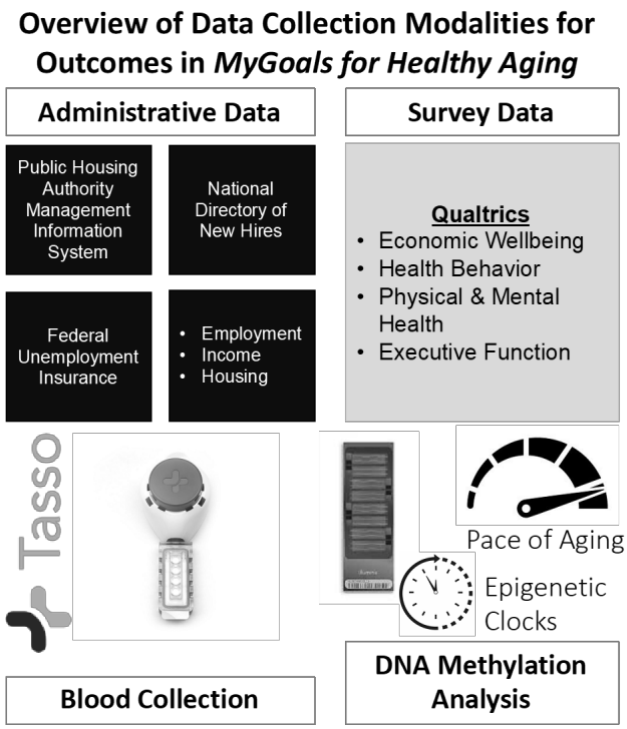
SECTION 4.3. STATISTICAL DESIGN AND POWER

This section elaborates on the survey, administrative, and biological measures of aging introduced in the **Approach** section. It provides information on the approach in terms of intervention timing, multiple outcome measures, the statistical model, contingency plans, and the power analysis. Details of primary and secondary outcome measures can also be found in **Section 4**.

MyGoals for Healthy Aging differs from *MyGoals for Employment Success* in two ways. (1) The duration of the *intervention* is three years rather than two; and (2) The total follow-up time is six years rather than two years. The longer intervention and follow-up time are both essential to test meaningful changes in aging-related outcomes, which are slower to manifest than economic changes.

We hypothesize that our intervention will improve social and economic well-being, and these improvements will impact healthy aging. Our metrics include measures of economic well-being and health behavior (**Aim 1**), mental and physical health and executive function (**Aim 2**) and biological aging (**Aim 3**) as well as long-term administrative data on income and mortality (**Aim 4**) collected from administrative data capture and results from administration of validated survey instruments, i.e., a Qualtrics survey that utilizes extensive skip questions six years post-enrollment (**Appendix**) and laboratory analysis of blood samples. The **Figure** at right provides a schematic overview of data collection modalities and the outcome domains derived from each.



Administrative data from the Public Housing Authority Management Information System are also replicated to allow tracking of participants who moved out of Section 8 housing. This includes information on family composition, welfare receipt, and employment by job type.

Measures of health insurance, diet, and exercise were not part of the original *MyGoals for Employment Success* Qualtrics survey. These will only be programmed into the survey if our study is funded. Therefore, these measures can be found as separate attachments to the **Appendix** file.

Please note again that our funding for cohort maintenance ended in 2020; thus, this is our final opportunity to conduct *MyGoals for Healthy Aging*.

4.3.1. Measures of economic wellbeing (Aim 1).

Measurements of economic wellbeing in *MyGoals for Healthy Aging* come from administrative data capture and survey measurements.

For both the treatment and control groups, the Public Housing Authority's Management Information System provides information on employment status, current address, housing subsidy information, and household composition. All employment and earnings in the formal labor market are tracked electronically using data from the National Directory of New Hires and the Unemployment Insurance datasets.

We will also collect detailed survey data on employment, income, welfare service receipt, economic hardship, schooling, and criminal justice system contact at 6 years post-randomization (project years 1-4). Employment questions on the survey cover a broad set of dimensions (e.g., employment) and domains (e.g., for pay work versus volunteer, job type, date of hire). Participants are probed regarding barriers to employment (e.g., childcare responsibilities or lack of transportation) and details relevant to capturing information possibly missing in the Unemployment Insurance administrative data (e.g., work in the informal sector for cash).

Data on income include employment earnings, earnings from others in the household, and welfare receipt. These questions are designed to allow for comparisons between survey and administrative data and obtain

details about each participant's total income otherwise unavailable from administrative data sources. Examples of other variables pertaining to income or wealth include details on home ownership and health insurance.

Detailed domains of economic hardship include whether the participant: 1) has cut the size of meals or skipped meals because s/he was unable to afford food, 2) had to move in with other people because of financial problems, 3) borrowed money from friends or family, 4) gone without a phone, 5) taken out a payday loan to cover bills, and 6) gone without needed medical care because s/he thought that the care would be unaffordable. Other questions can be found in the **Appendix**.

