score distribution.

Outcome 5. Conclusions for AR-tracked allegations reflect the low-risk nature of the AR track and signal that the AR track is being used appropriately.

Making Meaning of Results

By rule, "low-risk" was intentionally not defined to allow for caseworker professional discretion and tailoring to client context. This introduced decision making points for some mistreatment allegations that could be tracked to either AR or TR (e.g., caretaker neglect). Results in this section show caseworkers are applying sound decision making skills and the AR practice is not increasing harm.

Outcome 5 differs from the previous outcomes in that it speaks to strong implementation decisions, rather than an effect of the intervention itself. Measuring severity of harm and extent of impact helps assess whether the AR practice is being appropriately applied (i.e., reserved for low-risk cases) and is not increasing risk or harm. Differences in severity and extent of impact are explored for mistreatment allegations in general and by specific mistreatment type.

The dual-track model created a decision point for caseworkers to tailor their responses to the level of risk. The results demonstrate that the AR track is being applied appropriately and caseworkers are using sound decision making skills. In all instances measured in Outcome 5, cases with AR-track allegations are significantly less likely to include allegations with ratings of "severe" or "substantial" which reflects the AR track being reserved for low-risk cases.

Figure 19 compares the percentage of cases with a *severe* or *substantial* rating allegation between AR and AR-equivalent cases. Table 7 summarizes the severity level and extent of impact findings.

Figure 19. Severity of Impact

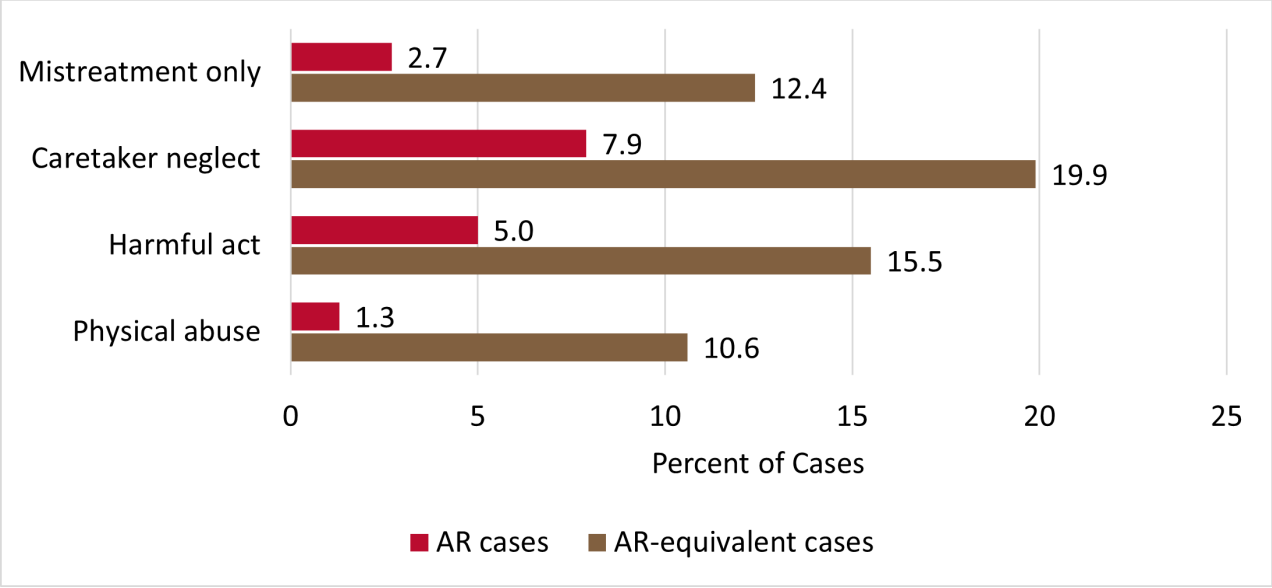


Table 7. Severity Level and Extent of Impact Outcome Group

Measure Name	QED Estimate	P-Value	95% Confidence Interval	Sample Size
Severe/substantial rating	-10.53% points	**<0.01**	[-13.27, -7.79]	1,884
Severe/substantial: Mistreatment only	-9.74% points	**<0.01**	[-12.21, -7.27]	1,640
Severe/substantial: Caretaker neglect	-12.01% points	**<0.01**	[-17.89, -6.13]	764
Severe/substantial: Harmful act	-10.45% points	**0.02**	[-18.99, -1.90]	234
Severe/substantial: Physical abuse	-9.29% points	**<0.01**	[-12.69, -5.90]	879

Outcome 5a. Severity Level and Extent of Impact for Cases with Mistreatment Allegations

Finding: On average, for cases with confirmed mistreatment allegations, AR cases were 10.53 percentage points less likely to have a *severe* or *substantial* rated allegation versus AR-equivalent cases ($p < 0.01$). The probability an AR case had a *severe* or *substantial* rating was 3.89%, and the probability an AR-equivalent case had a severe or substantial rating was 14.42%.

