STATISTICAL ANALYSIS PLAN

COMPARATIVE EFFECTIVENESS OF HEALTH SYSTEM-BASED VERSUS COMMUNITY-BASED DEMENTIA CARE/ A PRAGMATIC TRIAL OF THE EFFECTIVENESS AND COST-EFFECTIVENESS OF DEMENTIA CARE (D-CARE)

VERSION 4.0 FINAL
APRIL 9, 2024

FUNDING AGENCIES: PCORI/NIA
PCORI GRANT NUMBER: PCS-2017C1-6534
NIA GRANT NUMBER: 1 R01 AG061078-01

Prepared by:
Peter Peduzzi, PhD (blinded study biostatistician)
James Dziura, PhD (unblinded study biostatistician)
Erich Greene, PhD (unblinded study biostatistician)
Can Meng (unblinded study biostatistician)
Yunshan Xu (unblinded study biostatistician)
Hugo He (unblinded study biostatistician)
Jingchen Liang (unblinded study biostatistician)

ClinicalTrials.gov Identifier: NCT03786471

Table of Contents

1. STUDY SYNOPSIS

Title	COMPARATIVE EFFECTIVENESS OF HEALTH SYSTEM-BASED VERSUS COMMUNITY-BASED DEMENTIA CARE/ A PRAGMATIC TRIAL OF THE EFFECTIVENESS AND COST-EFFECTIVENESS OF DEMENTIA CARE
Study Design	The design is a pragmatic randomized 3-arm superiority trial. The unit of randomization is the patient/caregiver dyad.
Study Duration	6 years
Trial Sites	4 trial sites: Baylor, Scott, & White Health; Geisinger Health; University of Texas Medical Branch Health Care System; Wake Forest University
Objective	Conduct a pragmatic randomized trial to determine the comparative effectiveness and cost-effectiveness of two evidence-based models of comprehensive dementia care as well as the effectiveness and cost-effectiveness of both arms versus enhanced usual care.
Number of Subjects (Target)	2150: 1000 in each intervention arm and 150 in the usual care group
Main Inclusion Criteria	Community-living, including Assisted Living Facilities (cannot reside in a nursing home at the time of recruitment), persons with diagnosis of dementia established by a physician or other primary care provider (PCP), have family or friend caregiver(s) who speak English or Spanish, and have a PCP who is willing to partner with the program
Interventions	Health systems-based dementia care (based on the UCLA Alzheimer's and Dementia Care program) provided by a nurse practitioner or physician's assistant Dementia Care Specialist who works within the heath system. Community-based dementia care (using the BRI Benjamin Rose Institute Care Consultation model) provided by a social worker or nurse Care Consultant who works at a community-based organization (CBO). Both arms include structured assessments and creation of care or action plans, care coordination, and caregiver education and support, but they differ in key areas including the frequency and mode of communication with persons with dementia and caregivers, the degree of integration into the health system, including order-writing capability and use of the electronic health record. Usual care with consistent referral to Alzheimer's Association Helpline (1-800-272-3900) to speak to master's level consultants for decision-making support, crisis assistance, caregiver education, and referral to local programs and services.
Duration of Intervention and Follow-up	18 months
Primary Outcome	Behavioral symptoms and resulting caregiver distress as measured by the NPI-Q Severity and Modified Caregiver Strain Index scales.

Primary Analysis	All analyses will be according to intent to treat. The analysis of the two primary outcomes – NPI-Q Severity and Modified Caregiver Strain Index scores – will be done using a longitudinal repeated measures analysis based on maximum likelihood methods adjusted for the stratified randomization by site. Treatment differences will be summarized by least square means and multiplicity corrected confidence intervals.
Secondary Outcomes	NPI-Q Distress, caregiver unmet needs and confidence, and caregiver depressive symptoms.
Tertiary Outcomes	Patient long-term nursing home placement, functional status, cognition, goal attainment scaling (GAS), "days spent at home", Dementia Burden Scale-Caregiver, a composite measure of clinical benefit, Quality of Life of persons with dementia, Positive Aspects of Caregiving
Cost-effectiveness Analysis	Ratio of incremental net Medicare costs to incremental effects of the two primary outcomes, costs to Medicaid and consumers, changes in utilization by type of use, caregiver costs for spouses who have fee-for-service Medicare
Interim Analysis	Because this is a minimal risk trial with a relatively short follow-up period (18 months), no monitoring for efficacy or futility is proposed; only monitoring for safety will be done.

2. STUDY OBJECTIVES AND SPECIFIC AIMS

D-CARE is a pragmatic randomized clinical trial of 2150 participants at four diverse clinical trial sites (CTS) to compare the effectiveness and cost-effectiveness of 18 months of health systems-based dementia care (HSDC, based on the UCLA Alzheimer's and Dementia Care program) provided by a nurse practitioner Dementia Care Specialist (DCS) who works within the heath system versus community-based dementia care (CBDC, using the BRI Benjamin Rose Institute Care Consultation model) provided by a social worker or nurse Care Consultant who works at a CBO versus usual care (UC).