We collect data on enrollment in job training, non-degree programs, and degree programs. Finally, we capture details on criminal justice system interactions, including any contact, date of contact, felony convictions, and job denials for having a criminal record. These detailed economic wellbeing data will allow us to (1) develop a powerful composite economic wellbeing trial endpoint combining information across the many aspects of participants' economic lives possibly affected by the intervention and (2) explore in secondary analyses potential mediating pathways for the intervention's effects.

4.3.2. Survey measures of diet and exercise (Aim 1) and sleep, loneliness, psychological stress, depression, obesity, and health-related quality of life (Aim 2).

We will measure a range of health outcomes related to pathways of poverty by administering validated survey instruments by telephone. MDRC has found that surveys greater than 45-minutes in duration lead to increases in missing variables, particularly in socio-economically disadvantaged populations. Thus, shorter instruments were prioritized. The measures selected for inclusion in the survey are summarized in **Table 1** below. Participants will complete surveys 6 years post randomization (grant years 1–4)

Table 1. Survey measures for health insurance, diet, and exercise (**Aim 1**) as well as sleep, loneliness, stress, depression, obesity, and health-related quality of life (**Aim 2**).

Aim 1 Survey-measured Outcomes		
Diet	Eating at America's Table (EATS)	The Eating at America's Table (EATS) survey was developed by the National Institutes of Health as a nutrition survey. It is a brief, 7-item survey of fruit and vegetable consumption. We will use the "all day" scoring tool developed by the NIH to score the instrument.
Exercise	Behavioral Risk Factor Surveillance System (BRFSS) exercise measure.	The Behavioral Risk Factor Surveillance System (BRFSS) exercise measure is a single-item question on exercise. While not validated, our expert advisory panel noted that the purpose of the measure was to add a distal, descriptive outcome to assess whether the intervention may be influencing health behaviors. It asks, "During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?"
Aim 2 Survey-measured Outcomes		
Sleep	Sleep Quality Scale	The sleep quality scale is a single-item index of sleep quality that correlates highly with the gold-standard in the field, the Pittsburgh Sleep Quality Index [r=0.9 (162)].
Loneliness	UCLA 3-Item Loneliness Scale	The UCLA 3-Item Loneliness Scale is a brief measure of loneliness (163). The 3 items were obtained from a widely used and validated 20-item scale. The 3-item scale has been independently validated and is designed for inclusion in survey instruments.
Stress	Perceived Stress Scale	The Perceived Stress Scale (PSS) is a 10-item measure of psychological stress. The PSS has been extensively validated, including in ethnic minority populations (164, 165), is widely used, and responses strongly correlate with behavioral risk factors for health (e.g., smoking) and biological markers of stress, such as hsCRP, both of which are included as measures in our study.
Depression	Patient Health	The Patient Health Questionnaire (PHQ-9) is a 9-item measure of clinical depression that has been validated, including in ethnic minority populations (166–168), and is widely used.

	Questionnaire (PHQ-9)	
Obesity	BMI	Participants will be asked to self-report height and weight
Health-related Quality of Life	EuroQol 5D-5L	The widely used, including in one study by the PI (26), and well-validated EuroQol 5D-5L is a brief 5-item measure of health-related quality of life (169). Health-related quality of life is a multi- dimensional measure of health that is scaled from 0 (a state equal to death) to 1 (a state equal to perfect health). It is used in comparative effectiveness studies as the primary means of capturing morbidity across 5 domains. This measure is a primary outcome measure required for proper comparative evaluation of our study by the United States Department of Housing and Urban Development.

4.3.3. Executive function measures (Aim 2).

Our executive function assessment is meant to determine whether, on average, treated participants differ from control participants with respect to executive function. The BRIEF-A is an extensively validated measure of executive function, including in racially diverse populations (144). A team of expert consultants to MDRC developed a shorter version of the BRIEF-A tailored to the SMART Goal domains the coaches are trained to assess (102). In our preliminary analysis of year 2 survey data, the extracted measures were combined and assessed and T scores were computed. The questions were then validated in the *MyGoals* cohort and aligned with the executive function intervention. The consultants Richard Guare, PhD, Peg Dawson, PhD, and Colin Guare, PhD assisted MDRC to implement their executive function coaching program. They have extensive experience in evaluating ethnic minority populations in executive function domains (102) (examples shown in **Table 2)**. In the preliminary post-randomization year 2 survey data, the two BRIEF-A subscales (Behavioral Regulation and Metacognition and Negativity, Inconsistency, and Infrequency) are included. These were internally consistent (Cronbach's α = 0.92 and α = 0.87). We will repeat our assessment of the performance of the BRIEF-A in post-randomization year 6.

Table 2. Examples of executive function coaching domains contained in the BRIEF-A. These questions align with the executive function domains within which coaches are trained.

Component	Sample Questions*
Future Orientation	I know I need … a job … and really think I should work on finding one
Goal Development	I set long-term employment goals that I hope to achieve.
Overcoming Barriers	Even when I face challenges, I continue to pursue my … goals
Task Initiation	I have problems getting started on my own
Planning	People say I don't think before acting
Emotional Control	Lost your temper with someone other than friends or family
Completion	I have trouble finishing tasks (such as, chores and work)
Stress management	I get overwhelmed by large tasks
Diligence	I make careless errors when completing tasks
Futility	[Did] not to apply for a job because you didn't think you would get [it]
Perseverance	Missed an appointment … for a reason other than you were sick or ill.