Outcome 5b. Severity Level and Extent of Impact for Cases with Only Mistreatment Allegations

Finding: On average, for cases with confirmed allegations and only mistreatment allegations, AR cases were 9.74 percentage points less likely to have a *severe* or *substantial* rated allegation versus AR-equivalent cases ($p < 0.01$). The probability an AR case had a *severe* or *substantial* rating was 2.67%, and the probability an AR-equivalent case had a *severe* or *substantial* rating was 12.41%.

Outcome 5c. Severity Level and Extent of Impact for Cases with Caretaker Neglect Allegations

On average, for cases with confirmed caretaker neglect allegations, AR cases were 12.01 percentage points less likely to have a *severe* or *substantial* rated allegation versus AR-equivalent cases ($p < 0.01$). The probability an AR case had a *severe* or *substantial* rating was 7.90%, and the probability for an AR-equivalent case had a *severe* or *substantial* rating was 19.91%.

> *"Our AR tracked allegations for caretaker neglect, a lot of times that is a husband who is struggling…and likely they just need help. They just need Medicaid, or they just need respite."*
>
> **- Pilot County Caseworker**

Outcome 5d. Severity Level and Extent of Impact for Cases with Harmful Act Allegations

On average, for cases with confirmed harmful act allegations, AR cases were 10.45 percentage points less likely to have a *severe* or *substantial* rated allegation versus AR-equivalent cases ($p < 0.05$). The probability an AR case had a *severe* or *substantial* rating was 5.03%, and the probability an AR-equivalent case had a *severe* or *substantial* rating was 15.48%.

Outcome 5e. Severity Level and Extent of Impact for Cases with Physical Abuse Allegations

On average, for cases with confirmed physical abuse allegations, AR cases were 9.29 percentage points less likely to have a *severe* or *substantial* rated allegation versus AR-equivalent cases ($p < 0.01$). The probability an AR case had a *severe* or *substantial* rating was 1.31%, and the probability an AR-equivalent case had a *severe* or *substantial* rating was 10.61%.

Outcome 6. Differences in the composition of Alternative Response and Traditional Response allegations changed the distribution of case closure reasons.

A case can be closed for 12 reasons. Two reasons were previously analyzed as part of client engagement measures—contact refusal and service refusal. Together, these reasons only account for 17% of cases in the analytic sample. Determinations of client not at risk or no identified needs account for 47.1% of cases, while intervention complete accounts for 27.5% of cases. Of the remaining "other" closure reasons (8.3%), most are closed due to death of the client.

Making Meaning of Results

The low-risk nature of AR means more clients will have no identified needs or the acuity level will be lower. This is why looking at case closure reasons and other outcomes, in combination, is important to understand the full picture of the AR practice.

Both reasons of "client not at risk "and "no identified needs" indicate that AR cases were more likely to be closed (compared to equivalent cases in the pre-pilot period) because intervention was not required. As such, there is a smaller pool by which "intervention complete" could take place. These trends are consistent with the conclusions from Outcome 5 which showed the AR track is being used to handle low-risk allegations of mistreatment and all self-neglect cases. Figure 20 compares the percentage of cases by case closure reason for AR and AR-equivalent cases. Table 8 summarizes the case closure reason findings.