The study's Specific Aims are:

1. To engage 4 sets of health systems and community-based organizations, patients, families, and other stakeholders in a pragmatic clinical trial
2. To adapt and implement 2 dementia care interventions at each CTS, including training of DCS' and Care Consultants and pilot testing of the interventions, assessments, and outcomes
3. To recruit 2150 participants and randomize them to receiving HSDC (N=1000), CBDC (N=1000), or usual care (N=150)
4. To administer the interventions for 18 months for all participants
5. To collect self/caregiver-reported outcomes and use claims data to evaluate other secondary and tertiary outcomes
6. To compare the effectiveness of the 2 dementia care interventions versus each other and versus usual care
7. To conduct a cost-effectiveness analysis of the two intervention arms versus each other and versus usual care

8. To conduct analyses of tertiary outcomes of both interventions versus usual care including mortality, time spent at home, long-term nursing home placement, patient/caregiver satisfaction, goal attainment scaling (in the 2 active intervention groups) and comparing all three groups on several types of utilization and out-of-pocket expenses. We will assess physician satisfaction with dementia care provided by the 3 treatment groups.

3. RANDOMIZATION

At the time of the in-person or telephone (after the start of the 2020 COVID-19 pandemic) baseline assessment, caregiver and, if the person with dementia is capable, patient informed consent will be obtained. Upon successful completion of baseline measures, eligible persons will then be randomized to either HSDC or CBDC or UC in a 7:7:1 ratio. The randomization will be stratified by site using a permuted block design with a fixed block size of 15; using variable block sizes is not practical given the size of the block and the few numbers of participants to be enrolled in the usual care arm at each site. The randomization scheme will be computer-generated, and the allocation sequence will be concealed. Intervention assignment will be delivered by a web-based clinical trial management system. Because all the care processes employed in all arms reflect standard of care, the consent will focus only on the collection of data. This approach is the same as that employed in the PCORI- and NIA-funded Strategies to Reduce Injuries and Develop confidence in Elders (STRIDE) pragmatic multisite trial aimed at reducing serious falls-related injuries.

4. BLINDING

D-CARE is an unblinded trial; however, all baseline and follow-up data collection will be completed by blinded site staff.

At the Data Coordinating Center, Dr. Peduzzi will be the blinded statistician who will interface with the study leadership and attend the open session of the DSMB meetings; Dr. James Dziura, Dr. Erich Greene, Mr. Can Meng, Ms. Yunshan Xu, Mr. Hugo He and Mr. Jingchen Liang will be the unblinded statisticians who will prepare the randomization scheme and the closed DSMB report as well as attend the closed DSMB sessions.

5. OUTCOME MEASURES

5.1 Schedule of Administration

The primary, secondary, and tertiary outcomes are summarized in Table 1 along with their schedule of administration. Most of these outcomes are patient-reported and all are clinically meaningful to patients and caregivers as well as payers.

Table 1. Outcome Domains, Instruments and Timing

Measure	Administration	Month				
		Baseline (in-person)*	3 (phone)	6 (phone)	12 (phone)	18 (phone)
Sociodemographic characteristics		X				

Primary outcomes						
NPI-Q Severity (patient behaviors)	Questionnaire	X	X	X	X	X
Caregiver strain (MCSI)	Questionnaire	X	X	X	X	X
Secondary Outcomes						
NPI-Q Distress (caregiver)	Questionnaire	X	X	X	X	X
Caregiver depression (PHQ-8)	Questionnaire	X	X	X	X	X
Caregiver self-efficacy	Questionnaire	X		X		X
Tertiary Outcomes						
Cognition (MOCA)	Interview	X				X
Functional status (FAQ)	Interview	X				X
Functional status (Katz ADL)		X				X
Goal attainment scaling ***	Interview	X		X		X
Mortality	CMS					X
	Interview		X	X	X	X
Time spent at home	CMS					X
Inpatient hospital use**	CMS					X
Acute inpatient rehabilitation use**	CMS					X
Post-acute SNF use**	CMS					X
Hospice use**	CMS					X
Long-term nursing home use**	CMS					X
Caregiver Rating of Dementia Care Quality	Questionnaire				X	
Caregiver satisfaction with dementia care	Questionnaire		X		X	X
Dementia Burden Scale-Caregiver	Composite	X	X	X	X	X
Clinical Benefit	Composite	X	X	X	X	X
Quality of Life in Alzheimer's Disease	Questionnaire	X				X
Positive Aspects of Caregiving	Questionnaire	X		X		
Cost-effectiveness analysis	CMS					X

CMS=Center for Medicare and Medicaid Services and site-obtained claims data;

*Switched to telephone-collected after beginning of COVID-19 pandemic, with option for in-person visits

** Also used to calculate time spent at home

** Goal Attainment Scaling will be obtained only on a subset of caregivers (i.e., it will not be performed in caregivers in the usual care group nor on all caregivers in the two active intervention groups; this will be completed by an unblinded assessor (i.e., the Dementia Care Specialist or Care Consultant) with goal setting occurring at baseline, and goal attainment scaling occurring 6 months [6-12 months window] and 18 months [15-18 months window] after enrollment).

5.2 Primary Outcomes

There are two primary outcomes in D-CARE measuring dementia-related behavioral symptoms and caregiver distress at 3, 6, 12 and 18 months after baseline.

1. Severity of Dementia-related Behavioral Symptoms

The severity of symptoms of psychopathology in persons with dementia will be measured by the Neuro-Psychiatric Inventory Questionnaire - Severity (NPI-Q Severity). The NPI-Q Severity is a validated survey that assesses the caregiver's perception of the severity of 12 dementia-related psychiatric and behavioral symptoms. NPI-Q Severity scores range from 0-36 with higher scores indicating more severe symptoms.

