http://dx.doi.org/10.5665/sleep.2226

Effects of Continuous Positive Airway Pressure on Neurocognitive Function in Obstructive Sleep Apnea Patients: The Apnea Positive Pressure Long-term Efficacy Study (APPLES)

Clete A. Kushida, MD, PhD[1]; Deborah A. Nichols, MS[1]; Tyson H. Holmes, PhD[1]; Stuart F. Quan, MD[2,5]; James K. Walsh, PhD[3]; Daniel J. Gottlieb, MD, MPH[4,5]; Richard D. Simon Jr., MD[6]; Christian Guilleminault, MD[1]; David P. White, MD[5]; James L. Goodwin, PhD[2]; Paula K. Schweitzer, PhD[3]; Eileen B. Leary, RPSGT[1]; Pamela R. Hyde, MA[1]; Max Hirshkowitz, PhD[7]; Sylvan Green, MD[2]; Linda K. McEvoy, PhD[8]; Cynthia Chan, BS[9]; Alan Gevins, DSc[9]; Gary G. Kay, PhD[10]; Daniel A. Bloch, PhD[1]; Tami Crabtree, MS[11]; William C. Dement, MD, PhD[1]

[1]Stanford University, Stanford, CA; [2]University of Arizona, Tucson, AZ; [3]St. Luke's Hospital, Chesterfield, MO; [4]VA Boston Healthcare System, Boston, MA; [5]Brigham and Women's Hospital, Boston, MA; [6]Providence St. Mary Medical Center, Walla Walla, WA; [7]VAMC Sleep Center, Houston, TX; [8]University of California, San Diego, CA; [9]SAM Technology Inc. & The San Francisco Brain Research Institute, San Francisco, CA; [10]Georgetown University School of Medicine, Washington, DC; [11]Santa Rosa, CA

Study Objective: To determine the neurocognitive effects of continuous positive airway pressure (CPAP) therapy on patients with obstructive sleep apnea (OSA).

Design, Setting, and Participants: The Apnea Positive Pressure Long-term Efficacy Study (APPLES) was a 6-month, randomized, double-blind, 2-arm, sham-controlled, multicenter trial conducted at 5 U.S. university, hospital, or private practices. Of 1,516 participants enrolled, 1,105 were randomized, and 1,098 participants diagnosed with OSA contributed to the analysis of the primary outcome measures.

Intervention: Active or sham CPAP

Measurements: Three neurocognitive variables, each representing a neurocognitive domain: Pathfinder Number Test-Total Time (attention and psychomotor function [A/P]), Buschke Selective Reminding Test-Sum Recall (learning and memory [L/M]), and Sustained Working Memory Test-Overall Mid-Day Score (executive and frontal-lobe function [E/F])

Results: The primary neurocognitive analyses showed a difference between groups for only the E/F variable at the 2 month CPAP visit, but no difference at the 6 month CPAP visit or for the A/P or L/M variables at either the 2 or 6 month visits. When stratified by measures of OSA severity (AHI or oxygen saturation parameters), the primary E/F variable and one secondary E/F neurocognitive variable revealed transient differences between study arms for those with the most severe OSA. Participants in the active CPAP group had a significantly greater ability to remain awake whether measured subjectively by the Epworth Sleepiness Scale or objectively by the maintenance of wakefulness test.

Conclusions: CPAP treatment improved both subjectively and objectively measured sleepiness, especially in individuals with severe OSA (AHI > 30). CPAP use resulted in mild, transient improvement in the most sensitive measures of executive and frontal-lobe function for those with severe disease, which suggests the existence of a complex OSA-neurocognitive relationship.

Clinical Trial Information: Registered at clinicaltrials.gov. Identifier: NCT00051363.

Keywords: Obstructive sleep apnea, continuous positive airway pressure, neurocognitive function, randomized controlled trial, sleepiness

Citation: Kushida CA; Nichols DA; Holmes TH; Quan SF; Walsh JK; Gottlieb DJ; Simon RD; Guilleminault C; White DP; Goodwin JL; Schweitzer PK; Leary EB; Hyde PR; Hirshkowitz M; Green S; McEvoy LK; Chan C; Gevins A; Kay GG; Bloch DA; Crabtree T; Demen WC. Effects of continuous positive airway pressure on neurocognitive function in obstructive sleep apnea patients: the Apnea Positive Pressure Long-term Efficacy Study (APPLES). *SLEEP* 2012;35(12):1593-1602.

