# DIGITS Trial

## Statistical Analysis Plan (SAP)

## Version 1.0

**Original Approval: February 10, 2023**

**Updated: January 23, 2025**

**ClinicalTrials.gov Identifier: NCT05160233**

# Table of Contents

## LIST OF ABBREVIATIONS

| | |
|---|---|
| CONSORT | Consolidated Standards of Reporting Trials |
| DIGITS | Digital Treatments for Opioids and Other Substance Use Disorders in Primary Care |
| DSM | Diagnostic and Statistical Manual of Mental Disorders |
| EHR | Electronic Health Record |
| IRB | Institutional Review Board |
| ITT | Intent-To-Treat |
| KPWA | Kaiser Permanente Washington |
| MAR | Missing At Random |
| MI | Multiple Imputation |
| MI | Multiple Imputation with Chained Equations |
| OUD | Opioid Use Disorder |
| PC | Primary Care |
| PI | Principal Investigator |
| RCT | Randomized Controlled Trial |
| SAP | Statistical Analysis Plan |
| SD | Standard Deviation |
| SE | Standard Error |
| SUD | Substance Use Disorder |

## OVERVIEW OF CHANGES FROM PRIOR VERSION OF SAP

In this document, we note any additions or modifications to SAP Version 1.0 following its original creation. Rather than “up-version” the SAP to a new version number, for each change, we include an “Addendum” describing the change and rationale, and provide the date indicating when the change was made.

# 1.0 SUMMARY OF STUDY DESIGN AND PROCESSES

## 1.1 Study Objectives

The Digital Treatments for Opioids and Other Substance Use Disorders in Primary Care (DIGITS) Trial addresses a critical knowledge gap: how to best implement digital treatments for substance use disorders in primary care. We will study implementation of the reSET® and reSET-O® digital therapeutic platforms into primary care to potentially improve care for people with substance use disorders, and as a model for how to sustainably implement digital treatments into real-world healthcare.

This Statistical Analysis Plan (SAP) addresses the following study aim: estimate the effect of clinician-facing (practice facilitation) and patient-facing (health coaching) implementation strategies in increasing the reach and fidelity of a digital therapeutic for substance use disorders in primary care clinics. We hypothesize significantly higher reach among clinics randomized to practice facilitation strategy (**hypothesis 1**) and significantly higher fidelity among clinics randomized to the health coach strategy (**hypothesis 2**), compared to standard implementation. A secondary study aim to compare the population-level cost-effectiveness of each implementation strategy will be described in a separate analysis plan. Additional study hypotheses and secondary objectives are described in Sections 5.0.

## 1.2 Study Design and Randomization

The DIGITS Trial is a 2-by-2 factorial trial randomizing eligible primary care clinics at Kaiser Permanente Washington (KPWA) to four implementation approaches for digital treatments in primary care: (a) standard implementation, an evidence-based implementation strategy developed by our delivery system partners to implement a digital depression treatment; (b) standard with practice facilitation, adding a clinician-facing implementation strategy; (c) standard with a health coaching, adding a patient-facing implementation strategy to support reSET/reSET-O delivery, and (d) standard with both, adding both the clinician-facing and patient-facing implementation strategies.

**Randomization:** A randomization list will be created by the study biostatistician prior to the start of the trial using a computer-generated list of random numbers. Eligible clinics will be randomized 1:1:1:1 to each of the four intervention groups using permuted blocks of size 4, and 8 (to ensure approximately equal accrual into each arm of the study over time). Group assignments will be concealed in a password protected file by the biostatistical team until clinics become eligible.

## 1.3 Intervention groups

Details on the implementation strategies are described in the study Protocol. Briefly, all clinics receive the standard implementation strategy. Additionally, half are randomly assigned to receive a health coach. Independently, half are assigned to receive practice facilitation. This results in the following 4 trial arms:

| **Implementation Strategy** | **Arm 1** Standard Only | **Arm 2** Standard + Practice Facilitation | **Arm 3** Standard + Health Coach | **Arm 4** Standard + Both |
|---|---|---|---|---|
| Standard | YES | YES | YES | YES |
| Practice Facilitation | NO | YES | NO | YES |
| Health Coach | NO | NO | YES | YES |

### 1.4 Sample and Sample Size

KPWA primary care clinics are eligible if they have at least 1 clinician trained in the use of reSET and reSET-O and had not previously participated in a pilot project of the digital therapeutics. To maximize the number of clinics in the study, clinics could become eligible if they met criteria at any time between December 9, 2021 and August 11, 2022. Clinics that are not initially eligible could become eligible if a clinician completed training to prescribe reSET and reSET-O.

The number of primary care clinics available varies over time as some clinics close and new clinics open in the health system. As part of the pilot activities (described in Section 2.4), 27 clinics were identified as available to be randomized (after excluding 2 clinics that were selected to pilot the implementation strategies). In some cases a pair of clinics may be randomized together; in this case the pair of clinics will be handled as a single unit (we use the term “clinic” to refer to the randomized unit) for the purpose of defining the clinic-level outcomes (Section 4.0).