2. Caregiver Distress

The level of caregiver distress/strain as measured by the Modified Caregiver Strain Index (MCSI). The MCSI is a 13-item validated tool used to assess severity of caregiver strain. The index targets financial, physical, psychological, and social aspects of strain and is scored from 0 to 26 with higher scores indicating greater levels of strain.

5.3 Secondary Outcomes

There are three secondary outcomes in D-CARE:

1. NPI-Q Distress (measured at 3, 6, 12 and 18 months after baseline)

Distress of caregivers due to the symptoms of psychopathology in persons with dementia will be measured by the Neuro-Psychiatric Inventory Questionnaire - Distress (NPI-Q Distress). The NPI-Q Distress scale is a validated survey that assesses the level of distress experienced by the caregiver in response to dementia-related psychiatric and behavioral symptoms. NPI-Q Distress scores range from 0-60 with higher scores indicating more severe distress.

2. Caregiver Depression (measured at 3, 6, 12 and 18 months after baseline)

The severity of depression in caregivers as measured by the Patient Health Questionnaire (PHQ-8). PHQ-8 is a shortened version of the PHQ-9 (item 9 is omitted) that is often used when the instrument is administered by telephone and intervention cannot be adequately provided. The PHQ-9 is 9-item validated tool used to assess depressive symptoms in the caregiver using the DSM-IV criteria for major depression and is scored from 0-27 with scores >10 indicating moderate symptoms and scores >20 indicating severe depressive symptoms. Because of the infrequency of positive responses to item-9, its deletion of has minimal effect on scoring and identical scoring thresholds for depression severity can be used.

3. Caregiver Self-Efficacy (measured at 6 and 18 months after baseline)

Caregivers' ability to manage dementia-related problems and ability to access help is measured with a 4-item self-efficacy scale [range, 1 (strongly disagree) to 5 (strongly agree)] measuring the caregiver's self-efficacy for caring for the patient with dementia and for accessing help, including community resources. The scores for each of the 4 items are summed to produce an overall caregiver self-efficacy score ranging from 4-20 with higher scores indicating better caregiver self-efficacy; scores ≥ 15 indicate high self-efficacy.

5.5 Tertiary Outcomes (Exclusive of Cost-Effectiveness Measures)

The description of tertiary outcomes excludes two cost effectiveness measures that will be analyzed by the D-CARE health economist: Cost-Effectiveness Relative to Severity of Dementia-related Behavioral Symptoms and Cost-Effectiveness Relative to Caregiver Distress.

1. Cognition of Persons with Dementia (measured at 18 months after baseline)

Cognition as measured by the Montreal Cognitive Assessment (MoCA). MoCA is a validated widely used 30-point test (higher is better) of cognition that captures mild cognitive impairment as well as dementia. This will be collected at baseline and at the end of the study to document disease progression. To reduce respondent burden and missing data, we will use a shortened 3-item form for reporting study participant baseline characteristics and measuring decline in cognition, a tertiary outcome. Participants who score 8 or higher on the shortened version will receive the full 22-item telephone MOCA to determine whether they have capacity to provide informed consent.

2. Functional Status (measured at 18 months after baseline)

Functional status will be measured using the Functional Activities Questionnaire (FAQ) ranges from 0 to 30 with higher scores indicating more functional dependence. We will also measure activities of daily living (Katz ADL), which has been validated and is well-established in research and clinical use. Katz ADL ranges from 0-6, with higher scores indicating more functional independence.

3. Goal Attainment Scaling (measured at 6-12 months and 15-18 months after baseline)

Goal attainment scaling (GAS) asks patients and caregivers to select their most important goal and assesses their progress towards meeting it as a result of one of the study's intervention. *Goal attainment scaling*, defined as whether a person's individual goals are achieved as a result of the study intervention, will be measured using a 5-point goal attainment scale. GAS describes the person's expected level of goal achievement over a specified timeframe, ranging from much worse than expected (scored as -2) to much better than expected (scored as +2). Scales are dynamically set according to a person's needs, while measurement of attainment is standardized. Measurement of goal attainment will be performed with caregivers by clinicians involved in the intervention (Dementia Care Specialists and Care Consultants), who are unblinded, and can be performed over the telephone.

4. Mortality (measured over 18 months)

Mortality as measured by interviews at 3, 6, 12 and 18 months. Data will be verified using the Center for Medicare and Medicaid Services at 18 months. For participants who did not complete the 18-month interview, vital status will be investigated using the electronic health records (EHR).

5. Time Spent at Home

Time spent at home is calculated as 540 minus (1) the number of days in a health care facility, including hospital, nursing home, acute rehabilitation, and inpatient hospice, and (2) the number of days not alive.

6. Caregiver Rating of Dementia Care Quality (measured at 12 months)

The outcome is a count of the number of yesses. Caregiver Rating of Dementia Care Quality is a composite instrument of 10 items (with yes or no responses) from the Assessing Care of Vulnerable Elders (ACOVE), Physician Consortium for Performance Improvement (PCPI) and the American Academy of Neurology (AAN) quality measures. The outcome is a count of the number of yesses (range 0-10, higher counts indicate greater caregiver rating of dementia care quality).

7. Caregiver Satisfaction with Dementia Care (measured at 3, 12 and 18 months)

Caregiver's satisfaction with the dementia care program is measured using a 11-item questionnaire, modified from the University of California Los Angeles' Alzheimer's and Dementia Care program, and ranges from 11 to 55 (higher scores indicate greater caregiver satisfaction with the dementia care program).