Figure 20. Case Closure Reasons

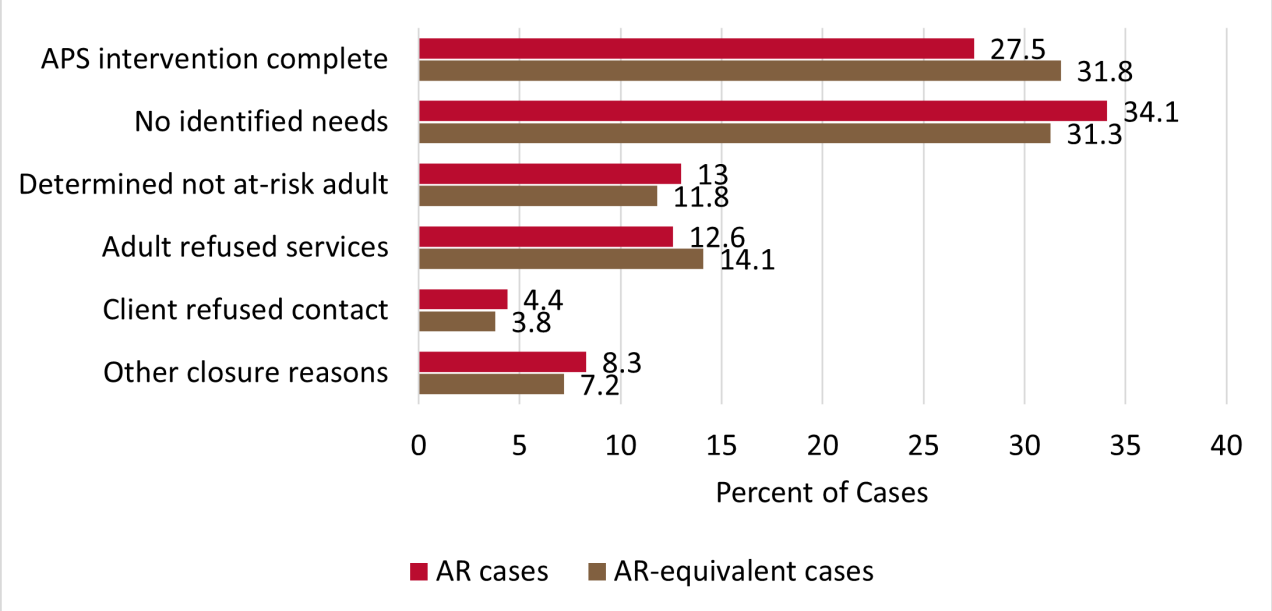


Table 8. Case Closure Reason Outcome Group

Measure Name	QED Estimate	P-Value	95% Confidence Interval	Sample Size
Intervention complete	-4.33% points	**<0.01**	[-6.14, -2.51]	14,641
Determined not at risk	1.28% points	0.06	[-0.04, 2.59]	14,641
No identified needs	2.80% points	**<0.01**	[0.01, 0.05]	14,641
Refused contact	0.61% points	0.14	[-0.19, 1.41]	14,641
Refused all services	-1.45% points	0.17	[-3.53, 0.64]	5,996
Services ineffective	0.09% points	0.75	[-0.47, 0.65]	5,996
Other closure reasons*	1.00% points	0.19	[-0.51, 2.50]	14,641

Note: "Other" closure reasons include incarceration, death, moved, self-neglect, mitigated prior, services not available, unable to locate.

Outcome 6a. Case Closed Because APS Intervention Was Completed

Finding: On average, the probability a case closed because the intervention was completed for AR cases was 4.33 percentage points lower compared to AR-equivalent cases ($p < 0.01$). The rate of this closure reason for AR cases was 27.46%, and the rate for AR-equivalent cases was 31.78%.

Outcome 6b. Case Closed Because Client Was Determined Not at Risk

On average, the probability a case closed because the client was determined not to be at risk for AR cases was 1.28 percentage points higher compared to AR-equivalent cases ($p < 0.10$). The rate of this closure reason for AR cases was 13.03%, and the rate for AR-equivalent cases was 11.75%.

Outcome 6c. Case Closed Because No Needs Were Identified

On average, the probability a case closed because the caseworker did not identify needs for AR cases was 2.80 percentage points higher compared to AR-equivalent cases ($p < 0.01$). The rate of this closure reason for AR cases was 34.13%, and the rate for AR-equivalent cases was 31.34%.

Inferential: Exploratory Results

Exploratory analyses dive deeper into the role of client support networks, track changes, and contact refusal. In the client support network analysis, we estimate the impact of more supports by track assignment. In the analysis of track changes, as well as contact refusal and case length, we report descriptive statistics.

> *"I've had people be able to involve their support network in that initial meeting which not only is helpful for them, but also helpful for me."*
>
> **- Pilot County Caseworker**

Support Networks Increase Engagement and Case Length

Making Meaning of Results

Across APS, caseworkers should heavily invest in involving client support networks to improve collaborative engagement, increase client agreement to a care plan, and accelerate the positive impacts observed with AR.

The size of a client's network could impact their tendency to refuse contact. This may interact with the AR philosophy of collaborative engagement and provides insight into the role of support networks as a protective factor to accelerate the positive impacts of APS intervention. Similarly, the size of a client's support network could have different impacts on case length. As such, exploratory analyses of support networks relies on estimating any differential impacts by interacting the number of support networks with the variable indicating the presence of an AR-track allegation.

Additional supports decrease the probability of client refusal. Refusal goes down by 1.10 percentage points per *each* additional support ($p < 0.01$). Importantly, this impact does not vary by track type; there is no statistically significant difference in client refusal between AR cases and AR-equivalent cases by the number of supports ($p = 0.995$). Qualitative narratives indicate this result may reflect improved trust that comes when support networks are included in the APS response and collaborative case planning with clients.

Additional supports do lengthen cases, however, adding 7.37 days to a case ($p < 0.01$). There is no statistically significant difference in case length between AR cases and AR-equivalent cases by the number of supports ($p = 0.216$). Adding supports to a case requires the caseworker make additional efforts to establish contact with the support and coordinate care. These results highlight a trade-off inherent in widening a client's support network. Lengthier case timelines are worth it if building support network capacity results in sustained safety and improved client health in the long term. Engaging support networks does not necessarily increase the burden on a caseworker, even if overall case length is increased, and a result of this upfront investment is the costs saving that can be realized via a lower likelihood of reinvolvement in APS.

Track Changes are Uncommon, Appreciated, and Increase Case Length

In the pilot period, an allegation can change tracks. Track changes are very uncommon though, with only 199 cases experiencing a track change in the analytic sample. There were a total 240 track changes (combined AR to TR and TR to AR), and most track changes were TR to AR (79.58%), illustrating a more conservative approach initially when handling a new case. Track changes are most common in cases with caretaker neglect (130 cases), exploitation (75 cases), or physical abuse allegations (62 cases). Cases with a single change take approximately 9 days longer compared to cases with no track change.