The sample of patients eligible for the trial include patients with a primary care visit at which they are age 18 or older to one of the randomized clinics during the patient accrual period (up to 1 year following randomization; defined below) who screen positive for cannabis or drug use on the day of the visit or in the prior year.

# 2.0 GENERAL PROCEDURES AND DEFINITIONS

## 2.1 Intent-to-treat (ITT) Analysis

Unless otherwise specified, all analyses will follow an intent-to-treat principle whereby clinics will be analyzed according to the intervention arm to which they were randomized regardless of the subsequent sequence of events.

## 2.2 Time Periods Anchored on Randomization

### 2.2.1 Study Day 1

For each clinic, the randomization date is defined as study day 1 (study day 0 is defined as the day before randomization).

### 2.2.2 Patient Accrual Period

This is the period from two weeks before study day 1 ("lead-in period") through the end of the clinic's active implementation period (defined next). We allow patients who have a visit two weeks before randomization to be included because, upon the beginning of implementation, providers may be motivated to provide this new treatment to recent patients they have seen in the clinic.

### 2.2.3 Active Implementation Period

This is the period from study day 1 until 12 months after, or for clinics that randomized late (after February 12, 2022) from study day 1 until a minimum of 6 months later. The clinics that randomized late do not have the full 12-month active implementation period to enable sufficient time for follow-up of secondary outcomes that require 18 months after the index visit.

## 2.3 Time Periods Anchored on Patient Visits

### 2.3.1 Index Visit

This is the first eligible visit during the Patient Accrual Period (defined above). If a patient's first eligible visit occurs during the "lead-in period" (defined in Section 2.2.2 above), then the index visit date is defined to be study day 1. If patients have eligible visits to multiple randomized clinics, then they are "assigned" to the clinic where they have their first eligible visit.

### 2.3.2 Follow-up Period

This is the period from the day of the index visit up to 18 months later, depending on the outcome (see Section 4.0).

## 2.4 Pilot Activities

### 2.4.1 Data Only Pilot

Prior to the trial, a secondary dataset—with data from before the start of the trial—was used to inform modifications and refinements to the trial design. This retrospective cohort included patients with a visit at which they were 18 years to one of KPWAs primary clinics during the period from January 1, 2017 to February 29, 2020 (ending before any clinics were randomized for the DIGITS trial). Specific goals of these pilot analyses included the following:

- Assess feasibility of the abstinence outcome (referred to as "effectiveness" from the original grant). We estimated the proportion of eligible patient who would have non-missing follow-up data during different time periods of interest. Based on these analyses, we determined that follow-up rates were lower than expected and that a longer follow-up of 18 months was needed for this measure.

- Evaluate statistical power under different candidate analytic samples. To limit the potential for identification bias (a type of selection bias that occurs when the intervention affects who is identified as eligible for inclusion in trial analyses)[1], our original plan was to use as our analytic sample patients who screen positive for cannabis or substance use ("screen-positive patients"). However, we sought to examine alternate candidate samples that avoid identification bias, including screen-positive patients, patients with a substance use disorder (SUD) diagnosis pre-randomization, and patients with a positive symptom checklist on the day of the visit or in the past year, to select the analytic sample with the largest power. As a result of the power evaluation, the screen-positive population appeared to have the highest relative power from among the samples considered, and thus no change was made to the choice of analytic sample.
    - Given the availability of more recent data (relative to the original grant) under the data-only pilot, we also updated our power calculations for the trial using the most accurate information on projected sample size available prior to trial launch. Specifically, we used data from a 1-year period from March 1, 2019 to February 29, 2020 on the number of clinics and average sample size within a clinic.

# 3.0 STUDY POPULATION

Patients are eligible for inclusion in the trial if:

1) Have a primary care visit in one of the DIGITS trial clinics during the Patient Accrual Period (defined above),
2) Screen positive for cannabis or drug use on the day of the visit or in the prior year,
3) Adult aged 18 years or older at time of visit

Visits meeting the above criteria are referred to as “eligible visits”. Patients who have requested through the health system to opt out of research are excluded.

## 3.1 Assignment of Patients to Clinics

Patients will be assigned to clinics based on clinic where they had their first eligible visit meeting the above criteria. This visit is referred to as the “index visit” (see Section 2.3.1).

# 4.0 OUTCOME MEASURES

## 4.1 Definition of Primary Outcome Measures

### 4.1.1 Reach

*Reach* is the clinic-level proportion of patients who initiate reSET or reSET-O. For this outcome to occur, a clinician must prescribe the digital therapeutic and the patient must complete at least one treatment module. This measure is calculated by first defining whether each eligible patient assigned to the clinic is reached, and then calculating the proportion over all eligible patients. For each patient, we will identify whether they were reached from the day of the "index" visit (defined in Section 2.3.1) through the end of the active implementation period plus a 2-week ("grace period"). We will allow for 2 weeks beyond the active implementation period so that patients who have their first qualifying visit at the end of the active implementation period will have at least 2 weeks to be reached. The impact of the choice of grace period will be explored within the 2-clinic pilot and may be modified; the final choice of grace period will be specified prior to obtaining any outcome data on the trial clinics.