8. Dementia Burden Scale - Caregiver (measured at 3, 6, 12 and 18 months after baseline)

Dementia Burden Scale-Caregiver (DBS-CG) is a composite of the NPI-Q Distress, MCSI, and PHQ-8 scales with items transformed linearly to be on a 0-100 possible range and then averaged, with higher scores indicating higher caregiver burden. The minimal clinically important difference (MCID) for the DBS-CG is 5 points.

9. Clinical Benefit (measured at 3, 6, 12 and 18 months after baseline)

Clinical benefit is a binary measure of patient symptoms using the NPI-Q severity scale (the only patient outcome anticipated to benefit from the program) and caregiver symptoms using the DBS-CG scale. Benefit on the NPI-Q severity scale is defined as having a score of ≤ 6 (the lowest tertile of symptoms) or improving by at least 3 points (the MCID). DBS-CG benefit is defined as having a score of ≤ 18.8 (the lowest tertile of symptoms) or improving by at least 5 points (the MCID). Clinical benefit is defined as having benefit on either or both component scales. Defining benefit in this manner captures both preventive (those who have few symptoms at baseline and do not deteriorate) and therapeutic (those who improve) benefit from the program.

10. Quality of Life in Alzheimer's Disease (measured at 18 months after baseline)

Quality of life will be measured at 18 months by the Quality of Life in Alzheimer's Disease (QOL-AD). The QOL-AD is a 13-item instrument scored 4-52 (higher scores indicate better quality of life) that can be administered to persons with dementia and caregivers. It has demonstrated sensitivity to psychosocial intervention correlates with health-utility measures, is widely translated and used internationally and can be used by people with MMSE scores as low as three.

11. Positive Aspects of Caregiving (PAC) (measured at 6 months after baseline)

PAC is measured with 11 items focusing on favorable aspects of the caregiving experience, with range 0-44 (most positive) and Cronbach's alpha 0.92. The items ask about their mental or affective state related to their caregiving experience. For example, providing help to care recipients has 'made me feel more useful', 'made me feel strong and confident', 'given more meaning to my life', 'enabled me to develop a more positive attitude toward life', and 'enabled me to learn new skills'.

12. CMS Patient Outcomes (measured over 18 months)

- 12a. Inpatient Days Spent at an Acute Care Hospital is defined as the number of days an individual is admitted to an acute care hospital.
- 12b. Inpatient Days Spent at an Inpatient Rehabilitation Facility is defined as the number of days an individual is admitted to an inpatient rehabilitation facility.
- 12c. Inpatient Days Spent at a Skilled Nursing Facility is defined as the number of days an individual is admitted to a Skilled Nursing Facility.
- 12d. Inpatient Days Spent at a Long-term Care Facility is defined as the number of days an individual is admitted to a Long-term Care Facility.
- 12e. Placement in Long-term Care Facility, which will be defined by (1) observed placement in a long-term care facility, or (2) the time (days) from enrollment to when individual is admitted to a Long-term Care Facility.
- 12f. Days Spent Receiving Hospice Benefit is defined as the number of days an individual is receiving hospice care regardless of location.

6. SAMPLE SIZE

6.1 Primary Outcomes

The assumptions used to determine sample size for testing differences among the three treatment arms for the two primary outcomes were: 1) type I error 0.05/6 = 0.0083 adjusted for 3 treatment comparisons (HSDC vs. CBDC, HSDC vs. UC and CBDC vs. UC) times 2 primary outcomes, 2) standard deviation (SD) for NPI-Q severity of 6.5 units and SD for MCSI of 6.7 units, 3) treatment difference for HSDC vs. CBDC of 1.5 units, 4) treatment difference for HSDC and CBDC vs. UC = 3, and 5) lost to follow up over 18 months because of death and dropout of 25%. The detectable treatment difference of 1.5 units for HSDC vs. CBDC is based on data reported in the literature suggesting a minimally clinical important difference (MCID) ranging from 2.8 to 4.0 (Mao et al., 2015). Given that both intervention arms will be receiving an intervention, we reduced the detectable difference in half to 1.5. Data from two studies (Callahan et al., 2006; Reuben, 2016) indicate that the expected benefit between HSDC and UC would be 3.2 units for NPI-Q Severity score. Based on these data we used a difference of 3 units for comparisons with UC. The estimates of the standard deviations (SD) for the 18-month NPI-Q Severity score and MCSI are 6.5 and 6.7, respectively, based on UCLA pilot data. An 18-month censoring rate of 25% was assumed because of death (15%) and dropout (10%). The

death rate was taken from fee-for-service claims data on 274 UCLA patients with a mean survival time of 1.5 years.

Sample size was first determined for the comparison between HSDC and CBDC because the effect size is smaller. The goal was to achieve at least 90% overall power for testing both outcomes. Testing each outcome at 95% power with a sample size of 1000 subjects per treatment group (adjusted for censoring) gives at least 90% overall power for testing both outcomes. Given these sample sizes, the sample size is 150 (for UC adjusted for censoring) for comparisons with the two interventions giving an overall power of at least 90%. For conservativism, we assumed the same effect size for both interventions compared with UC even though we hypothesize that HSDC will be superior to CBDC. The total sample size for the trial is 2150 individuals (adjusted for censoring).