Qualitative narratives show that caseworkers appreciate the option to change tracks as new details emerge in the case and that a longer case length is a worthwhile trade-off for flexibility. The low frequency of track changes supports the solid decision making by counties during initial response, especially when paired with fidelity data on initial track assignment. Decision making during

Review, Evaluate and Direct (RED) team is critical to sound decision making and best practices in RED team should be continuously promoted. Allowing for track changes in any direction, and without limit, structurally supports the AR philosophy of tailoring response to level of risk, while not increasing workload.

Contact Refusal and Case Length Were Higher in Rural Counties

On average, 3.85% of urban clients refused contact, while 7.21% of rural clients refused contact among pilot counties. Higher refusal rates likely link back to lower rates of a scheduled initial visit (a key driver of change underlying the AR philosophy) in rural counties, as well as higher rates of frail elderly clients who may have a cultural mindset of autonomy in aging. Encouraging best practices for engagement, such as Motivational Interviewing, is thus vital in rural counties. Importantly, there was no statistically significant difference in the contact refusal rates between AR and AR-equivalent cases by rural or urban setting ($p = 0.295$) among the QED sample. This is not surprising since the matching process intentionally created equivalence along county, which is a strong proxy for rural-urban status. On average, rural cases took 10 days longer, with no statistically significant difference in case length between AR and AR-equivalent cases by rural or urban setting ($p = 0.551$).

Legal Authority

The Colorado Office of Public Guardianship (OPG) is a public agency established by the Colorado General Assembly in 2017 within the Judicial Department. In 2023, SB23-064 extended the office indefinitely and requires the office to operate in every judicial district in the state by December 31, 2030. The Colorado OPG provides guardianship services for indigent and incapacitated adults when other guardianship possibilities are exhausted. Given this, the OPG and APS may sometimes hold interest in a shared population. As such, we provide a spotlight on county-held legal authority.

Spotlight on Legal Authority

This analysis includes both county and non-county legal authority. However, among APS cases in the 15 pilot counties, **a county-held legal authority is very rare**. Of the 9,790 screened-in cases opened and closed in the pilot period, **only 14 cases had County Legal Authority** listed as a support. The type of authority is guardianship—temporary or permanent. As such, **results below largely reflect non-county held legal authority**.

For the 1,042 cases with only AR-tracked allegations where a support had *any* legal authority (conservatorship, guardianship, medical proxy, representative payee, power of attorney), 43.28% closed because no needs were identified, 34.64% closed because the intervention was completed, 6.05% closed because the client was not at risk, and 5.76% closed due to death.

Nearly 45% of cases had at least one professional support and one family support. The types of professionals included Home Health Agency/Providers, Social Work Practitioners, Nurses, and Law Enforcement. The types of family supports included aunts, brothers, and sons. Cases with one or more professional supports is similar to cases with one or more family supports at ~72%.

Recommendations

> **The effective repeal date of SB21-118 is July 1, 2027. Prior to this date, CDHS must make recommendations to the General Assembly on the future of AR in Colorado.**
>
> **Based on favorable findings from the 2-year outcomes evaluation, the AR practice should be recommended for statewide scaling with adequate resourcing for strong implementation.**

According to SB21-118, participating pilot counties may continue to implement a dual-track model through June 30, 2027, as the effective repeal date of the legislation is July 1, 2027. Final reporting to the General Assembly is due in the January 2026 legislative session by CDHS. Findings from the 2-year outcome evaluation show the AR practice is having a positive impact on at-risk adults in Colorado by reducing re-entry and case length through collaborative engagement. Coupled with qualitative feedback showing strong support for the AR practice, our recommendations focus on what it will take to expand the dual-track model statewide with integrity, to ensure all at-risk adults in Colorado can benefit from this APS innovation.

Based on favorable evaluation findings alongside support by implementing partners, the AR practice should be recommended for statewide scaling.

Areas for Priority Rule Change

Rule promulgation will need to happen to establish the regulations governing a statewide dual-track model, such as criteria governing track assignment and timelines for APS intervention. During rule promulgation, we recommend revisiting the initial response timeline for low-risk cases of mistreatment and self-neglect, to allow caseworkers and clients adequate time to establish and document the collaborative relationship and ensure the positive effects of the AR practice are maximized. Rules should also make clear what counts as initial contact and expand methods to include in-person as well as text, phone call, and other communication mediums such as email.

Phased Rollout

Innovations that become permanent practice are benefited by a phased rollout that can ensure the state and counties have adequate time to build the structural and cultural conditions necessary for success. The state needs time to promulgate rules and build a robust training system based on learnings from pilot implementation; county staff need time to engage training on rules and best practices that support effectiveness; and both counties and the state need space to develop policies and processes for implementation. A phased rollout would involve CDHS setting a feasible timeline (e.g., 3 to 5 years) for counties to fully opt-in to a dual-track model, and then counties assessing their readiness and the preparation it will take for them to successfully engage the AR practice. A phased rollout (versus a global one-time adoption across the state) is important not just

to ensure technical aspects can be effectively adopted, but also to inspire culture change by ensuring authentic state-county partnership and reducing caseworker overload.