### 4.1.2 Fidelity

*Fidelity* is the clinic-level mean number of weeks in which patients use the digital therapeutic as recommended. This includes completing a recommended 4 or more modules per week while under the care of a clinician (visit in the past 30 days). This measure is calculated by first calculating the number of weeks' fidelity for each eligible patient assigned to the clinic (patients never reached have 0 weeks' fidelity) and then averaging over the number of eligible patients. The maximum number of weeks' fidelity for each patient is 12. We note that the prescription length is 90 days for reSET (~12.9 weeks) and 84 days for reSET-O, but we use 12 weeks for both outcomes for consistency. To identify whether a week was one with fidelity, we identify each patient's reSET activation date and examine for module use during the following 12 weeks. Possible followup for all patients extends up to 14 weeks after the end of the active implementation period (defined in Section 2.2.3). This 14 week period includes the 2-week "grace period" defined above for Reach, and allows for 12 weeks of app use for patients who are reached on the last day of the grace period. We note that the denominator is all eligible patients assigned to the clinic (rather than using a denominator of patients that are reached) to avoid denominator—or identification—bias.[1]

## 4.2 Definitions of Secondary Measures

Additional secondary measures will be analyzed to further describe the impact of the implementation models on outcome measures reflecting processes of care, implementation, and effectiveness. The following table describes each of the measures that will be analyzed as described in the statistical analysis section below.

**Table 4.2: Primary, Secondary, and other Outcomes [a]**

| **Primary Outcome Measures** |
|---|
| **1. Reach of the digital therapeutic to patients in the primary care clinic.** The proportion of patients who initiate the digital therapeutic, defined by instances in which patients open the app, enter the prescription code, and use a treatment module. |
| **2. Fidelity of patients' use of the digital therapeutic to clinical recommendations.** Mean number of weeks during patients' 12-week prescription in which patients use 4 app modules/week and have visited a clinician in the past 30 days. [b,c] |
| **Secondary Outcome Measures** |

| |
|---|
| **1. Engagement.** Mean number of months in which patients make ≥1 visit for substance use disorder [b,c] |
| **2. Economic costs.** Costs from the perspective of a health system and payer including implementation, direct intervention, operating, and other indirect health care costs. This measure will be used to calculate the population-level cost effectiveness of increasing reach, fidelity, and engagement.<br>The statistical analysis plan for economic costs is being developed separately. |
| **Other Outcome Measures** |
| **1. Reach-2.** The proportion of patients prescribed reSET or reSET-O |
| **2. Fidelity-2.** Mean number of weeks in which patients use at least 1 module/week [c] |
| **3. Sustainment.** The proportion of patients who are reached during the sustainment period<br>**Addendum** (June 17, 2024): This outcome is no longer being analyzed due to infeasibility, because the digital therapeutic vendor filed for bankruptcy in April 2023 |
| **4. Substance use.** The proportion of patients who are reached and reduce their substance use [d]<br>**Addendum** (January 13, 2025): Due to low numbers of patients who were reached, this outcome is presented as the proportion of patients reached who reduce their substance use; we omitted the planned inferential statistics that would have compared these proportions. |
| **5. Abstinence.** The proportion of patients who are reached and are abstinent from substances. [d]<br>**Addendum** (January 13, 2025): Due to low numbers of patients who were reached, this outcome is presented as the proportion of patients reached who are abstinent from all substances; we omitted the planned inferential statistics that would have compared these proportions. |
| **Other Explanatory and Sensitivity Analysis Measures** |
| *Adoption*. The proportion of health care providers prescribing the digital therapeutic, overall and by provider type |
| *Adoption-2.* The mean number of months in which providers access clinician dashboards |
| *Reach-3*. Proportion of patients who download and unlock the digital therapeutic |
| *Fidelity-3*:[e] Mean number of weeks in which the patients use 4 modules per week but without the requirement that they visit a clinician [c] |
| *Fidelity-4*:[e] Mean number of modules completed over the 12-week prescription |
| *Substance use-2.* The proportion of patients who are reached and reduce their substance use, as measured by self-report data collected by the digital therapeutic<br>**Addendum** (July 8, 2024): This outcome is no longer being analyzed due to low numbers of patients who were reached, and because among those who were reached there were low numbers of weeks with engagement (in which patients contributed self-report data). |
| *Abstinence-2.* The proportion of patients who are reached and are abstinent from substances, as measured by self-report data collected by the digital therapeutic<br>**Addendum** (July 8, 2024): This outcome is no longer being analyzed for the same reason as the preceding outcome measure. |

| |
|---|
| *Abstinence-3.* Abstinence verified by urine drug screens among patients prescribed the digital therapeutic for opioid use disorder, based on EHR data<br>**Addendum** (July 8, 2024): This outcome is no longer being analyzed due to the low numbers of patients prescribed reSET-O |

EHR=electronic health record

[a] Primary, secondary, and other pre-specified outcome measures are described as registered on ClinicalTrials.Gov. The other explanatory and sensitivity analysis measures were not registered.