Table 2: Power and Sample Size for the Two Primary Outcomes

Comparison	Difference to be detected	Power for MCSI	Power for NPI-Q Severity	Overall Power	Adjusted Sample Sizes
HSDC vs. CBDC	1.5	95%	95%	90%	1000 per group
HSDC or CBDC vs. UC	3.0	95%	95%	90%	1000 HSDC and CBDC vs. 150 UC

The inflation of sample size for censoring due to deaths and losses is conservative because the data will be analyzed using longitudinal methods under the MAR assumption that include all available data on each participant. Considering the first patient follow-up for the two primary outcomes is at 3 months, the effective censoring rate for analytical purposes is approximately 5% for the primary analysis (i.e., those without any primary outcome follow-up data).

Heterogeneity of treatment effects (HTE) for the two primary outcomes will be assessed in six subgroups of participants: high vs. low patient function by FAQ, high vs. low patient function by ADL, high vs low NPI-Q Severity, high vs low MCSI at baseline, spouse caregiver vs. other caregiver, and white non-Latino vs. nonwhite or Latino. The study will have approximately 90% power to detect interactions (differences in treatment effects between subgroup categories) of about 3 units for both NPI-Q severity and MCSI for the HSDC vs. CBDC comparisons. Interactions on the order of about 6 units can be detected with approximately 90% power for comparisons of HSDC and CBDC with UC for the two primary outcomes. These detectable interactions assume a conservative Bonferroni adjusted type I error of 0.05/(6x3x2) = 0.0014 per outcome to control for multiplicity. (In the analysis, we will conduct one global test of interaction per subgroup for 6 subgroups). Using an anchor-based approach that compares changes in NPI-Q measures with other clinical measures, the MCID for the severity scale is 3.2 (Mao et al., 2015). Hence, based on this range of MCID, the study will have adequate power to detect clinically meaningful differences in subgroups for comparisons between the two interventions. However, only large interactions can be detected for the comparisons with UC because of the smaller sample size in this arm. Although equal subgroup sample sizes were assumed in the calculations, power will not be greatly affected if the imbalance in subgroup sample sizes is not more extreme than a ratio of 2:1.

6.2 Secondary and Tertiary Outcomes

For the three secondary outcomes (NPI-Q Distress, caregiver depression (PHQ-8), Caregiver self-efficacy), we determined the detectable effect sizes assuming a Bonferroni adjusted type I error of 0.05/(3x3) = 0.006 to control for multiplicity. The detectable effect sizes with 90% power for the three continuous measures (NPI-Q Distress, caregiver depression (PHQ-8), Caregiver self-efficacy) are on the order of 0.20 for testing HSDC vs. CBDC. For testing HSDC/CBDC vs. UC, there will be approximately 90% power to detect effect sizes on the order of 0.40. Thus, we will have good power to detect small to moderate effects for the three continuous outcomes.

Power determinations for tertiary outcomes were not performed because these are considered exploratory. Sample size and power were calculated using Power and Sample Size (PASS) software (Kaysville, UT, 2013).

The goal attainment scaling (GAS) tertiary outcome will be collected only from caregivers (in the two active intervention groups) enrolled through December 15th, 2020. Goals will be set at baseline and will be assessed at two follow-up time points: at 6-12 months and 15-18 months after enrollment. After considering the expected loss to follow up, we anticipate 80% power (appropriate for a tertiary outcome) to detect a 10% difference between HSDC and CBDC arms.

7. INTERIM MONITORING

Interim monitoring will focus on safety, recruitment, adherence to protocol, baseline comparability of treatment groups, data quality, and uptake of the assigned intervention. A set of monitoring tables will be generated for this purpose (see DSMB Tables, Listings and Figures). Because this is a minimal risk trial and the follow-up period is relatively short (18 months), no monitoring for efficacy or futility is being proposed.

7.1 Safety Monitoring

Because this is a minimal risk trial, we will ascertain self-report (by caregiver) of SAEs of death and hospitalizations at each outcome assessment time point, which will be monitored by the DSMB. At the end of the follow-up, we will confirm any between-group differences in self-reported SAEs by examining data on mortality and hospitalizations from CMS data files and site-obtained utilization data. In addition, if a caregiver's, PHQ-8 score is >14, Research Assistants will immediately (that day) notify the site's clinical director (or covering physician) who can decide on next appropriate steps.

8. DATA COLLECTION AND DATA FREEZE

Data collected from date of study initiation (6/18/2019) to last date of 18-month study follow-up interview (08/21/2023) will be used. The trial database for primary, secondary and safety outcomes was closed/frozen on 4/9/2024 and considered ready for final analysis. Data lock for tertiary outcomes will be done at a later date, e.g., after resolution of fidelity data and receipt of CMS data.

9. ANALYSIS PLAN

This section describes the analysis of the primary, secondary and tertiary outcomes, including safety. The analytic plan will be modified as necessary up to final data lock and before study unblinding. Results will be reported to clinicaltrials.gov within one year of obtaining the primary outcome on the last patient. The trial has been registered on clinicaltrials.gov: NCT03786471.

Data analysis will be performed by the unblinded statisticians from the Data Coordinating Center. Current versions of SAS and R will be used for all analyses.

9.1 General Approach.

Nominal and ordinal categorical variables will be summarized using frequencies and percentages. Continuous variables will be summarized with the following descriptive statistics: N, mean, standard deviation, median, minimum, maximum, interquartile range, and range. Kaplan-Meier methodology will be used to summarize time-to-event variables (e.g., mortality). If estimable, 25^{th}, 50^{th} (median), and 75^{th} percentiles will be presented. No imputation of missing data will be performed for the primary analysis of the primary and secondary outcomes. Diagnostic tests and sensitivity analyses will be performed. Parametric distributional assumptions will be checked. If assumptions fail, other distributions will be considered prior to transformations and non-parametric methods.