Adequate Resourcing

The AR practice represents collaboration at its core—and that collaboration is true for caseworkers to clients, as well as between counties and the state. Implementing a dual-track model with fidelity requires adequate resourcing at CDHS and at the county level. Specifically, at the state level, the AR Specialist position should become a permanent position at CDHS, dedicated funds are needed to establish a robust statewide training approach, and fidelity monitoring should be absorbed by the ARD team to continuously drive quality improvement and performance management. At the county level, counties should assess their workforce for readiness to adopt a dual-track model; ensure caseworkers and supervisors have the skills and attitude (e.g., learning mindset) necessary for success; and provide professional development opportunities where needed. A leading professional development opportunity is Motivational Interviewing.

Advancing Partnerships for the Aging Population

While evidence building for AR was focused firstly on APS response, evaluation results also have implications for the aging population across units at CDHS, including the State Unit on Aging. As Colorado and the nation grapple with how best to care for this rapidly growing community, it is imperative that prevention and intervention services reflect their unique conditions and challenges. For example, in-depth data on self-neglect generated through this pilot can inform best practices for aging adults and clearly identify that more resources are needed across program levels and with cross-system investment. For example, leveraging Prop 123 funding to support affordable housing investments that match the needs of the older adult population. The state's first-ever Multi-Sector Plan on Aging provides a prime roadmap to advance partnerships and leverage results of the 2-year outcomes evaluation toward statewide infrastructure in system- and community-level services and supports for older adults.

Within CDHS, a prime area for partnership is in outreach and service provision among shared populations across the State Unit on Aging and APS. This collaboration can be facilitated by the newly developed Aging and Adult Protective Services organizational theory of change, created in collaboration with the Colorado Lab as the state's Coordinating Entity for Evidence-Based Decision Making. For example, supports provided by the Area Agencies on Aging can act to prevent initial involvement in APS, help stabilize clients involved in APS when a case is opened, and create continuity into the future once an APS case is closed in order to help reduce repeat involvement.

Applying the Evidence-Based Decision Making Approach

Colorado is recognized nationally as a leader in using evidence to strengthen budget decisions, inform legislative policy, and deliver better results for Colorado communities. To accelerate progress, the Colorado Lab released a 5-year vision for Colorado's approach to EBDM across branches of state government. The vision was developed in collaboration with members of the executive and legislative branches, including Joint Budget Committee (JBC) members and staff; the

Governor's Offices of State Planning and Budgeting (OSPB), Operations, and Information Technology; General Assembly members; leaders from several state agencies; and the Colorado Evidence-Based Policy Collaborative. The vision aims to align decision makers across government in their approach to using and building research evidence to drive effective state investments, continuous improvement, innovation, and improved outcomes.

Colorado's EBDM approach (Figure 21) exists at the intersection of the best available research evidence, decision-makers' expertise, and community needs and implementation context. This approach recognizes that research evidence is not the only contributing factor to policy and budget decisions. EBDM provides a leading-edge framework to activate results from the 2-year rigorous evaluation of the AR practice, including for multiple cross-system use cases (e.g., meeting statutory requirements of House Bill 24-1428: Evidence in Budgeting) and populations (e.g., informing prevention services for the aging population). State and county APS—and their partners—are encouraged to adopt the EBDM framework and continuously use pilot evaluation findings toward quality improvement and achieving sustained change among Colorado communities.

Figure 21. Evidence-Based Decision Making Approach

Finally, a key component of the EBDM approach is not only using existing evidence for action but also building evidence and promoting learning communities. The evaluation showed several areas of the CAPS data system that could be enhanced to improve reliability and completeness of data (e.g., reduce missing data), and help APS answer critical questions of the effects of their programming, and for whom and under what conditions.

Looking Ahead

The Colorado Lab will support CDHS in final reporting to the General Assembly. We will also work closely with the state and pilot counties to move results into action and activate recommendations available from this rigorous evaluation.

> *"AR allows the person on the other end [the client] to be more open to services, whereas if we just show up there [without scheduling], with no warning, we put that block up of like, 'No, you guys are going to make me leave my home or you're going to make me do things I don't want to do.' AR lets us focus on the relationship and be more person-centered."*
>
> **- Pilot County Caseworker**