[b] Visits must indicate that the clinician coded an International Classification of Disease (ICD-10) substance use disorder diagnosis.

[c] Fidelity and engagement are measured while a prescription is active, even if the prescription starts before but ends after their clinic's active implementation period is complete.

[d] Pre-specified substance use and abstinence are collected via self-report as part of routine annual screening for cannabis, alcohol, and other drug use. The assessment period includes up to 18 months after each patient's qualifying visit to help ensure a follow-up measure is collected. These will be analyzed among patients who become eligible before the last 3 months of active implementation (to help ensure the patients could complete the 12-week prescription before the outcomes were measured). Substance use will also be assessed separately by each type (i.e., drug-specific reductions in the frequency of cannabis use, illicit drug and prescription medications use, and alcohol use separately). We will consider follow-up screening scores from 3-18 months after the "index" visit, and if a patient has multiple follow-up scores, we will use the score closest to 12 months.

[e] One secondary outcome (Fidelity-2) was listed twice by mistake in the published DIGITS protocol paper (the same measure was also listed as "Fidelity-3").[2] We therefore deleted this redundant outcome and re-numbered the outcome names (Fidelity-3 in this table is listed as Fidelity-4 in the protocol paper, and Fidelity-4 is listed as Fidelity-5).

# 5.0 ANALYSES OF OUTCOME MEASURES

## 5.1 Analysis of Primary Outcome Measures

### 5.1.1 Evaluation of Primary Hypotheses

For each of the two primary clinic-level outcomes (reach and fidelity), we will fit a separate linear regression model to estimate the main effect of each factor level (practice facilitation, health coach) using contrast (effect) coding, which ensures accurate interpretation of the effect estimates given the factorial design (see Appendix C for details).[3, 4] We will test **hypotheses 1** and **2** by applying a (two-sided) Wald test to the appropriate contrast from the regression model. Because robust standard errors (SEs) can perform poorly with small sample sizes,[5] we plan to apply the standard least squares estimator for calculating SEs.

To control the familywise type 1 error rate of the co-primary hypotheses at 0.05, we will use the Holm procedure.[6] Under the Holm method, which is uniformly more powerful than the standard Bonferroni correction, unadjusted P-values from $M$ different tests are first sorted from smallest to largest (let $P_{(i)}$ denote the $i$th smallest P-value). Statistical significance for each test is then evaluated sequentially with a significance threshold of $\alpha/(M-i+1)$ where $\alpha \coloneqq 0.05$ denotes the nominal (familywise) type 1 error rate. Once a null hypothesis fails to be rejected hypothesis testing is stopped. Thus, for our 2 primary outcomes, we would compare the smallest P value with a significance threshold of 0.025; if non-significant stop. Otherwise, if significant then compare the second P value with a threshold of 0.05 to assess significance.

### 5.1.2 Evaluation of Secondary Hypotheses

We secondarily hypothesize that (a) practice facilitation is superior to health coaching in increasing reach and (b) health coaching is superior to practice facilitation in increasing fidelity. To test these hypothesis, we will compare the main effect of practice facilitation to the main effect of health coaching (see Appendix C).

We also will examine whether the two enhanced strategies together are superior to standard implementation by comparing clinics with both enhanced strategies to clinics with neither. Finally, we will examine interaction effects between the two enhanced strategies: we will estimate the interaction term (and construct 95% confidence intervals [CIs]), which corresponds to the difference in the effect of a given implementation strategy due to the addition of the second implementation strategy (Appendix C).

We plan to report unadjusted P values and construct 95% CIs for parameters corresponding to each hypothesis above; findings will be interpreted cautiously unless highly statistically significant given issues of multiple testing.

### 5.1.3 Sensitivity Analyses for Primary Hypotheses

We will conduct several sensitivity analyses to examine the robustness of results to our analytic approach. If, due to random chance, factor levels are imbalanced in any baseline characteristics (e.g., clinic size), we will perform sensitivity analyses where we include these characteristics in the regression models (one at a time, given the small number of clinics).

Additionally, following reporting guidelines,[7, 8] we will examine whether the proportion of screen-positive patients (the measure defining the study population) differs across arms. To assess for identification bias,[1] we plan to conduct sensitivity analyses in which we consider alternative population denominators: all patients with a SUD diagnosis, and all patients with visits regardless of whether they screened positive for substance use.

> **Addendum** (January 23, 2025): For the sensitivity analysis among patients with an SUD diagnosis, we restricted the primary sample to those who also had a documented SUD

diagnosis. We were unable to conduct the analysis among all patients with an SUD diagnosis—regardless of whether they screened positive for cannabis or drug use—as initially planned due to the study programmer leaving the study towards the end of the study and subsequent programmers' efforts were focused on other aspects of the study.