*Likert scale asking frequency (e.g., "Once per Week") or agreement (e.g., "Strongly Agree").

4.3.4. Biological aging measures (Aim 3).

Biological aging is the gradual and progressive loss of system integrity that occurs as we grow older, driving risk for disease, disability, and mortality (61). Biological processes of aging are ongoing from early life and are affected by the environment (63, 67, 170–172); psychosocial stress, environmental toxicants, and other insults can accelerate biological processes of aging. These processes originate with the accumulation of molecular changes, including loss of proteostasis, mitochondrial dysfunction, telomere attrition, cellular senescence, and changes to DNA methylation; all termed the "hallmarks of aging" (10). Critically, the hallmarks of aging are modifiable; in animals, interventions that slow or reverse accumulation of aging hallmarks can extend healthy lifespan (173). Efforts are now underway to translate this science into therapies to slow human aging with the goal of preventing or delaying multiple chronic diseases (9, 71). A range of methods has been developed to quantify the biological processes of aging in humans enabling us to test interventions (90).

Pace of aging methods were developed specifically to provide surrogate endpoints for RCTs testing therapies proven to slow aging and prevent disease in model organisms (117, 174, 175). Such surrogate endpoint measures are essential because aging-related diseases take decades to develop. Therefore, short-term readouts on intervention effectiveness are a critical priority. Over the past decade, DNA methylation has emerged as the molecular substrate most promising for analysis of biological aging in RCTs (88).

DNA methylation clocks. The most prominent DNA methylation-based measures of aging are known as "clocks"—because they are strikingly accurate at predicting the mean chronological age (the amount of time that has passed since birth) of a sample population. Differences between biological age, measured by DNA methylation clocks, and true chronological age of an individual can quantify the extent of biological aging that the person has experienced (92).

Table 3. Quantifications of biological aging for the *MyGoals for Healthy Aging* Study

Measurement Development	Analysis and Interpretation	Hypothesis
GrimAge Clock. The GrimAge was developed from DNA methylation analysis of mortality in the Framingham Heart Study Offspring Cohort. First, DNA methylation algorithms were developed to predict levels of a panel of blood proteins. Second, a meta-algorithm was developed to predict mortality from DNA-methylation-predicted levels of the blood proteins, DNA-methylation-predicted levels of tobacco exposure, sex, and chronological age. Third, meta-algorithm predictions were transformed to take on values equivalent to chronological age by normalizing the distribution of predicted values to match the distribution of chronological age values. The final GrimAge algorithm is applied to whole-genome DNA methylation data and produces a single clock-age value that represents the age at which a person's mortality risk would be "average" within the Framingham cohort.	For analysis, GrimAge is first regressed on chronological age and residual values are predicted. Residual values >0 indicate more advanced ("older") biological age and increased risk for disease and death. Residual values <0 indicate less-advanced ("younger") biological age and reduced risk. Models testing effects on GrimAge will include a minimum set of covariates including chronological age and sex.	**Hypothesis:** GrimAge residual values in the *MyGoals* treatment group will be "younger" than GrimAge residual values in the control group.
DunedinPoAm Pace of Aging. DunedinPoAm was developed from DNA methylation analysis of decline in organ system integrity in the Dunedin Longitudinal Study 1972-3 birth cohort. First, mixed-effects growth models were used to test within-individual change in 18 organ-system function tests* across ages 26-38 years. (An under-review update of the measure includes changes modeled through age 45 years). Second, for each cohort member, the slopes of change were combined across the 18 models to form a single composite measure of how much faster or slower that person's body was deteriorating relative to the cohort norm. This measure is termed Pace of Aging. Third, Pace of Aging was modeled from DNA methylation data collected at the end of the follow-up interval to derive the DunedinPoAm DNA methylation algorithm. The final DunedinPoAm DNA methylation algorithm produces a value that represents the ratio of the total decline in system integrity experienced by the individual during a calendar year over the average decline in the Dunedin cohort.	**Analysis and Interpretation.** Values >1 indicate a faster pace of aging and increased risk for disease and death. Values <1 indicate a slower pace of aging and reduced risk. Models testing effects on DunedinPoAm will include a minimum set of covariates including chronological age and sex.	**Hypothesis:** DunedinPoAm pace of aging values will be "slower" in the *MyGoals* treatment group as compared to the control group.

