Supplementary Appendix

Supplement to: Villareal DT, Aguirre L, Gurney AB, et al. Aerobic or resistance exercise, or both, in dieting obese older adults. N Engl J Med 2017;376:1943-55. DOI: 10.1056/NEJMoa1616338

Table of Contents

Investigators

1. Dennis T. Villareal, M.D.,[1,2]
2. Lina Aguirre, MD[3,4]
3. A. Burke Gurney, Ph.D., P.T.[5]
4. Debra L. Waters, Ph.D.[3,7]
5. David R. Sinacore, Ph.D. P.T.[8]
6. Elizabeth Colombo, M.D., Ph.D.[3.4]
7. Reina Armamento-Villareal, MD,[1,2]
8. Clifford Qualls, PhD[6]

[1]Division of Endocrinology, Diabetes, and Metabolism, Baylor College of Medicine, Houston, TX

[2]Center for Translational Research on Inflammatory Diseases, Michael E DeBakey VA Medical Center, Houston, TX

[3]Medicine Care Line, New Mexico VA Health Care System, Albuquerque, NM

[4]Department of Internal Medicine, [5]Division of Physical Therapy, [6]Department of Mathematics and Statistics, University of New México School of Medicine, Albuquerque, NM

[7] Department of Medicine, School of Physiotherapy, University of Otago, Dunedin, New Zealand