## 5.2 Analyses of Secondary and Other Outcome Measures

Analyses of engagement, the other pre-specified reach and fidelity measures, and other outcomes will follow the same analytic procedures as the primary outcomes (e.g., linear regression of the clinic-level measures). We hypothesize significantly higher engagement among sites randomized to practice facilitation and/or health coaching (Arms 2-4 from Section 1.3) versus those with neither (i.e., standard implementation only; Arm 1). For secondary outcome analyses we do not plan to adjust P values for multiple comparisons. However, given the potential for false positive findings when multiple testing is conducted any findings will be interpreted cautiously.

### 5.2.1 Analyses of Sustainment Outcomes

We will describe reach over time during the sustainment period graphically by study arm. If reach is greater than 5% during sustainment (overall or in any study arm), we will conduct secondary analyses allowing the intervention effect to vary over time. Specifically, we will subdivide the sustainment period into discrete time intervals (e.g., 4-month windows) and include interaction terms with intervention group; repeated measures over time within a clinic will be accounted for using a mixed-effects model with clinic-specific random intercepts.

## 5.3 Effect Modification and Subgroup Analyses

We plan to conduct analyses of the following subgroups: sex (male, female) and race and ethnicity (final subgroups to be determined after examining the distribution of sample sizes by groups). We will calculate subgroup-specific values of the primary outcomes (e.g., reach among males) for each clinic. These clinic- and subgroup-specific measures will be modeled using a parallel approach as for the primary analysis, additionally including an interaction term between the demographic subgroup indicator and each of the intervention groups (patient coach, practice facilitation). To account for correlation of subgroup-specific measures within a clinic, these models will also include a clinic-specific random intercept. We will examine whether the intervention effect differs across each demographic factor and estimate subgroup-specific intervention effects. Any such comparisons will likely be underpowered and must be interpreted with caution.

Given that clinic-level estimates of reach will be more highly variable for smaller subgroups, we plan to only estimate intervention effects for subgroups that comprise at least 20% of the full sample (and at least 10% of the sample in each clinic). If we do not have a sufficient sample size within a particular race and ethnicity subgroup, we will consider collapsing these subgroups to create a larger group. Analyses will be descriptive as we do not have hypotheses on how intervention effects might differ across these subgroups.

To describe potential disparities across smaller groups (comprising at least 5% of the sample) we will describe the mean clinic-level outcome (and construct 95% CIs) by arm.

## 5.4 Exploratory Patient-level Analyses

In addition to estimating intervention effects at the clinic level, we will also conduct analyses to estimate reach, fidelity, abstinence, and substance use outcomes across patients assigned to clinics within each of the 4 study arms. Additional exploratory analyses will also estimate the mean outcomes (e.g., proportion reached, mean weeks' fidelity) stratified by sex and SUD type to provide data about implementation effectiveness in specific subgroups.

**Addendum** (August 5, 2024): No analyses were conducted by SUD type due to the small number of patients reached.

# 6.0 POWER CONSIDERATIONS

In factorial trials, main effects are analyzed by examining mean responses at one factor level and at a contrasting factor level, collapsed across all levels of the other factors.[3, 4, 9] For example, to examine the main effect of practice facilitation we will compare outcomes for Arms 2 and 4 to Arms 1 and 3 described in Section 1.3. Thus, power is based on the sample size (here number of clinics) per factor level, not per randomization arm. Because outcomes are ascertained from the EHR, we assume we will have complete outcome data on all patients and attrition will not affect final sample size (see section 8.1 on missing data).

Minimal detectable differences for fixed 80% power were estimated based on 27 primary care clinics (total number of operating clinics at the time of the trial pilot after excluding the 2 pilot clinics) and a two-sided type 1 error rate of 0.025 for the co-primary outcomes of reach and fidelity (to control the familywise type 1 error at 0.05). Calculations are based on a sample of screen-positive patients, and we assume based on baseline data that 16.6% of these patients will be eligible for reSET, as indicated by having a positive DSM checklist (≥2 SUD symptoms); we refer to these patients as "eligible" in this Section of the SAP. We further assume a standard deviation (SD) across clinics of the proportion of patients reached of 0.016 based on preliminary data on the clinic-level proportions of patients reached by digital depression treatment implemented previously at KPWA available at the time of the trial pilot (described in Section 2.4.1).

## 6.1 Reach

Under the above assumptions, we estimated that we will have >0.80 power to detect an increase of 2 percentage points in the clinic-level percentage of screen-positive patients who are reached in clinics with external facilitation compared to clinics without. This corresponds to a 12 percentage point increase among those eligible (e.g., from 4% to 16%). Calculations were based on a two-sample t-test and an SD of 0.016 as described above.

## 6.2 Fidelity

We estimated that we will have >0.80 power to detect an increase in the clinic-level mean number of weeks fidelity of 0.088 among screen-positive patients, assuming a SD across clinics of 0.07. To provide context for this effect size, we present the corresponding mean change in weeks fidelity among patients who are reached, under different assumptions on how the health coach could increase reach:

| Scenario of increased reach in clinics with a health coach* | % of patients reached with a health coach | | Detectable change (with 80% power) in mean weeks' fidelity among reached patients |
|---|---|---|---|
| | Among screen-positive patients | Among eligible patients** | |
| No increase | 1.3 | 8.0 | 6.6 |
| 25% increase | 1.7 | 10.0 | 4.5 |
| 50% increase | 2.0 | 12.0 | 3.1 |

*All scenarios assume the following for clinics without a health coach: 1.3% reached among screen-positive patients (8% among eligible), and a mean of 4 weeks fidelity among reached patients

**Assuming 16.6% of screen-positive patients are eligible for reSET/reSET-O

For example, if reach among clinics with a health coach increases by 25% (from 1.3% to 1.7%) among screen-positive patients (from 8% to 10% among those eligible) this difference of 0.088 weeks fidelity among screen-positive patients corresponds to an increase in the mean number of weeks of 4.5 (from 4 to 8.5 weeks) among those reached.