Appendix A: Description of Fidelity of Implementation Indicators

Fidelity Indicator	Data Type	Data Source & Definition	Not Met	Approaching	Met
1. Initial Track Assignment Evidence that initial track assignment is consistently and appropriately applied by leads and supervisors.	Quantitative	Colorado Adult Protective Services (CAPS): How are considerations being used and which track are cases assigned to. Identify instances in which either 1) a case is tracked to Alternative Response (AR) and all considerations are marked as "yes" for all allegations (indicating allegations may not be low risk) or (2) a case is tracked to Traditional Response (TR) and all considerations are marked "no" for all allegations (indicating allegations are likely to be low risk).	<70% of allegations or cases	70%–90% of allegations or cases	>90% of allegations or cases
	Qualitative	AR Pilot project learning log. CAPS narrative fields for considerations.	Narrative examples indicate very inconsistent and disparate use of considerations fields; considerations are not usually appropriately applied.	Narrative examples indicate mostly consistent and equitable use of considerations fields; considerations are usually appropriately applied.	Narrative examples indicate consistent and equitable use of consideration fields; considerations are appropriately applied.
2. Initial Response Evidence that the option to schedule an initial visit is being exercised consistently and appropriately.	Quantitative	CAPS: To understand what proportion of cases with only AR-tracked allegations have an initial visit that is scheduled vs. unannounced; look at data by allegation type for the case.	<50% of cases	50%– 0% of cases	>70% of cases
	Qualitative	AR pilot project learning log. CAPS: When caseworkers choose to do an unscheduled visit, do they have a rationale for why and does the rationale meet standards for appropriate use? I.e., It is appropriate to use a non-scheduled visit for cases with only allegation(s) of self-neglect that meet the criteria for an emergency (visit within 24 hours of report).	Narrative examples indicate inconsistent and disparate rationale for an unscheduled visit; rationale is not usually appropriate.	Narrative examples indicate mostly consistent and equitable rationale for an unscheduled visit; rationale is usually appropriate.	Narrative examples indicate consistent and equitable rationale for an unscheduled visit; rationale is appropriate.

Fidelity Indicator	Data Type	Data Source & Definition	Not Met	Approaching	Met
3. Track Changes Evidence that use of the track change option is being exercised judiciously and consistently.	Quantitative	CAPS: To understand whether 1) more than two track changes occur within an allegation (regardless of direction); and 2) final track change occurs at the end of an investigation (i.e., compare date of final track change to "Date of Finding").	<50% of allegations	50%–70% of allegations	>70% of allegations
	Qualitative	AR pilot project learning log. CAPS narrative fields for justification of track change. Use these fields to understand whether a track change is justified by a change in risk level.	Narrative examples indicate inconsistent, disparate, and low to no justification of track change related to change in risk level.	Narrative examples indicate mostly consistent, equitable, and moderate justification of track change related to change in risk level.	Narrative examples indicate consistent, equitable, and extensive justification of track change related to change in risk level.
4. Investigation and Conclusion Evidence that a determination of a conclusion is being consistently and robustly applied (for closed cases).	Quantitative	Administrative Review Division (ARD) Assessment/ Investigation Measure 19: Were the findings/conclusions supported by evidence? Assess the percent of "Yes" responses among closed cases with at least one AR-tracked allegation.	<50% of cases	50%–70% of cases	>70% of cases
	Qualitative	AR Pilot project learning log. CAPS narrative field for case closure summary. Use case closure summaries to a) identify qualitative examples to share with other pilot counties, b) provide coaching on including a conclusion in the narrative field, and c) identify the extent of impact.	Narrative examples indicate inconsistent and disparate use of case closure summaries, low to no inclusion of conclusion and extent of impact.	Narrative examples indicate mostly consistent and equitable use of case closure summaries, moderate inclusion of conclusion and extent of impact.	Narrative examples indicate consistent and equitable use of case closure summaries, extensive inclusion of conclusion and extent of impact.

Fidelity Indicator	Data Type	Data Source & Definition	Not Met	Approaching	Met
5. Matching Needs to Services Evidence that services in the case plan are being matched to client needs and their families.	Quantitative	ARD Case Planning & Provisioning of Services Measure 3: Were services necessary to mitigate risk factors with unmitigated significant impacts identified and added to the case plan in CAPS? [Abbreviated as 'Client Services Accurately Identified']. Assess services at the case level (not allegation level) for cases with only AR-tracked allegations; assess measure separately for cases with both AR- and TR-tracked allegations.	<50% of cases	50%–70% of cases	>70% of cases
	Qualitative	AR Pilot project learning log. Case closure summary in CAPS may also provide insight.	Narrative examples indicate inconsistent matching of services in the case plan to client and family needs.	Narrative examples indicate mostly consistent matching of services in the case plan to client and family needs.	Narrative examples indicate consistent matching of services in the case plan to client and family needs.
6. Use of Data Evidence that data are being used to improve Adult Protective Services practice, drive outcomes, and assure the equitable reach of AR.	Quantitative	Colorado Department of Human Services attendance records. County-level attendance at county rep meetings and learning sessions.	Attended <60% of learning sessions and county meetings.	Attended 60%–80% of learning sessions and county meetings.	Attended >80% of learning sessions and county meetings.
	Qualitative	AR Pilot project learning log. Documentation of actionable decisions arising from learning sessions and bi-annual memos.	Narrative examples indicate actionable decisions rarely address improvements to AR practice, and/or case outcomes, and/or equity in AR reach.	Narrative examples indicate actionable decisions sometimes address improvements to AR practice, and/or case outcomes, and/or equity in AR reach.	Narrative examples indicate actionable decisions consistently address improvements to AR practice, and/or case outcomes, and/or equity in AR reach.