9.2 Comparability of Baseline Characteristics.

Distributions of baseline demographic and clinical characteristics will be summarized. Comparability for continuous variables will be examined graphically and by summary statistics (means, medians, quartiles, etc.). Categorical variables will be examined by calculating frequency distributions.

9.3 Analysis of Primary Outcomes

All analyses will be according to intent to treat or as-randomized. The analysis of the two primary outcomes – NPI-Q Severity and MCSI scores – will be done using a constrained longitudinal repeated measures analysis (Fitzmaurice et al., 2011) based on maximum likelihood methods adjusted for the stratified randomization by site. Individuals missing both baseline and all follow-up scores for the primary outcomes will be excluded from the analysis; data for PLWD who had a change in caregiver will be censored at the time of caregiver change. Treatment differences, averaged over all follow-up time points, will be summarized by least square means and 97.5% confidence intervals. Significance testing for each outcome will be done by the Hochberg (1988) procedure using an overall type I error of 2.5% (2-sided) to control for the 3 pairwise treatment comparisons, i.e., testing each outcome at 2.5% will maintain the overall type I error at 5% (2-sided). This analysis will be implemented as a mixed effects repeated measures model which uses all available records under the assumption that visit-specific data are missing at random (MAR). Inclusion of the baseline, 3, 6, 12, and 18-month data as outcomes in the model will assist in meeting the MAR assumption. We will test for the treatment by time interaction at the 10% significance level. If significant, we will also report treatment differences at each time point. We will also examine for treatment by site interactions, although there may not be power to detect them. Sensitivity analyses will investigate the MAR assumption, such as methods that jointly model missingness and outcome distributions, e.g., pattern mixture models (Henderson et al., 2000; National Research Council, 2010).

We will conduct an exploratory analysis to test whether giving any intervention is superior to usual care if there is no significant difference between the two interventions, and if neither intervention is superior to usual care for either or both outcomes. We will test this hypothesis using the same longitudinal repeated measures analysis described for the primary outcomes at the 0.05 (2-sided) level of significance uncorrected for multiplicity. The difference between giving any intervention and usual care will be reported as least square means with 95% confidence intervals.

9.3.1 Heterogeneity of Treatment Effects (HTE)

HTE for the primary outcomes will be assessed in six key subgroups of participants: high vs. low patient function by FAQ, high vs. low patient function by ADL, high vs low NPI-Q Severity, high vs. low MCSI at baseline, spouse caregiver vs. other caregiver, and white non-Latino vs. nonwhite or Latino (defined by PLWD for NPI-Q Severity and by caregiver for MCSI outcome). Determination of cutpoints for HTE subgroups is presented in Appendix A. These subgroup analyses are aimed at determining whether there is differential effectiveness of the interventions among patients with more and less severe disease, perhaps elucidating a preventive (if effect is greater among those with less severe disease) versus therapeutic (if effect is greater among those with more severe disease) effect. Similarly, the high vs. low distress subgroup comparison will determine whether the intervention is more effective among those with more distress at baseline (a therapeutic benefit) versus less distress (a preventive benefit). Evidence of HTE will be based on tests of interaction within the longitudinal model structure described above. Because six tests of interaction will be conducted for each outcome, we will use the Hochberg procedure to control for multiplicity using an overall type I error of 2.5% (2-sided) for each outcome; subgroup treatment differences will be reported using 99% confidence intervals.

Exploratory analyses to assess HTE for the primary outcomes will also be conducted in demographic subgroups (age, race and ethnicity) as well as the effects of the COVID-19 pandemic and not adjusted for multiplicity.

9.4 Analysis of Secondary Outcomes

The analysis of the continuous secondary outcomes – NPI-Q Distress, caregiver depression (PHQ-8), Caregiver self-efficacy– will be analyzed like the primary outcomes. The Hochberg procedure will be used to control for the testing of each secondary outcome using an overall type I error of 5% (2-sided); treatment differences will be reported using 95% confidence intervals.

9.5 Analysis of Tertiary Outcomes

The approach to the analysis of these outcomes is briefly presented; however, the actual analytic methods used will in most cases depend on the distribution of the data. A more complete analytic plan will be finalized at the time of data lock for these outcomes and the SAP will be amended accordingly. Before any analyses are conducted, data will be summarized by means, medians, ranges and box plots.

Continuous Outcome Scales

- Montreal Cognitive Assessment (MoCA) measured at 18 months, range 0-12
- Functional status measured at 18 months by Functional Activities Questionnaire (FAQ) , range 0-30, and activities of daily living (Katz ADL), range 0-6.
- Dementia Burden Scale-Caregiver (DBS-CG) measured at 3, 6, 12 and 18 months is a composite of the NPI-Q Distress, MCSI, and PHQ-9 scales with items transformed linearly to be on a 0-100 possible range and then averaged with higher scores indicating higher caregiver burden.
- Quality of Life in Alzheimer's Disease (QOL-AD) measured at 18 months, range 4-52.
- Caregiver’s Satisfaction with Dementia Care measured at 3, 12 and 18 months, range 11-55.
- Positive Aspects of Caregiving (PAC) measured at 6 months, range 0-44.

These data will be analyzed like the primary outcomes using longitudinal models.