*The organ-system function tests included in the Pace of Aging were VO2max (cardiorespiratory fitness), FEV1 and FEV1/FVC ratio (lung function), mean arterial pressure, body mass index, waist hip ratio, gum health, leukocyte telomere length, BUN, creatinine clearance, C-reactive protein, HbA1C, lipoprotein (a), APoB100/ApoA1 ratio, triglycerides, HDL cholesterol, total cholesterol, white blood cell count

The focus of DNA methylation clocks was originally to predict chronological age (176, 177) but has evolved to predict age-related health declines, including disease and death (178–180). The most recent clocks, which are better mortality predictors than earlier generations, consistently indicate more advanced biological aging in lower SES and minoritized populations (19, 95, 97–99, 181). Recently, MPI Belsky developed a distinct measure focused on the pace of biological aging in younger populations rather than static biological age (23), further advancing the field. This new measure, DunedinPoAm, is more akin to a speedometer than a clock. It measures the speed of aging in an individual over the recent past and was designed specifically for use in RCTs because of its unique sensitivity to short-term changes in aging. DunedinPoAm is comparable to the 2nd-generation DNA methylation clocks as a predictor of morbidity and mortality (23, 93, 182). It is also more

sensitive to the impacts of socioeconomic factors than traditional clocks (23, 100, 22). Critically, the DunedinPoAm measure is predictive of morbidity and mortality across race/ethnic groups and also clearly identifies health disparities in the pace of biological aging (22, 99, 100, 182).

MyGoals for Healthy Aging will test hypotheses using two DNA methylation measures of aging: the GrimAge DNA methylation clock (180) and the DunedinPoAm pace of aging (23) because evidence indicates these are the best measures for quantifying social determinants of health effects on biological aging. They also show evidence of predicting the course of AD/ADRD (see Approach D.3.2). GrimAge and DunedinPoAm are described in detail in **Table 3** However, because whole-genome DNA methylation data can be used to compute many different aging measures, we will continue to monitor the literature and implement the measures with the strongest evidence at the time we conduct our analysis.

In addition to DNA methylation, *MyGoals* will use the blood samples to measure: (i) genome-wide single nucleotide polymorphisms (SNPs) to control for artifacts in the DNA methylation data (91, 183, 184); and (ii) C-reactive protein and hemoglobin A1c to measure proximal risk factors for chronic disease.

4.3.5. Influence of the intervention timing on the statistical approach. In March 2017, the first of the 1,798 participants was enrolled, administered the baseline survey, and began the intervention. Because enrollment was conducted on a rolling basis, a small portion of the participants are still receiving the 3-year intervention at the time of grant review (**Table 4,** left). We will also collect data for the 6th year post-randomization timepoint on a rolling basis beginning with the first participant enrolled. This way, each participant will have received approximately the same 6 years of follow-up post the 3-year intervention; thus, *our statistical approach is not impacted by rolling enrollment in our primary analyses*.

4.3.6. Approach to multiple outcome measures. Social policy experiments produce a wide array of potential medical and non-medical impacts and typically do not have a single outcome measure of interest. In *MyGoals for Healthy Aging*, our primary outcomes consist of 6 broad outcome domains (two per specific aim). Our analytic plan is to test hypotheses at the domain level. For domains comprising multiple measures, we will use the seemingly-unrelated regression approach, established by Kling et al. (146), to conduct joint analysis of the multiple outcomes to compute a single statistical test of association based on familywise adjusted p-values. Additionally, we will create aggregate indices, e.g., a modification of their economic self-sufficiency index. This index averages multiple measures of incentive income, earnings, employment, and public assistance (146).

Table 4. Enrollment start date, end date, and analytic sample by program site and treatment status. MyGoals for Healthy Aging, 2017-2019.

Site	Enrollment Start Date	Enrollment End Date	Program Status	Analytic Sample*
Baltimore	March, 2017	September, 2018	Treatment	373
			Control	375
Houston	February, 2017	July, 2019	Treatment	527
			Control	523

*Total after withdrawals (7 in Baltimore 3 in Houston).

In Aim 1, the domains are economic wellbeing (employment, income, housing, health insurance, and crime) and health behavior (diet, exercise, sleep).

In Aim 2, the domains are physical/mental health (sleep, loneliness, psychological stress, depression, obesity, health-related quality of life, C-reactive protein, and HbA1C) and executive functioning.

In Aim 3, the domains are the pace of biological aging (DunedinPoAm) and biological age (GrimAge Clock).

4.3.7. Approach to developing pre-specified models. We have budgeted time for our team to develop a package of pre-specified models as part of the program ramp-up in project year 1. These will be listed on our project website. Dr. Manly will develop pre-specified models for the BRIEF-A. Dr. Belsky will develop pre-specified models for the aging measures and biomarkers. Dr. Muennig will focus on broader measures of the social determinants of health using validated survey measures.

4.3.8. Statistical approach to primary and secondary outcome measures. All primary outcome variables are continuous. The statistical approach is therefore similar for each, and we describe them together.

(1) We will use two-sample t-tests to compare differences in mean values between the treatment and control groups. Because our study is an RCT, this produces unbiased results but does not offer the precision or flexibility of linear regression models. Such models can, for example, reduce noise associated with random variation in participant characteristics between the treatment and control groups.

Thus, for outcomes measured at both baseline and follow-up (e.g., the BRIEF), we will **(2)** calculate the change from baseline to follow-up and use the difference as the outcome variable in a regression analysis. There is only 1 participant per household, so we do not need to account for intra-household effects in the model.

(3) We will utilize an *intention-to-treat* (ITT) approach for each outcome of interest to measure the effect of eligibility for the intervention on health using the final domain-level measures of economic wellbeing, health behavior, physical and mental health, executive function (BRIEF-A), biological age (GrimAge Clock), and pace of aging (DunedinPoAm).