Calculations were based on a two-sample t-test and assumed a SD of the mean number of weeks among screen-positive patients of 0.07. This SD was estimated based on the SD of the proportion reached by a digital depression treatment at KPWA of 0.016, an average sample size per clinic of 639 screen-positive patients, an assumed mean of 4.0 weeks fidelity among patients reached in clinics without a health coach, and a SD of 3.6 weeks fidelity among patients reached. These latter values were based on publicly-released secondary datasets[10, 11] of the multisite TES/reSET RCT that indicated that patients who initiated the treatment (n=226) had a mean=8.0 (SD=3.6, range=0-12) weeks of fidelity, defined by completing ≥ 4 modules/week. We assumed half that level of engagement in clinics without a coach. Calculations for how the SD of the clinic-level mean number of weeks fidelity was derived from these parameters are in Appendix B.

# 7.0 DESCRIPTIVE ANALYSES

## 7.1 Analyses of Demographic and Baseline Data

The demographic variables for this study include age, sex, race/ethnicity, zip code-based characteristics (e.g., census variables), and type of insurance. The baseline clinical characteristics include diagnoses of medical conditions, mental health disorders, and substance use disorders. Clinic-level variables include aggregated measures of patient-level variables (e.g., proportion female) as well as clinic size.

Descriptive statistics for baseline and demographic variables will be presented for the randomized clinics and for participants assigned to the clinics, overall and separately for each of the intervention arms. Descriptive statistics will include N, mean, standard deviation, median, 25th and 75th percentiles, minimum and maximum for continuous variables and proportions and percentages for categorical variables. Since randomization is expected to produce balance at baseline across arms of the trial, statistical comparisons of treatment groups with respect to baseline characteristics will not be conducted. The updated CONSORT statement for parallel group randomized trials no longer recommends formal testing of statistical significance of differences between baseline characteristics.[12]

## 7.1 Crossover

We will examine crossover of patients between trial arms. Specifically, we will report the proportion of eligible patients assigned to a clinic from one arm who also had an eligible visit to a clinic from each of the other arms. Additionally, we will describe crossover among patients who are prescribed reSET or reSET-O.

## 7.2 Graphical Analysis of Reach Over Time

We will plot the proportion of patients reached over time since the clinic was randomized (e.g., within 3-month periods) by clinic and by study arm. The unit of time will be selected to allow for a sufficient precision of reach within the time window.

# 8.0 OTHER CONSIDERATIONS

## 8.1 Missing Data

Given that the primary and secondary outcomes rely on EHR data from the reSET and reSET-O app, and insurance claims data, if there is no evidence of a particular event, such as evidence of reach using app data, we will assume that the event did not occur.

For the primary outcomes of reach and fidelity, we expect patients who are enrolled in the KPWA health plan at the time of the index visit (and who remain enrolled over the follow-up period) will have complete capture of these outcome measures. Given our clinic-level outcome, we will examine whether clinic-level measures describing data capture (e.g., mean length of enrollment after the index visit, proportion of patients enrolled as of the index visit) vary across the four study arms. If any of these measures vary across arms (using a threshold of $P < 0.05$), then we will adjust for the measure with the strongest association in the outcome model (note we are limited in the number of variables that can be adjusted for given the small number of clinics). Other measures of data capture that differ across arms will be adjusted for separately in sensitivity analyses.

The secondary engagement outcome and other secondary and sensitivity measures of reach and fidelity will apply the same approach as for reach and fidelity.

For the other outcome measures of substance use and abstinence, we rely on screening data collected as part of routine clinical care. Patients will be missing follow-up screening outcome data if they receive care outside the health system or if they do not have a score documented in the EHR during the follow-up period (up to 18 months after the index visit). As above, we will examine whether the clinic-level proportions of patients missing a follow-up score differs across intervention groups. In addition, we plan to examine whether any clinic-level covariates (e.g., proportion of patients within different age groups) is associated with the clinic-level proportion of patients missing follow-up screening data. If we find that missingness differs across intervention groups (based on $P < 0.05$), then we will adjust for the baseline covariate most strongly associated with missingness in the outcome model. Other baseline covariates predictive of missingness will be adjusted for separately in sensitivity analyses. If conclusions differ across sensitivity analyses that adjust for different baseline covariates, or if we observe a large difference in missingness across arms (e.g., mean proportion missing differs by >5%), then we will consider applying multiple imputation (MI) methods.[13] Specifically, we plan use multiple imputation with chained equations (MICE)[14] to construct 25 imputed data sets—at a patient level—and then aggregate the patient-level imputed datasets to generate the clinic-level outcomes; Rubin's rules will be applied to accurately construct standard errors.[13] We will include in the imputation model any covariates that are associated with the patient-level outcome or with missing outcome information. This approach assumes that outcome data are "missing at random" (MAR).[13]