[8]The Program in Physical Therapy, Washington University School of Medicine, MO

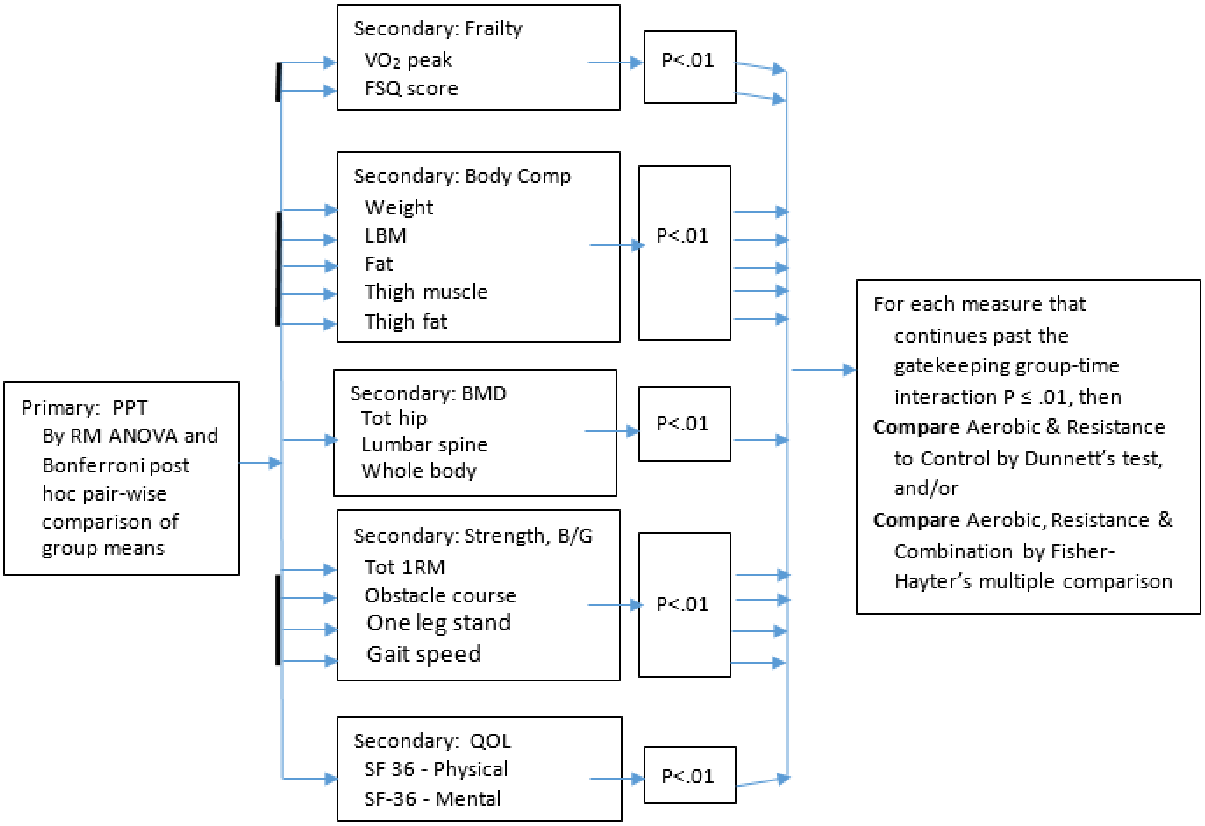
Statistical Analysis Rationale

A strategy can be used to analyze our data. We formalized this strategy by using a gatekeeping strategy.[1;2] We used only multiple comparison methods that maintained the family-wise error rate (FWER) ≤ 0.05. In the gatekeeping strategy, hypotheses were grouped in ordered families and sequential rules were set out that define rejection and continuation. The first family consisted of the one primary hypothesis (about PPT) and we required a significant interaction (α=.05) and used Bonferroni multiple comparison method for a strong post hoc pair-wise comparison among the four (4) study arms (Control, Aerobic, Resistance, and Combination) with α=.05. While our design included a Control group, the emphasis of our four specific aims in our protocol was the comparisons among the three (3) intervention arms; a limited purpose for the comparisons to Control was to show that at least one of the interventions (Aerobic, Resistance) was better than that of the Control. Thus we required at least one difference with Control and at least one difference among the interventions for the primary outcome to continue to the testing of the secondary outcomes (parallel gatekeeping).[2] We considered our secondary measures as pooled in the 5 domains (see labels in **Tables 2** and **S3**). A gatekeeping strategy required a p-value <.01 for the omnibus interaction for a domain measure in order to continue to the post hoc testing of this measure: namely, comparison to Control and among the intervention groups. (The omnibus test tests the "overall" significance of the model; it tests whether the explained variance in a set of data is significantly greater than the unexplained variance, overall). We used Fisher-Hayter's pair-wise comparison[3] method among the means of the intervention groups; it has more power than other strong methods (e.g., Holm [4]) and yet it maintains FWER $\alpha \leq .05$. We tested the differences to Control with the usual Dunnett's test[5] (also maintains FWER). Thus, our procedure could be described as tree gatekeeping based on the Fisher-Hayter test[2;6] Tree gatekeeping procedures can account for logical restrictions among multiple hypotheses as well as incorporate serial and parallel gatekeeping procedures for analysis of hierarchically ordered multiple hypotheses (see **Supplementary Figure S1.** Gatekeeping Strategy).

To explain further, Fisher-Hayter post hoc multiple comparison method is a modification of and is more liberal than Tukey's studentized range test (as called Tukey's honest significant difference, HSD) Fisher-Hayter method of pair-wise comparisons of group means is more conservative than Fisher's least significant difference (LSD) method. Fisher-Hayter maintains FWER≤0.05, and is more powerful than Tukey's HSD[7-9] and Holm.[4] The Hayter's modification requires a significant omnibus test (step 1) as protection of the subsequent pair-wise comparisons (step 2). If we proceeded past the gatekeeping to the analysis of the 5 domains of secondary measures, we considered the secondary measures to be parallel hypotheses since we did not necessarily expect them all to show difference among the 3 interventions. For each measure we required a significant omnibus test, and for the final analysis of the measure, used Dunnett's test for comparison to Control l (k=2, α=0.05) and Fisher-Hayter method of pair-wise comparison for the 3 interventions (k=3, α=0.05).

It should be pointed out in **Table 2** and **S3** that the 3 columns of p-values for comparisons among the interventions are the same as though we used Fisher's LSD. This is due to the fact that for k=3 means, Fisher-Hayter is the same as Fisher's LSD. This is proved in Theorem 1 of Hayter[13], and one can verify this by comparing the table of critical values for the two methods. So both Fisher-Hayter and Fisher's LSD method with k=3 means maintained FWER ≤0.05. However, the p-values in Table 2 marked with a ‖ were not considered significant per the Fisher-Hayter method because the step 1 omnibus test (interaction) among the 3 means was not significant.

Sensitivity analyses that validated the statistical approach taken included multiple imputation for missing fitness data (which confirmed a similar pattern of results). Analyses also included logistic regression to determine whether data were consistent with an assumption that data were missing at random (data were consistent with the assumption that data were missing at random).