Binary Outcomes

- Clinical Benefit (measured at 3, 6, 12 and 18 months) is a **binary measure** of patient symptoms using the NPI-Q severity scale (the only patient outcome anticipated to benefit from the program) and caregiver symptoms using the DBS-CG scale. Benefit on the NPI-Q severity scale is defined as having a score of ≤ 6 (the lowest tertile of symptoms) or improving by at least 3 points (the MCID). DBS-CG benefit is defined as having a score of ≤ 18.8 (the lowest tertile of symptoms) or improving by at least 5 points (the MCID). Clinical benefit is defined as having benefit on either or both component scales.

  The specific endpoint for analysis remains under consideration; candidates include attainment of clinical benefit at or before a specified assessment, (discrete) time to first observed benefit, and the full set of measured clinical benefit data. Depending on the choice of endpoint, discrete time survival models, logistic regression, and GEE and mixed models for longitudinal repeated measures will be considered.

- Goal attainment scaling (measured at 6-12 and 15-18 months), defined as whether a person's individual goals are achieved as a result of the study intervention, will be measured using a 5-point goal. GAS describes the person's expected level of goal achievement over a specified timeframe, ranging from much worse than expected (scored as -2) to much better than expected (scored as +2). Scales are dynamically set according to a person's needs, while measurement of attainment is standardized. GAS is only determined for the two intervention arms: HSDC and CBDC.
  - Primary GAS outcome is the proportion who attained first identified goal at either 6 or 18 months.
  - An exploratory GAS outcome will be a repeated measures analysis of the proportion who hit target at 6 and 18 months.

GEE and mixed models for longitudinal repeated measures data will be considered.

Count Outcome

- Caregiver Rating of Dementia Care Quality (measured at 12 months), a 10-item, yes/no instrument that contains content from ACOVE/PCPI/AAN quality measures. The outcome is a count of the number of yesses.

These data will be analyzed by Poisson, Negative Binomial, hurdle models or rank based methods depending on the distribution of the counts.

Time spent at home is calculated as 540 minus (1) the number of days in a health care facility, including hospital, nursing home, acute rehabilitation, and inpatient hospice, and (2) the number of days not alive. The analytic approach will depend on the distribution of the outcome. As an initial approach to the analysis, we will consider a negative binomial model assuming a potential follow-up time of 540 days. Alternatively, it may be necessary to instead analyze the complement of number of days spent at home (days dead plus days spent in a health care facility – hospital, nursing home, acute rehabilitation, and inpatient hospice) and/or use a zero-inflated model.

Mortality will be analyzed by the Cox regression model adjusted for site for time to event data; mortality rates will be summarized by Kaplan-Meier estimates.

CMS data for participants include the number of days spent in:

- Acute Care Hospital
- Inpatient Rehabilitation Facility
- Skilled Nursing Facility
- Long-term Care Facility (including time to placement)
- Hospice care

These data will be summarized by counts, means, medians and ranges. Because of the possibility of many zero counts, Poisson, Negative binomial and hurdle models will be considered. Also, because of censoring we will also consider analyzing the data similarly to time spent at home. Time to placement in a long-term care facility will be analyzed using survival methods.

9.6 Missing Data

Several strategies will be implemented to address the issue of missing data during this study. Prevention is the most obvious and effective manner to control bias and loss of power from missing data (National Research Council, 2010) We will also minimize missing data and other errors during data collection. This protocol will follow the intent-to-treat principle, requiring follow-up of all participants randomized regardless of the actual treatment received (Lachin, 2000) The primary source of missing data will result from death since we expect relatively few dropouts (< 10%). Whenever a participant drops out of the study, we will document the specific reason for dropout, who decided that the participant would drop out, and whether the dropout involves some or all types of participation. For example, some participants may drop out of the patient/caregiver-reported outcomes but permit use of claims data to determine some secondary and tertiary outcomes.

The follow-up period of 18 months was set to capture the time-period during which treatment benefits would be maximized while minimizing the impact of deaths. Also, the timing and frequency of follow-up assessments at 3, 6, 12, and 18 months was designed to minimize missing primary outcome data on participants by having measurements at early time points.

Because the study population has dementia, the primary source of data collection (including primary and all secondary outcomes) will be from caregivers. To minimize missing data, all follow-up data will be collected by telephone, making it much easier, compared to in-person visits, for participants (and caregivers) to complete the outcomes assessments. Because most instruments, including the two primary outcome measures, are self-administered, we can also offer the options of collecting these data by fax, secure website, or secure email, which will also reduce the likelihood of missing data. We will also employ a risk-based approach to trial monitoring with timely data entry combined with weekly missing data reports that will trigger protocols for tracking and obtaining missing data items or outcome assessments data (FDA Guidance for Industry, 2013).

Despite these prevention efforts, it is reasonable to assume missing data will occur. Our proposed primary and secondary analyses make use of all available data and are valid under the assumption that missing data will be missing at random (MAR) (Diggle, et al., 2002; Molenberghs et al., 2004). We will evaluate the plausibility of this assumption by determining the

extent of missing data and use logistic regression to identify factors associated with missing data. Sensitivity analyses will investigate the MAR assumption, such as methods that jointly model missingness and outcome distributions, e.g., pattern mixture models (Henderson et al., 2000; National Research Council, 2010).

9.7 Analysis of Safety

Safety analyses will involve tabulating the occurrence of serious adverse events (deaths and hospitalizations) and unanticipated problems among the three treatment groups. Hospitalizations will be summarized by counts, frequency distributions and event rates per person year of follow-up. Mortality rates will be calculated by the method of Kaplan-Meier. Differences in mortality rates will be analyzed by Cox regression model and differences in hospitalization rates by either Poisson or negative binomial regression models adjusted for site. A p-value of 0.05 (2-sided) will be used for the safety analyses.