## 9.0 SAFETY AND INTERIM ANALYSES

Due to the nature of this study—testing implementation interventions, with all care provided by the health systems and using only secondary data—there are no formal interim analyses of safety performed. Further, all clinical care—and therefore responsibility for the quality of care—in this cluster-randomized pragmatic quality improvement trial is provided by the health systems. Therefore, no formal interim analyses linked to stopping rules will be conducted. Since all care is provided by the health system, not the study, it would not be appropriate to intervene at the patient or provider level for any safety issue.

## REFERENCES

1. Bobb JF, Qiu H, Matthews AG, McCormack J, Bradley KA. Addressing identification bias in the design and analysis of cluster-randomized pragmatic trials: a case study. *Trials*. Mar 23 2020;21(1):289. doi:10.1186/s13063-020-4148-z
2. Glass JE, Dorsey CN, Beatty T, et al. Study protocol for a factorial-randomized controlled trial evaluating the implementation, costs, effectiveness, and sustainment of digital therapeutics for substance use disorder in primary care (DIGITS Trial). *Implement Sci*. Feb 1 2023;18(1):3. doi:10.1186/s13012-022-01258-9
3. Kugler KC, Trail JB, Dziak JJ, Collins LM. Effect coding versus dummy coding in analysis of data from factorial experiments. *University Park, PA: The Methodology Center, Pennsylvania State University*. 2012;
4. Collins LM, Dziak JJ, Kugler KC, Trail JB. Factorial experiments: efficient tools for evaluation of intervention components. *Am J Prev Med*. 2014/10// 2014;47(4):498-504. doi:10.1016/j.amepre.2014.06.021
5. Imbens GW, Kolesar M. Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics*. 2016;98(4):701-712.
6. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*. 1979:65-70.
7. Eldridge S, Kerry S, Torgerson DJ. Bias in identifying and recruiting participants in cluster randomised trials: what can be done? *Bmj*. 2009;339
8. Eldridge S, Campbell M, Campbell M, et al. Revised Cochrane risk of bias tool for randomized trials (RoB 2.0): additional considerations for cluster-randomized trials. 2016;
9. Wu CJ, Hamada MS. *Experiments: planning, analysis, and optimization*. vol 552. John Wiley & Sons; 2011.
10. Campbell AN, Nunes EV, Matthews AG, et al. Internet-delivered treatment for substance abuse: a multisite randomized controlled trial. *Am J Psychiatry*. Jun 2014;171(6):683-90. doi:10.1176/appi.ajp.2014.13081055
11. Nunes EV. Data from: Web-delivery of Evidence-Based, Psychosocial Treatment for Substance Use Disorders (CTN-0044 Data). 2014.
12. Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMC medicine*. Mar 24 2010;8:18. doi:10.1186/1741-7015-8-18
13. Little RJ, Rubin DB. *Statistical analysis with missing data*. vol 793. John Wiley & Sons; 2019.
14. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*. 2011;30(4):377-399.

# APPENDICES

## Appendix A. Modifications to the Protocol

From the time of funding of the original DIGITS grant proposal, pilot work (described in Section 2.4) led to modifications to the original study design and SAP. Adaptations related to the intervention are described in the Study Protocol; here we focus on modifications pertaining to the SAP.

Changes prior to trial launch

- Changes to eligibility criteria: because we learned that not all clinics would have a provider who could prescribe reSET and reSET-O, eligibility criteria were modified to require that a clinic have such a provider prior to become eligible for the study.
- Changes in randomization approach: rather than randomize all primary care clinics at trial launch, given the modified eligibility criteria for clinics to be enrolled in the study, we randomize clinics in stages as they become eligible over time. To balance the 4 intervention arms over time we employ a permuted block randomizations. Given the small number of clinics, 4 intervention arms, and need to use permuted blocks, we do not incorporate any additional stratification variables (e.g., we do not stratify on clinic size as had been originally planned).
- Power calculations were updated using more recent data from KPWA clinics available prior to trial launch as part of the data-only pilot (see Section 2.4.1) on: the number of clinics, average sample size per clinic, and preliminary data on the clinic-level proportions of patients reached by digital depression treatment implemented previously at KPWA.

Modifications after trial launch

- One secondary outcome was listed twice by mistake in the published DIGITS protocol paper.[2] We therefore deleted this redundant outcome and re-numbered the outcome names listed in Table 4.2.