Secondary: Frailty
VO₂ peak
FSQ score
P<.01
Secondary: Body Comp
Weight
LBM
Fat
Thigh muscle
Thigh fat
P<.01
Primary: PPT
By RM ANOVA and Bonferroni post hoc pair-wise comparison of group means
Secondary: BMD
Tot hip
Lumbar spine
Whole body
P<.01
Secondary: Strength, B/G
Tot 1RM
Obstacle course
One leg stand
Gait speed
P<.01
Secondary: QOL
SF 36 - Physical
SF-36 - Mental
P<.01
For each measure that continues past the gatekeeping group-time interaction P ≤ .01, then
Compare Aerobic & Resistance to Control by Dunnett's test, and/or
Compare Aerobic, Resistance & Combination by Fisher-Hayter's multiple comparison

Supplementary Figure S1. Gatekeeping Strategy

The three columns to the left represent gatekeeping functions for the secondary measures: the first column is for the PPT and is tested by the group- time interaction in PROC MIXED in SAS; if this primary interaction is significant then proceed to the second column consisting of five secondary domains containing 16 individual measures which are tested with the Bonferroni correction for five multiple comparisons; the domain is significant if any one of its individual measures has a significant group-time interaction (parallel hypotheses); if any of the 16 measures is significant then proceed for that row to the fourth column for two select hypotheses against the control group, tested with the Dunnett's correction for multiple comparisons; if either of the hypotheses about the Aerobic or Resistance with Control by Dunnett's tests is significant then proceed for that individual measure to the three pair-wise comparisons among the Aerobic, Resistance and the Combination groups tested for the group- time interaction by Fisher-Hayter's method.

Abbreviations: PPT = Physical Performance Test, VO_{2peak} = Peak Oxygen Consumption, FSQ = Functional Status Questionnaire, Body Comp = Body Composition, BMD = Bone Mineral Density, B/G = Balance/Gait, QOL = Quality of Life. Groups: Control, Aerobic = Weight management and aerobic training, Resistance = Weight management and resistance training, Combination = Weight management and combined aerobic and resistance training.

Supplemental Table S1. Summary of Serious Adverse Events, Exercise Related Adverse Events and Procedure Related Adverse Events

	Control (n=40)	**Aerobic (n=40)**	**Resistance (n=40)**	**Combination (n=40)**
Serious Adverse Events				
	Appendicitis	Wrist fracture	Congestive heart failure	Non-ST-elevation myocardial infarction
	Incarcerated umbilical hernia	Endometrial adenocarcinoma	Cholecystitis	Hypotension
		Hospitalization for flu symptoms	Liver cyst	
			Colon cancer	
Exercise Related				
		Fell during exercise	Atrial fibrillation	Shoulder injury
		Left shoulder pain	Left shoulder pain	Left knee pain (n=2)
		Back pain with sciatica	Right knee pain from torn meniscus	Spinal stenosis exacerbation
				Hip pain
Procedure Related				
		Hand & forehead abrasion from fall	Fall after physical performance testing	
			Tripped and fell after blood draw	

Groups: Control, Aerobic = Weight management and aerobic training, Resistance = Weight management and resistance training, Combination = Weight management and combined aerobic and resistance training

Supplemental Table S2. Adverse Events by Systems using MedDRA

	Control (n=40)	**Aerobic (n=40)**	**Resistance (n=40)**	**Combination (n=40)**	P-value
System Organ Class, no. (%)					
Cardiac disorders	0 (0)	0 (0)	2 (5.0)	1 (2.5)	.62
Ear and labyrinthine disorders	0 (0)	0 (0)	0 (0)	1 (2.5)	1.0
Eye disorders	0 (0)	0 (0)	0 (0)	1 (2.5)	1.0
Gastrointestinal disorders	1 (2.5)	1 (2.5)	1 (2.5)	1 (2.5)	1.0
General disorders	1 (2.5)	0 (0)	2 (5.0)	2 (5.0)	.76
Hepatobiliary disorders	0 (0)	0 (0)	1 (2.5)	1 (2.5)	1.0
Infections and Infestations	2 (5.0)	1 (2.5)	1 (2.5)	2 (5.0)	1.0
Injury, poisoning, and procedural complications	1 (2.5)	5 (12.5)	7 (17.5)	3 (7.5)	.13
Metabolism and Nutritional disorders	0 (0)	0 (0)	0 (0)	1 (2.5)	1.0
Musculoskeletal and connective tissues disorders	2 (5.0)	8 (20)	4 (10)	5 (12.5)	.24
Neoplasms, benign, malignant, and unspecified	0 (0)	1 (2.5)	3 (7.5)	1 (2.5)	.40
Nervous system disorders	0 (0)	0 (0)	1 (2.5)	1 (2.5)	1.0
Psychiatric disorders	0 (0)	0 (0)	1 (2.5)	0 (0)	1.0
Respiratory, thoracic, and mediastinal disorders	1 (2.5)	1 (2.5)	1 (2.5)	0 (0)	1.0
Skin and subcutaneous tissue disorders	1 (2.5)	1 (2.5)	1 (2.5)	1 (2.5)	1.0
Surgical and medical procedures	0 (0)	1 (2.5)	0 (0)	0 (0)	1.0
Vascular disorders	0 (0)	0 (0)	0 (0)	1 (2.5)	1.0