Fidelity Indicator	Data Type	Data Source & Definition	Not Met	Approaching	Met
7. Continuing Education/Professional Development Evidence that county staff participate in mandatory and voluntary trainings and professional development opportunities on AR and apply training to practice and CAPS use.	Quantitative	Evidence that county staff participate in mandatory and voluntary trainings and professional development opportunities on AR and apply training to practice and CAPS use.	Participated in <60% of continuing education opportunities and voluntary AR trainings.	Participated in 60%–80% of continuing education opportunities and voluntary AR trainings.	Participated in >80% of continuing education opportunities and voluntary AR trainings.
	Qualitative	AR Pilot project learning log. Understand to what extend staff apply new knowledge gained from participating in AR trainings to their practice and to CAPS use.	Narrative examples indicate inconsistent, disparate, and low application of training to practice and CAPS use.	Narrative examples indicate mostly consistent, equitable, and moderate application of new knowledge to practice and CAPS use.	Narrative examples indicate consistent, equitable, and extensive application of new knowledge to practice and CAPS use.

Appendix B: Validation Checks

Common Support

To understand whether the analytical sample contains enough cases from which to make valid comparisons, we checked for common support. Common support requires that we have many cases in the pre-pilot period with high propensity scores making them good matches to cases in the pilot period. Common support also requires many cases in the pilot period with low propensity scores. In this context, we are looking for substantial overlap in the distribution of the propensity scores between cases in the pilot period and cases in the pre-pilot period. Common support is visually checked by plotting the distribution of pre-pilot and pilot propensity scores and verifying there is sizeable overlap in the distributions and that there is no bunching near 0 or 1. Distributions of propensity scores (available upon request) demonstrate that there is substantial overlap in the distributions and the absence of truncation at 0 or 1. Naturally, due to how we constructed the analytic sample, pre-pilot cases tend to have lower propensities and pilot cases, higher propensities. We see sizable density for both groups between propensity scores of 0.2 and 0.6.

Predictive Power of the Propensity Score

After weighting, statistical equivalence across a large set of covariates is a sign that pre-pilot cases and pilot cases are similar on average; thus, differences between them can be inferred as causal. If the covariates in the analytic sample are properly balanced, controlling for propensity score in a regression, estimating the presence of at least one Alternative Response (AR) allegation in a case on the matching variables, should remove any statistically significant relationships. Without including the propensity score as a control, the regression estimates large and statistically significant coefficients. Thirty-one covariates are statistically significant, accounting for one-third of all estimated covariates. With the inclusion of the propensity score in the model, none of the matching covariates are statistically significant and the magnitudes of the coefficients shrink. This indicates that the propensity score is correctly subsuming the information contained in the covariates.

Equivalence of the Matching Variables at Baseline

Comparing means of the matching variables across pre-pilot and pilot samples before and after weighting is used to ensure cases from the pre-pilot sample are observationally equivalent to AR cases. A series of difference-in-means tests find that prior to weighting, differences in incidence of caretaker neglect, exploitation, harmful act, physical abuse, self-neglect, presence of an initial mental illness, substance abuse, emergency/immediate harm, and most of the age interactions are statistically significant ($p < 0.01$). Differences for other matching variables (e.g., sexual abuse, age, presence of medical conditions, and other age interactions) are also statistically significant ($p < 0.05$). After weighting, the same tests found that none of the differences are statistically significant at any conventional level. Overall, the balance tests indicate the weighting scheme based on propensity scores was successful at generating balance across the two samples and reinforces the notion that the estimates produced are doubly robust. Adding covariates in order to model the outcome can improve statistical precision by reducing unexplained variance in the outcome; we include covariates in the estimation of the impact of AR as a robustness check.

Appendix C: Outcome Measures

1. Repeat Involvement

a. **Repeat involvement**: A repeat involvement occurs when a client has a screened-in case opened and closed and then has a subsequent screened-in case opened within 6 months of the prior case closing. Repeat involvement can occur regardless of the number of cases a client has.

b. **Repeat self-neglect**: A repeat involvement case in which both the first and second cases had an allegation of self-neglect. This measure differs from overall repeat involvement because, to be included in this measure, a client's first case must have had a self-neglect allegation and must have opened a second case within 6 months of initial case closure.

c. **Repeat mistreatment**: A repeat involvement case in which both the first and second cases had an allegation of mistreatment. This measure differs from overall repeat involvement because, to be included in this measure, a client's first case must have had a mistreatment allegation and must have opened a second case within 6 months of initial case closure.

2. Case Length

a. **Total case length**: The number of days between when a report was received and when the case was closed.

b. **Report received to initial response**: The number of days between when a report was received and when a caseworker either made initial contact or attempted to make initial contact, in person or over the phone.

c. **Report received to baseline assessment completion**: The number of days between when a report was received and when the baseline assessment was completed. To be included in this measure, a case must have a complete baseline assessment. Clients missing a completed assessment could have refused contact, were unable to be contacted, died, or experienced another circumstance that prevented the completion of the assessment.

d. **Initial response to baseline assessment completion**: The number of days between when a caseworker either made initial contact or attempted to make initial contact, and when the baseline assessment was completed. To be included in this measure, a case must have a complete baseline assessment.

e. **Report received to last finding or conclusion**: The number of days between when a report was received and when a caseworker determined the last finding or conclusion on a case.

f. **Initial response to last finding or conclusion**: The number of days between when a caseworker either made initial contact or attempted to make initial contact, and when a caseworker determined the last finding or conclusion on a case.