## Appendix B. Derivation of the Standard Deviation of the Fidelity Outcome

Let $W_{ij}$ be the number of weeks' fidelity for screen-positive patient $j$ at clinic $i$. Assume that the mean and variance of $W_{ij}$ among patients who are reached are given by $\mu_1$ and $\sigma_1^2$, respectively. Let $p_i$ be the clinic-specific proportion of patients reached at clinic $i$ and assume that the mean and variance of $p_i$ is given by $\mu_r$ and $\sigma_r^2$, respectively. Then it can be shown that the variance of $W_{ij}$ is given by $Var(W_{ij}) = (\sigma_1^2 + \mu_1^2)\mu_r - \mu_1^2\mu_r^2$.

Our clinic-level outcome if interest $W_i = \frac{1}{n}\sum_{j=1}^{n} W_{ij}$is the mean across patients in the clinic of number of weeks' fidelity (for simplicity, we assume a constant clinic size of $n$). It can be shown (using the variance formula for the sum of correlated random variables) that

$$Var(W_i) = \frac{(\sigma_1^2 + \mu_1^2)\mu_r - \mu_1^2\mu_r^2 + (n-1)\mu_1^2\sigma_r^2}{n}$$

Plugging in the assumed parameter values as described in Section 6.2 ($\mu_r = 0.0166$, $\sigma_r^2 = 0.0163^2$, $\mu_1 = 4$, and $\sigma_1^2 = 3.6^2$, and $n = 639$), we have $Var(W_i) = 0.0050$ and $SD(W_i) = 0.07$. (We note that since $n$ is relatively large, the above is approximately equal to $Var(W_i) \approx \mu_1^2\sigma_r^2$ and $SD(W_i) \approx 0.065$.)

## Appendix C. Details on Contrast Coding

We consider models of the form $E(Y) = \beta_0 + \beta_f F + \beta_c C + \beta_{fc} FC$ where $Y$ is the clinic-level outcome (proportion reached), $C$ is a factor variable for health coaching and $F$ is a factor variable for practice facilitation. Under standard 'dummy' coding, the factor variables are coded

as 0 and 1 (where 0 indicates absence of the condition and 1 indicates presence of the condition); consequently, $\beta_c$ is the effect of practice coaching in the absence of practice facilitation and $\beta_f$ is the impact of practice facilitation in the absence of health coaching.

An alternative coding method is to code $C$ and $F$ as -½ or ½ to indicate absence or presence of the condition, respectively. Under this approach we can express the mean outcome for each study arm as a function of the model coefficients as in the following table:

| Study Arm | $F$ | $C$ | $E(Y)$ |
|---|---|---|---|
| (1) Standard implementation | -½ | -½ | $\mu_1 = \beta_0 - \frac{1}{2}\beta_f - \frac{1}{2}\beta_c + \frac{1}{4}\beta_{fc}$ |
| (2) practice facilitation only | ½ | -½ | $\mu_2 = \beta_0 + \frac{1}{2}\beta_f - \frac{1}{2}\beta_c - \frac{1}{4}\beta_{fc}$ |
| (3) health coaching only | -½ | ½ | $\mu_3 = \beta_0 - \frac{1}{2}\beta_f + \frac{1}{2}\beta_c - \frac{1}{4}\beta_{fc}$ |
| (4) both practice facilitation and health coaching | ½ | ½ | $\mu_4 = \beta_0 + \frac{1}{2}\beta_f + \frac{1}{2}\beta_c + \frac{1}{4}\beta_{fc}$ |

We can then calculate the intervention effects of changing each condition on the mean outcome as the other condition stays the same. Specifically, we have:

- The change in the mean outcome when adding practice facilitation is
    - $\mu_2 - \mu_1 = \beta_f - \frac{1}{2}\beta_{fc}$ in the absence of health coaching ($C = -½$) and
    - $\mu_4 - \mu_3 = \beta_f + \frac{1}{2}\beta_{fc}$ in the presence of health coaching ($C = ½$).
    - Thus the effect of practice facilitation averaged over conditions with and without health coaching is given by $\beta_f = \frac{1}{2}[(\mu_4 - \mu_3) + (\mu_2 - \mu_1)]$. This is referred to as the “main effect” of practice facilitation.
    - And the difference in the effect of practice facilitation with versus without health coaching (interaction) is given by $\beta_{fc} = [(\mu_4 - \mu_3) - (\mu_2 - \mu_1)]$
- The change in the mean outcome when adding health coaching is
    - $\mu_3 - \mu_1 = \beta_c - \frac{1}{2}\beta_{fc}$ in the absence of practice facilitation ($F = -½$) and
    - $\mu_4 - \mu_2 = \beta_c + \frac{1}{2}\beta_{fc}$ in the presence of practice facilitation ($F = ½$).
    - Thus the effect of health coaching averaged over conditions with and without practice facilitation is given by $\beta_c = \frac{1}{2}[(\mu_4 - \mu_2) + (\mu_3 - \mu_1)]$. This is referred to as the “main effect” of health coaching.
    - And the difference in the effect of health coaching with versus without practice facilitation (interaction) is given by $\beta_{fc} = [(\mu_4 - \mu_2) - (\mu_3 - \mu_1)]$
- The change in the mean outcome when adding both practice facilitation and health coaching is

- $\mu_4 - \mu_1 = \beta_f + \beta_c$