Groups: Control, Aerobic = Weight management and aerobic training, Resistance = Weight management and resistance training, Combination = Weight management and combined aerobic and resistance training

Supplemental Table S3. Effect of Specific Exercise Modes, Added to Diet-induced Weight loss, on Outcomes*

Outcome Variables	Control (n=40)	Aerobic (n=40)	Resistance (n=40)	Combination (n=40)	P Value†					
					Group-Time Interaction	Aerobic vs. Control	Resistance vs. Control	Aerobic vs. Resistance	Combination vs. Aerobic	Combination vs Resistance
Primary outcome										
PPT score										
Baseline	28.6±0.5	29.3±0.3	28.8±0.4	27.9±0.4						
Change at 3 months	0.4±0.3	2.9±0.4‡	3.2±0.4‡	4.0±0.4‡						
Change at 6 months	1.0±0.4	3.9±0.4‡	3.9±0.4‡	5.5±0.4‡	<.001	<.001	<.001	.87	.002	.004
Secondary outcomes										
Other frailty measures										
VO_{2peak} (ml/kg/min)										
Baseline	17.0±0.5	17.6±0.5	17.0±0.6	17.2±0.6						
Change at 6 months	0.1±0.3	3.3±0.3‡	1.3±0.3‡	3.1±0.3‡	<.001	<.001	.007	<.001	.63	.001
FSQ score										
Baseline	29.8±0.5	30.1±0.5	29.3±0.6	29.8±0.6						
Change at 6 months	0.4±0.3	2.0±0.3‡	2.3±0.3‡	3.6±0.3‡	<.001	.002	<.001	.46	.005	.03
Body composition										
Body weight (kg)										
Baseline	97.9±2.9	96.9±2.3	101.8±2.9	99.0±2.9						
Change at 6 months	-0.9±0.5	-9.0±0.6‡	-8.5±0.5‡	-8.5±0.5‡	<.001	<.001	<.001	.76§	.72§	.96§
Lean mass (kg)										
Baseline	54.9±2.3	55.0±1.9	58.1±2.3	56.5±1.8						
Change at 6 months	0.0±0.2	-2.7±0.3‡	-1.0±0.3¶	-1.7±0.3‡	<.001	<.001	.03	.001	.047	.20
Fat mass (kg)										
Baseline	43.0±1.5	41.9±1.3	44.3±1.5	42.5±1.6						
Change at 6 months	-0.9±0.4	-6.3±0.5‡	-7.3±0.4‡	-7.0±0.5‡	<.001	<.001	<.001	.18§	.36§	.67§
Thigh muscle (cm^3)										
Baseline	1302±63	1234±62	1190±48	1186±66						
Change at 6 months	10±7	-77±7‡	-23±7¶	-40±7‡	<.001	<.001	.008	<.001	.005	.21
Thigh fat (cm^3)										
Baseline	1774±132	1700±98	1848±108	1784±125						
Change at 6 months	-2±36	-260±35‡	-280±35‡	-288±35‡	<.001	<.001	<.001	.64§	.61§	.97§
BMD at total hip										
Total hip (gm/cm^2)										
Baseline	1.031±.025	1.018±.019	1.047±.022	1.010±.025						
Change at 6 months	.004±.004	-.027±.004‡	- .006±.004	-.014±.004¶	.001	<.001	.37	.005	.04	.43
Lumbar spine (gm/cm^2)										
Baseline	1.141±.033	1.118±.022	1.144±.033	1.157±.033						

Change at 6 months	.010±.006	.002±.006	.008±.006	.008±.005						
Whole body (gm/cm^2)										
Baseline	1.134±.023	1.118±.022	1.115±.021	1.120±.023						
Change at 6 months	-.001±.005	-.003±.005	.005±.005	.002±.005						
Strength, balance, gait										
Total 1RM (kg) ‖										
Baseline	269±21	265±20	288±30	272±16						
Change at 6 months	2±5	5.0±5	49±5‡	48±5‡	<.001	.99	<.001	<.001	<.001	.82
Obstacle course (sec)										
Baseline	15.9±0.7	15.5±0.8	16.4±0.6	17.0±1.0						
Change at 6 months	0.0±0.3	-1.5±0.4‡	-2.2±0.3‡	-2.9±0.3‡	<.001	.02	<.001	.20	.01	.20
One leg stance (sec)										
Baseline	7.7±0.7	6.7±0.8	6.0±0.6	7.9±1.0						
Change at 6 months	-0.8±0.8	2.5±0.9¶	3.6±0.8¶	5.9±0.8‡	.001	.07	.002	.40§	.02§	.13§
Gait speed (m/min)										
Baseline	75.0±2.6	74.6±2.0	74.3±2.3	68.8±2.2						
Change at 6 months	-0.5±1.3	8.1±1.3‡	9.3±1.3‡	12.1±1.3‡	<.001	<.001	<.001	.56	.03	.09
Quality of Life										
SF-36, physical score										
Baseline	47.0±1.7	48.6±1.4	51.0±1.5	45.9±1.6						
Change at 6 months	-1.6±0.8	6.5±0.7‡	7.4±0.8‡	9.5±0.7‡	<.001	<.001	<.001	.74	.02	.049
SF-36, mental score										
Baseline	42.7±0.7	43.4±0. 9	42.7±0.9	45.1±0.9						
Change at 6 months	-0.7±0.6	1.9±0.5	1.9±0.6	2.6±0.5						