For primary outcome measures, which are all continuous in our study, we will use a linear regression model:

$$y_i = \beta_0 + \beta_1 Treatment_i + X_i^T \beta_2 + \epsilon_i$$

where y_i is the outcome of interest for individual i, $Treatment_i$ is equal to 1 for the treatment group and 0 for control group, and X_i represents covariates, which are included to increase the precision. Covariates will include demographic information, e.g., age, race, gender, education level, household composition, and work history at enrollment. They will also include baseline measures of the outcome of interest, when available.

Finally, because our analysis of primary and secondary outcomes assumes that *MyGoals for Healthy Aging* is a poly-intervention that serves as a single treatment (and each component works through a similar set of pathways), it is appropriate to use random selection in the treatment group as an instrumental variable (IV) approach to capture treatment-on-treated (TOT) effects. TOT effects are particularly useful from an academic standpoint, because they allow for more precise quantification of the impact of other social policies on health and for measuring health impacts in real-world *MyGoals for Healthy Aging* implementations that are either more or less successful than ours at improving economic outcomes.

For continuous outcomes, we will consider a two stage least squares estimation as follows:

$$y_i = \beta_0 + \beta_1 D_i + X_i^T \beta_2 + \epsilon_i$$

where y_i is the outcome of interest for individual i, D_i is equal to the ITT parameter estimated in Formula 1 divided by the regression-adjusted compliance rate, and X_i represents the same set of covariates as above.

To mitigate potential bias due to differential missingness in outcome values, we will apply inverse probability weighting, i.e., complete cases will be re-weighed by the estimated probability of missingness given baseline covariate values. This approach to missingness implicitly assumes that outcomes are “missing at random” (MAR), which is a reasonable assumption given the rich covariate information available. In a sensitivity analysis, we will evaluate the robustness of our estimates to violations of the MAR assumption using a pattern mixture model or instrumental variables approach (185). Missing covariate entries will be accommodated by multiple imputation.

We will examine **outliers** and, where possible, use tests of plausibility to remove values that are clearly erroneous. If deletions are necessary, they will be documented along with justification for omission. We will also present analyses showing how any such outlier would have impacted the final estimation. All other outliers will remain in the analysis, but will also be tested using sensitivity testing.

In devising an analytical approach, we will pre-specify analyses based not only upon desired analyses but also upon a number of **contingency plans** in the event of problems with the experiment (e.g., differences in the demographic characteristics of the participants at follow-up or changes in economic conditions). In addition to the outcomes described above, we will compute various composite risk scores. If we do not meet targets for enrollment or obtain consent, we will analyze both sites together for primary outcomes and limit our subgroup analyses.

4.3.9. Sex as a biological variable. Our analytic approach considers sex as a biological variable in three ways. **(1)** We will include sex as a covariate in our regression models testing treatment effects to account for potential differences in the numbers of men and women at follow-up between the intervention and control groups. **(2)** We will conduct stratified analysis, repeating tests of treatment effects separately by sex to explore/describe any differences in apparent effects of treatment. **(3)** We will formally test for differences in treatment effects by including sex*treatment group product terms in our regressions to test if treatment effects differ between men and women.

4.3.10. Approach to reporting null effects. In the event of null or difficult to explain findings related to health and cognition, the team will attempt to publish such findings in peer-reviewed journals. We will also post papers to preprint servers such as the MedRxiv to ensure open access of findings regardless of journal paywall policies. In addition to preprints, unpublished findings will be published on the planned *MyGoals for Healthy Aging* website at Columbia University. This will reduce both duplicative efforts by other teams and publication bias. Funds have been allocated for these efforts.

4.3.11. Statistical approach to creating the public-use dataset. To complete **Aim 4**, we need to not only create a dataset for public use but also set up the infrastructure for electronic data linkages and access to biological specimens. **Aim 4** adheres to the NIH's "Across the Lifespan Policy," which seeks to capture outcomes at all points in the human lifespan.

The primary statistical challenge in developing such a dataset is developing a matching algorithm that will reliably link identifiers in each subset of data. We will use the comprehensive identifiers for all participants (e.g., name, date of birth, and Social Security Number) to match participant records to death certificate records. We will then submit this record to the National Death Index at the Centers for Disease Control and Prevention for matching. We anticipate that only ~32 participants will die by the end of the study period, a number vastly underpowered for computing study outcomes. **Thus, while the deaths captured within the performance period will not be used for study outcomes, they will be used to ensure that future studies using the NDI linkage will be simple to administer and will be reliably matched using tested algorithms.**

The dataset we create will be stored electronically in a secure location at MDRC for future matching. We anticipate creating this initial linkage in PY 5. The linkage activities are free of charge for R01-funded investigators (NOT-OD-20-057), so we anticipate being able to update the linkage on a regular basis.

4.3.12. Statistical approach to secondary analyses. *MyGoals for Healthy Aging* will produce data not only useful for causal inference but also invaluable for insights into the biopsychosocial determinants of health among a highly disadvantaged population. Such data would have great utility even if our proposed intervention fails to produce an impact because they would uniquely apply to a highly disadvantaged population and would contain.