REFERENCES

Callahan CM, Boustani MA, Unverzagt FW, Austrom MG, Damush TM, Perkins AJ, Fultz BA, Hui SL, Counsell SR, Hendrie HC. Effectiveness of collaborative care for older adults with Alzheimer disease in primary care: a randomized controlled trial. Journal of the American Medical Association. 295(18):2148-57,2006. PubMed PMID: 16684985.

Diggle PJ, Heagerty P, Liang K-Y, Zeger S. Analysis of Longitudinal Data. Oxford Science publications, Clarendon Press, Oxford, 2002 (2nd ed.)

FDA Guidance for Industry. Oversight of Clinical Investigations - A Risk-Based Approach to Monitoring. August 2013.

Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis. Hoboken, NJ: Wiley, 2011.

Henderson R, Diggle P, Dobson A. Joint modeling of longitudinal measurements and event time data. Biostatistics.1:465–480, 2000.

Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 75:800-802, 1988.

Lachin JM. Statistical considerations in the intent-to-treat principle. Controlled Clinical Trials. 21:167-89, 2000.

Mao HF, Kuo CA, Huang WN, Cummings JL, Hwang TJ. Values of the Minimal Clinically Important Difference for the Neuropsychiatric Inventory Questionnaire in Individuals with Dementia. Journal of the American Geriatric Society. 63(7):1448-52, 2015. doi:10.1111/jgs.13473. Epub 2015 Jun 5. PubMed PMID: 26046666.

Molenberghs G, Thijs H, Jansen I, et al. Analyzing incomplete longitudinal clinical trial data. Biostatistics. 5(3):445-64, 2004.

National Research Council. The Prevention and Treatment of Missing Data in Clinical Trials. Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press, 2010.

Reuben DB. Final Progress report for 1C1CMS330982 “UCLA Alzheimer’s and Dementia Care: Comprehensive coordinated, patient-centered” submitted to CMS March 31, 2016.

APPENDIX A. DETERMINATION OF CUTPOINTS FOR HTE SUBGROUPS
Heterogeneity of Treatment Effects (HTE)

HTE for the primary outcomes will be assessed in six key subgroups of participants:

1. higher vs. lower patient function by FAQ
2. independent vs. dependent patient function by Katz ADL
3. more vs. less NPI-Q Severity
4. more vs. less MCSI at baseline
5. spouse caregiver vs. other caregiver
6. white non-Latino vs. nonwhite or Latino

Determination of Subgroup Cutpoints

1. Higher vs. Lower Patient Function by FAQ

   Patient functional status is measured using the Functional Activities Questionnaire (FAQ) and Katz Activities of Daily Living (ADL).

   - Scoring: 0-30, higher score indicating more functional dependence
   - Determined Cutpoint: 19 (65.73%, upper tertile)
   - Categories: higher patient function (0-19) vs. lower patient function (20-30)

2. Independent vs. Dependent Function by Katz ADL

   - Scoring: 0-6, higher score indicating more functional independence
   - Determined Cutpoint: 6
   - Categories: dependent patient function (0-5) vs. independent patient function (6)

3. More vs. Less NPI-Q Severity
   The severity of symptoms of psychopathology in persons with dementia will be measured by the Neuro-Psychiatric Inventory Questionnaire – Severity (NPI-Q Severity).

   - Scoring: 0-36, higher score indicating more severe PLWD symptoms
   - Determined Cutpoint: 12 (66.21%, upper tertile)
   - Categories: less severe (0-12) vs. more severe (13-36)

4. More vs. Less MCSI at baseline
   Caregiver distress/strain is measured using the modified caregiver strain index (MCSI).

   - Scoring: 0-26, higher score indicates greater level of strain
   - Determined Cutpoint: 13 (66.50%, upper tertile)
   - Categories: less strain (0-13) vs. more strain (14-26)

5. Spouse caregiver vs. Other caregiver
   Caregiver relationship is self-reported during baseline interviews.

6. White Non-Latino vs. Nonwhite or Latino
   Ethnicity is self-reported during baseline interviews.

- Categories: less severe (0-12) vs. more severe (13-36)

4. More vs. Less MCSI at baseline
   Caregiver distress/strain is measured using the modified caregiver strain index (MCSI).

   - Scoring: 0-26, higher score indicates greater level of strain
   - Determined Cutpoint: 13 (66.50%, upper tertile)
   - Categories: less strain (0-13) vs. more strain (14-26)

APPENDIX B: DSMB TABLES, LISTINGS AND FIGURES

The general types of table, listings and figures (TLFs) for the Open DSMB Report will include

1. Intake graph showing cumulative observed and expected number screened and enrolled
2. Consort diagram showing flow of screening and recruitment data
3. Reasons for exclusion overall and by site
4. Entry characteristics of enrolled participants overall and by site
5. Completeness of follow-up overall and by site: person-years of follow-up, losses and withdrawals
6. Fidelity of treatment
7. Frequency of serious adverse events (SAEs) overall and by site: mortality and hospitalizations
8. Listing of SAEs
9. Protocol Deviations

The closed DSMB Report will mirror the Open Report and present data by randomized treatment group designated as A, B and C. Additional tables will include information about completeness of primary and secondary outcome data. Because there is no interim monitoring for efficacy and futility, no outcome data will be presented to the DSMB by treatment unless requested by the DSMB.