* Plus–minus values for the change scores are the least-squares adjusted means ±SE from the repeated-measures analyses of variance; plus–minus values for the baseline values are the observed means ±SE. Scores on the Physical Performance Test (PPT) (primary outcome) range from 0 to 36, with higher scores indicating better physical function; the minimal clinically important difference is 1.8. Peak oxygen consumption (VO_{2peak}) was assessed during graded treadmill walking. Scores on the Functional Status Questionnaire (FSQ) range from 0 to 36, with higher scores indicating better function. BMD denotes bone mineral density.

† P values for the changes from baseline to 6 months in between-group comparisons were calculated with the use of mixed-model repeated-measures analyses of variance (with baseline values and sex as covariates) and are reported when the overall P value was lower than 0.05 for the interaction among the four groups over time. In a Bonferroni correction to adjust for the multiple comparisons in the PPT score (in which the P values were multiplied by 5 for the comparison with an alpha level of 0.05), the corrected P values were 0.01 for the combination group versus the aerobic group and 0.02 for the combination group versus the resistance group. In accordance with a gatekeeping strategy, a significant group-by-time interaction (P<0.01) and at least one significant difference between an exercise group and the control group and at least one significant difference among the exercise groups in the change in PPT score were required to continue to testing of the secondary outcomes; comparisons of the exercise groups with the control group were performed with Dunnett's test and comparisons among the intervention groups were performed with the Fisher-Hayter test. Secondary analyses included a comparison between the combination group and the control group; all P values were less than 0.05.

‡ P<0.001 for the comparison of the value at the follow-up time with the baseline value within the group, as calculated with the use of mixed-model repeated-measures analysis of variance.

§ The between-group differences were not significant in the last three columns because the group-by-time interaction for the three intervention groups (step 1 of the Fisher-Hayter test) was not significant; the family-wise error rate of 0.05 or less is maintained for the Fisher–Hayter multiple comparison procedure.

¶ P<0.01 for the comparison of the value at the follow-up time with the baseline value within the group, as calculated with the use of mixed-model repeated-measures analysis of variance.

‖ Total one-repetition maximum (1RM) is the total of the maximum weight a participant can lift, in one attempt, in the biceps curl, bench press, seated row, knee extension, knee flexion, and leg press.

References

(1) Dmitrienko A, Tamhane AC. Mixtures of multiple testing procedures for gatekeeping applications in clinical trials. *Stat Med* 2011;30:1473-1488.

(2) Dmitrienko A, Tamhane AC, Liu L, Wiens BL. A note on tree gatekeeping procedures in clinical trials. *Stat Med* 2008;27:3446-3451.

(3) Hayter AJ. The maximum familywise error rate of Fisher's least significant difference test. Biometrics 12, 1000-1004. 1986.

(4) Holm S. A simple sequentially rejective multiple test procedure. Scand.J.Statist. 6, 65-70. 1979.

(5) Dunnett CW. Pairwise multiple comparisons in the homogeneous vriance, unequal sample size case. Journal of the American Statistical Association 75, 789-795. 1980.

(6) Brechenmacher T, Xu J, Dmitrienko A, Tamhane AC. A mixture gatekeeping procedure based on the Hommel test for clinical trial applications. *J Biopharm Stat* 2011;21:748-767.

(7) Tukey J.W. Comparing individual means in the analysis of variance. Biometrics 5, 99-114. 2016.

(8) Tukey JW. The problem of multiple comparisons. Unpublished manuscript, Department of Statistics, Princeton University, 1953.

(9) Kramer CY. Extension of multiple range tests to group means with unequal numbers of replications. Biometrics 12, 309-310. 1956.