1. Variation in exposure to different components of the treatment. The *MyGoals for Healthy Aging* intervention comprises several components: income support, employment incentives, and executive function coaching. Some participants, because of personal preferences and/or when they enter/exit the intervention, will receive only some components. We will test whether treatment effects are different for subsets of participants exposed to different doses of each treatment component. Specifically, participants vary in their employment periods and usage of their executive function coaches. We will utilize this variation in treatment dosing to test whether treatment effects vary depending on how much of each intervention component the participant receives. Within our general regression framework, this test will be executed by adding product terms that test heterogeneity in the overall treatment effect according to the doses of the different intervention components a participant received.

2. Subgroup effects. We plan to explore the health impacts associated with variation in economic treatment effects: 1) by geographic subgroups (i.e., site), and 2) by sociodemographic subgroups (age, gender, race, prior labor force attachment, education level). Baltimore-site participants are expected to experience larger treatment effects than Houston-site participants because they are more disadvantaged than their Houston counterparts (**Table 5**, right). In terms of sociodemographic subgroups, we will explore whether relatively more disadvantaged participant subgroups experience stronger treatment effects. We will also investigate whether up-take of the interventions varies by subgroup (e.g., in our prior RCT, *Paycheck Plus*, women had better uptake of the intervention than men).

Table 5. Housing subsidies by site, MyGoals for Healthy Aging.

Subsidy	Baltimore	Houston
$1 - $599	6%	18%
$600 - $899	33%	38%
$900 - $1199	32%	28%
$1200 or more	29%	15%

3. Dose-response effects. Treated participants who stop engaging in any aspect of the program are allowed to re-engage any time they wish during the 3-year program period, potentially allowing us to estimate variation in exposure to the program as a whole.

4. External Validity. Economic data are collected continuously via electronic linkages before and throughout the period of performance. These data are captured even for participants who drop out of the trial before data

collection within *MyGoals for Healthy Aging*. These data will therefore enable us to test if this dropout is systematic with respect to participant characteristics and allow us to assess the study's external validity.

4.3.13. Statistical Power for Aims 1 and 2. We estimated power using the Minimal Detectable Effect Size (MDES).

Assumptions. All analyses assume a 1:1 assignment for treatment and control and a power of 80%. Although the attrition rate to date has been less than 20%, we conservatively assume a combination of attrition and missing/incomplete data will yield an effective sample size of 1200 individuals (i.e., 66% of the original cohort). In the power analysis, we calculated the minimal detectable difference between the treatment and control groups in the change of outcome measures at follow-up. With our existing sample size, we estimate that we can, at each site, detect 0.16 standard deviations (SD) of difference at a significance level of 0.05 using a two-sided test.

A treatment effect of this size after 6 years of post-randomization follow-up (3 years of intervention and 3 additional years of follow-up) is plausible. We analyzed 38 US welfare experiments conducted between 1965 and 2019 in a meta-analysis (25) and determined that the follow-up duration in *MyGoals for Healthy Aging* will be 25% longer than the median of the analyzed studies (72 mo for *MyGoals* vs. 54 mo for the median study). Also, *MyGoals for Healthy Aging* uses an economic intervention based on the *Work Rewards* RCT and is thus expected to generate an income dose larger than 90% of the other analyzed studies.

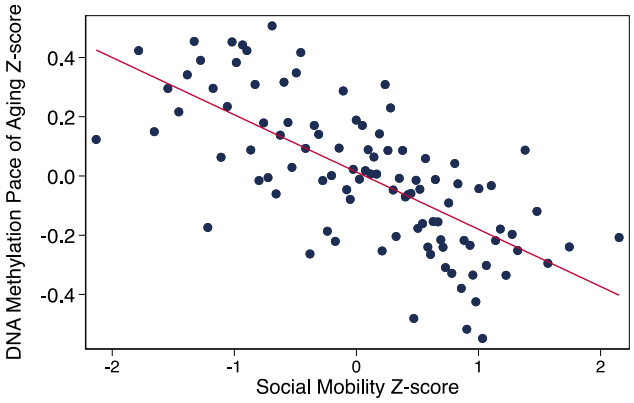
Power to Detect Effects on Health Outcome Measures. We used preliminary data from our meta-analysis (25), our ongoing *Paycheck Plus* study with MDRC (26, 186), and laboratory measures from the National Health and Nutrition Examination Survey (NHANES) to generate estimates of means (M) and standard deviations (SD) for the outcome measures in participants matched on age, sex, and race/ethnicity to our *MyGoals for Healthy Aging* participants. Based on these estimates, we computed changes in absolute units equivalent to our minimum detectable effect of d=0.16 for BMI, EuroQol 5D, HbA1C, and CRP. These are shown in **Table 6** (right). For the psychometric scales included in our outcome battery (the Perceived Stress Scale, the Patient Health Questionnaire PHQ-9, the Sleep Quality Index, the UCLA Loneliness scale, and the BRIEF-A), we estimate a change <10% of the mean value represents a minimum detectable effect.