3. Client Engagement

a. **Client refused contact**: When a client refuses contact, the caseworker closes the case with a closure reason of "client refused contact." Refusal typically occurs early in the life of a case, but a client can refuse contact at any point. In the analytic sample, there were 595 cases (4.06%) with this closure reason.

b. **Client refused all services**: When a client refuses *all* services, the caseworker closes the case with a closure reason of "client refused all service". To be considered in this subsample, a client must have engaged long enough to have been offered services and determined to have a need for services. Services were not offered in situations where a client refused contact earlier, there were no identified needs, the assessment determined the client was not at risk, the client was not able to be located, or client death amongst other reasons. Consequently, outcome measures analyzing service uptake will be based on a non-random subsample of cases. In the analytic sample, there were 791 cases (5.39%) with this closure reason.

c. **Percent of services refused**: Of the clients who were offered services, the percentage of services refused is calculated by taking the number of services refused and dividing it by the number of services offered. This measure differs from the previous measure because it is not binary and intended to capture subtler changes in client engagement.

d. **All services ineffective**: Of the clients who were offered services, if all services were determined ineffective, the caseworkers closed the case with a closure reason of "services ineffective." In the analytic sample, there were 59 cases (0.40%) with this closure reason.

e. **Total number of client communications**: A communication defined as either a monthly contact or an interview. Department policy mandates approximately one contact per month not to exceed 35 days, but it is possible that caseworkers and clients communicate more frequently. When a caseworker adds an interview, it is because they interviewed the client. When a monthly contact is added, this could be the result of a communication with the caseworker or another professional like nursing staff or a case officer.[ix]

f. **Average number of client communications**: Average number of communications per month is calculated by taking the total number of client monthly contacts plus interviews and dividing it by the number of months a case was open.

4. Client Safety

a. **Safety score**: Client safety score is an aggregate measure based on the weighted risk and protective factors. The scores are bounded between 0 (least safe) and 100 (most safe). The construction of the safety score measure implies that clients with many risk factors, but also many mitigating factors, will score similarly to clients with few risk factors and few mitigating factors. To be included in the safety score analysis, a client must have completed

[ix] In the original file containing client communications, a trivial six cases had a communication that was not a monthly contact or an interview. All six were in the pre-pilot period.

a baseline assessment and reassessment (maintaining engagement, determined to be at risk, and progressed through the case process until closure).

b. **Risk score**: Unlike the safety score, risk score does not account for the presence of mitigating and protective factors. The scores are bounded between 0 (least risk) and 100 (most risk). The risk score is a component of the safety score. To be included in the risk score analysis, a client must have completed a baseline assessment and reassessment (maintaining engagement, determined to be at risk, and progressed through the case process until closure).

5. Allegation Severity Level and Extent of Impact

a. **Rating of severe or substantial harm**: A severity level is entered when a finding of "Substantiated" or "Mistreatment Occurred - Not Culpable" is entered on a Traditional Response-track allegation. With Alternative Response-track allegations, an extent of impact level is entered when a "Yes" conclusion is entered, indicating whether or not the allegation as alleged was true. If so, an impact level (using a 3-point scale) is then entered to confirm the level of harm, if any to the client. Following Adult Protective Services (APS) guidance and the logit model, we treat the highest severity level, "severe," and highest extent of impact level, "substantial," the same. A binary variable is created to differentiate the highest severity and extent of impact level from the two lesser levels.

6. Case Closure Reason

a. **Case closed because APS intervention was completed**: When the interventions drawn up in the case plan are completed, the caseworker closes the case with a closure reason of "APS intervention complete." To reach this closure designation, "APS intervenes, completes an investigation and assessment, and develops a case plan. APS works to implement services and/or works with others to implement services that mitigate risk, even if unable to mitigate all needs." In the analytic sample, there were 4,418 cases (30.12%) with this closure reason.

b. **Case closed because client was determined not at risk**: To reach this closure designation, "The case has been assigned but upon assessment, the client does not meet criteria as an at-risk adult." In the analytic sample, there were 1,803 cases (12.29%) with this closure reason.

c. **Case closed because no needs were identified**: To reach this closure designation, "The investigation and assessment were completed and there are no identified significant unmitigated needs." In the analytic sample, there were 4,753 cases (32.40%) with this closure reason.

Endnotes

[1] Brownson, R. C., Shelton, R. C., Geng, E. H., & Glasgow, R. E. (2022). Revisiting concepts of evidence in implementation science. *Implementation Science: IS, 17*(1), Article 26. https://doi.org/10.1186/s13012-022-01201-y

[2] Creswell, J.D., & Cresswell, J.D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Thousand Oaks, CA: SAGE Publications, Inc.

[3] Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.2307/2335942

[4] Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189. http://www.jstor.org/stable/1555493

[5] Robins, J.M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, *90*(429), 122–29. https://doi.org/10.1080/01621459.1995.10476494

[6] Heckman, J., & Navarro-Lozano, S. (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *The Review of Economics and Statistics, 86*(1), 30-57. https://doi.org/10.1162/003465304323023660