Table 6. The proposed primary outcome measures for the 5 year survey and their associated Minimum Detectible Effect Size (MDES) analysis by Aim.* We can detect a 0.16 standard deviation (SD) change at each site.

	Change in Outcome Equivalent to d=0.16
Body Mass Index (BMI)†	1.4
EuroQol 5D (EQ5D)†	0.05
Hemoglobin A1c (HgA1c)†	0.15%
C-Reactive Protein (CRP)†	0.10 mg/L

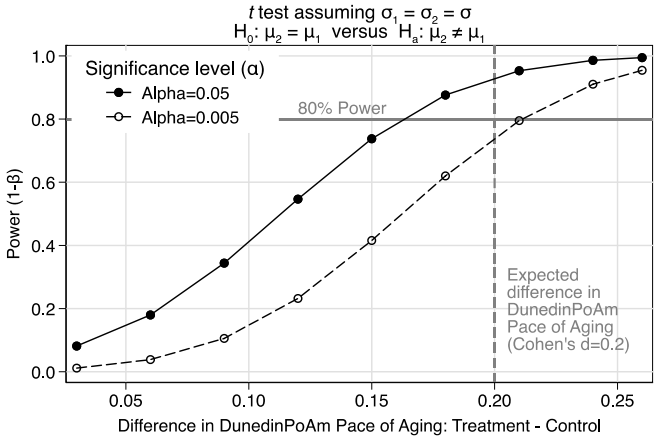
4.3.14. Statistical Power for Aim 3. We have >80% power at an alpha threshold of 0.05 to detect an effect-size of Cohen's d=0.2. That effect-size is plausible; in blood DNA methylation analysis of the DunedinPoAm measure in the E-Risk cohort (23) and US Health and Retirement Study (**Figure 1**) and in blood DNA

Figure 1. Association of upward socioeconomic mobility with slower DunedinPoAm Pace of Aging in the US Health and Retirement Study (effect-size r=0.2)



US Health and Retirement Study Data (n=3,931)
Plotted points show average X and Y coordinates for bins of ~40 participants. Regression line estimated from un-binned data.

Figure 2. Statistical power for treatment-effect analysis in the *MyGoals for Healthy Aging* RCT.



Parameters: N = 1200, N_1 = 600, N_2 = 600, μ_1 = 0, σ = 1

methylation analysis of DunedinPoAm in the Texas-Twin Project cohort (100), the effect-sizes for a one standard-deviation difference in household socioeconomic status were 0.2. At conventional levels of statistical significance (i.e., at the alpha=0.05 threshold), we are powered >80% to detect even smaller effects (**Figure 2**).

4.3.15. Statistical Power for Aim 4. All participants have consented for electronic data follow-up, but a small number of participants (typically 2% in other MDRC experiments) will have missing data, transposed data, or will otherwise generate a questionable match. We therefore assume 98% follow-up (n=1,762). A two-sided log rank test with an overall sample size of 1,762 subjects (1:1 allocation) achieves 80% power at a 0.05 significance level to detect a 5% difference in mortality hazard at six years of follow-up.

4.3.16. Clinical significance and health gains required for policy implementation.

Demonstrated clinically-significant health benefits are essential to establish a sufficient cost-benefit ratio for program implementation. MDRC's evaluation of *Work Rewards* found that the combination of FSS+incentives produced a 38% increase in earnings. This effect led to an overall positive net present value for participants, but not for the government budget. However, health impacts were not part of this calculus. Small to moderate effect sizes for outcomes such as depression or health-related quality of life (HRQL) can substantially increase the economic returns of an intervention.

As small changes in health in a broadly applied economic intervention may produce meaningful changes in population health, we assume a 5% threshold for clinical significance. Relative to the population mean, this corresponds to Cohen's d=0.15, or roughly equivalent to our minimum detectable effect size.

4.3.17. Limitations and contingency plans.

Limitations. Our decades of combined experience with logistical planning in social experiments, including the NIA-funded *Paycheck Plus* (5 R01 AG054466), greatly increases our chance of success. The intervention phase of *MyGoals* is nearing completion and thus far has been successful with respect to randomization and intervention uptake. Nevertheless, the study has some limitations.

First, may be difficult to determine which aspects of the program (income, employment, social support, or executive function) are most important for health. This is an unavoidable limitation of a multifaceted intervention and a near universal limitation of social policy experiments. For example, randomization to Medicaid versus no insurance impacts both income (by reducing out of pocket expenditures) and access to health care. Randomization to Section 8 housing vouchers can impact health via an array of mechanisms, not limited to exposure to crime, better schooling, more privileged social networks, and higher paying jobs. Our primary goal, therefore, is to evaluate, in its entirely, a policy package that is a candidate to replace *FSS*, the largest welfare program for Section 8 housing voucher recipients. It may nevertheless be possible to estimate the independent effect of some components of the intervention depending on whether there is variation in exposure to different components among the treatment group. Moreover, methylation algorithms under development may eventually reveal unique patterns specific to reductions in material hardship relative to improvements in social support. These algorithms can be applied to our data as they our developed.

Second, as with any social experiment, the participants themselves cannot be blinded to treatment status. It is not clear whether perceived exposure to a social intervention might alter objective outcomes such as blood sugar in the same way that a placebo pill might. However, it is not possible for control group participants to receive the treatment.

Third, voluntary enrollment is a threat to external validity. However, the characteristics of the parent population (i.e. the complete population of unemployed public housing residents) are known as they are available through the PHA database at each site. We therefore will examine effect sizes after adjusting for differences in these characteristics as a sensitivity analysis. Moreover, our inclusion criteria are relevant to the group being targeted by the *MyGoals* policy: unemployed persons of working age.

Finally, participants were enrolled on a rolling basis. This presents a threat because the job market can change over time. It also presents an opportunity because it becomes possible to examine the association between labor market characteristics and program uptake over time.

Contingency plans. Multiple forms of flexibility are built into the study.

First, we can control when we initiate data collection. Our expert panel does not expect intervention fade out on health impacts. Thus, if there are logistical issues associated with ramp-up, it is possible to simply delay data collection efforts. This buffer is built into the study (see **TIMELINE** attachment).

Second, our weekly team meetings will report successful completions and will shift financial resources as needed to cover study priorities. This dynamism, coupled with UI data, will allow us to make decisions in near real-time, so we can change course as needed. Like any RCT, we will be operating under budgetary constraints. If a larger than anticipated proportion of participants are lost to follow-up at the 6-year survey, extended contact tracing can quickly consume resources.

We have a wide array of measures of biological aging to choose from. This along with our detailed statistical power analyses will allow us to alter measure collection should we need to divert resources to address attrition or differential attrition. For our primary statistical power analyses, we assumed economic impacts are roughly similar to those observed in *Work Rewards* and a follow-up rate of 70%. If attrition is much larger than expected, we will still have UI, PHA, and vital status by cause of death for the whole sample. Our primary outcome measure, DunedinPoAm, is sensitive to both the effects of poverty and to interventions that slow aging, and prior RCTs of social policies suggest health outcomes are uniquely responsive to changes in income. *MyGoals for Healthy Aging* is powered to detect a Cohen's d effect size of d=0.23 (at an alpha of 0.05) even if we only capture 33% of the parent sample of 1,798 participants.

Third, we will deploy predictive analytics during year-2 data collection to identify participants least likely to be to successfully reached. These analytics, also used in the *Paycheck Plus* RCT, would likely be based upon factors such as the number of outreach attempts, the type of outreach (e.g., phone versus a home visit), and time from outreach to response. Applying predictive analytics to the cohort allows us to limit data collection efforts to those participants for whom outreach efforts are most cost-effective.

Fourth, to address attrition, we use intensive contact tracing for those who have left public housing and are no longer in HUD's MIS system. We also use electronic data systems to capture some economic and housing variables for all participants. This will help identify differences between the characteristics of the treatment and control group after accounting for attrition. Because administrative data are not subject to attrition, they provide a concrete reference point against which missing participants and missing data can be compared and corrected.

Alternative plans in case blood collection from *MyGoals for Healthy Aging* participants is unsuccessful. We have not previously collected blood samples from *MyGoals* participants. If significant numbers of participants decline blood collection, we will collect saliva sample for DNA extraction and DNA methylation analysis. We are confident this approach will be successful. In the *Paycheck Plus* Trial, which targets a sample with similar race/ethnic composition and similar economic disadvantages, MPI Muennig obtained self-collected saliva samples from >95% of those who were sent kits. Of these kits, 95% yielded DNA suitable for methylation analysis in Co-I Kobor's lab. DNA methylation analysis of these samples is ongoing. *In sum, collection of saliva DNA samples from MyGoals participants is feasible.*

However, as DunedinPoAm was developed from analysis of blood DNA methylation data, it is unclear if saliva DNA will yield similar results. DNA methylation states determine cell-type and therefore vary across tissues in the body. Saliva DNA comes from a mixture of leukocytes (the same cells from which blood DNA is derived) and buccal epithelial cells. We conducted analysis of DNA methylation data from matched saliva and blood samples from n=21 individuals (Gene Expression Omnibus accession GSE111165) to evaluate whether DunedinPoAm values were similar across tissues. We used well-established algorithms (187–189) to correct for the cell composition of saliva and blood samples. In the full cell-composition-adjusted dataset, the correlation between blood- and saliva-measured DunedinPoAm was r=0.6. However, when we restricted analysis to the n=14 paired samples run in the same assay batch, the correlation was r=0.8. This correlation approaches the test-retest ICC for blood DNAm in DunedinPoAm (190). (The test-retest reliabilities of DNA methylation measures of aging are 0.8-0.9, although our study will be able to take advantage of promising new methods that may increase these values, approaching 1.0 (190, 191)) Critically, using existing methods, we observed the same effect-size of r=0.2 for associations of household socioeconomic status with DunedinPoAm in a saliva-DNA methylation-based analysis of young people in the US (100), in blood-DNA methylation analysis of young people in the UK (23), and in older adults in the US (see **Figure 1**, **Statistical Power for Aim 3**). These data establish that DunedinPoAm can be measured from saliva DNA methylation.