INTRODUCTION

Obstructive sleep apnea (OSA) is a common sleep-related breathing disorder estimated to affect more than 14 million Americans[1]; comprehensive data are lacking on the impact of OSA on the neurocognitive domains of attention and psychomotor function, learning and memory, and executive and frontal-lobe function. Continuous positive airway pressure (CPAP) therapy is in widespread use,[2] yet its efficacy in providing significant long-term neurocognitive and other functional benefits to OSA patients has not been systematically investigated. The National Heart, Lung, and Blood Institute (NHLBI)-supported Apnea Positive Pressure Long-term Efficacy Study (APPLES) is a randomized, double-blind, 2-arm, sham-controlled, multicenter, long-term (6 months) trial of CPAP therapy, designed to provide adequate statistical power to assess its efficacy on neurocognitive function in patients with OSA across a range of disease severity.

A commentary on this article appears in this issue on page 1585.

Submitted for publication December, 2011
Submitted in final revised form May, 2012
Accepted for publication June, 2012

Address correspondence to: Clete A. Kushida, MD, PhD, RST, RPSGT, FAASM, Stanford Sleep Medicine Center, 450 Broadway Street, MC 5704, Pavillion C, 2nd Floor, Redwood City, CA 94063; Tel: (650) 721-7560; Fax: (650) 721-3465; E-mail: clete@stanford.edu

METHODS

Participants

APPLES was conducted at 5 Clinical Centers: Stanford University, Stanford, CA; University of Arizona, Tucson, AZ; Providence St. Mary Medical Center, Walla Walla, WA; St. Luke's Hospital, Chesterfield, MO; and Brigham and Women's Hospital, Boston, MA. The protocol[3] was approved by the institutional review board (IRB) at each site; the first participant was enrolled in 11/2003 and the final completion month was 8/2008.

The inclusion criteria[3] were a diagnosis of OSA[4] with an apnea-hypopnea index (AHI) ≥ 10 and age ≥ 18 years. The

primary exclusion criteria[3] were: (1) prior OSA treatment with CPAP or surgery; (2) anyone in the household with current/past CPAP use; (3) sleepiness-related automobile accident within past year; (4) oxygen saturation < 75% for > 10% of the diagnostic polysomnogram (PSG) total sleep time; and/or (5) conditions (including known neurocognitive impairment), disorders, medications, or substances that could potentially affect neurocognitive function and/or alertness.

Study Design

Sample size[3] was calculated to permit detection of treatment effects at least as large as those estimated from two pilot studies, with 90% power and a type I error rate of 5%. In the pilot studies, the Pathfinder Number Test had the smallest estimated effect size of 0.2, which translates to a difference of 26 msec in reaction time between the Active and Sham CPAP groups. Allowing for 3 interim analyses and 20% dropout,[5,6] this effect size provided a randomization target of 1,100 participants (Appendix Section 1A).

The Data Coordinating Center (DCC) used a computerized permuted block design[3] to randomize 1,105 participants to active vs. sham CPAP (REMstar Pro, Philips Respironics, Inc.) devices; the sham CPAP device closely simulates the airflow through the exhalation port and the operating noise of the active CPAP device.[7] Randomization was stratified by gender, race (white vs. non-white), and OSA severity (mild, 10.0-15.0 respiratory events per hour of sleep; moderate, 15.1-30.0; severe, > 30; using American Academy of Sleep Medicine Task Force [1999] OSA diagnostic criteria).[4] A biased coin (7:3) was implemented for blocks of 30 when the difference in percentage randomized to active vs. sham at a given site was > 7%. Participants and most personnel were blinded[3] to treatment assignments, with the exception of site coordinators, PSG technologists, and the database administrator/data manager.

Participants were studied up to 6 months over 11 visits (Figure 1) and were compensated up to $500 for study completion. All data from sites were linked to a unique subject code and were securely transferred and archived by the DCC using a custom-designed Internet-based data management system that facilitated extensive quality control procedures.[3]

CPAP adherence[3] was objectively assessed using Encore Pro SmartCard (Philips Respironics, Inc.) data. Site staff contacted participants twice within the first week after starting CPAP to ensure use and manage any problems, and regularly thereafter to discuss CPAP nonadherence (< 4 h of use/night).

Efficacy and Safety Evaluations

The primary outcomes[3] were 3 neurocognitive variables, each representing a neurocognitive domain: (1) Pathfinder Number Test-Total Time (PFN-TOTL) assesses attention and psychomotor function (A/P), and comprises the total time for the participant to scan, locate, and connect numbers in sequence (computer analog of Trail Making Test Part A); (2) Buschke Selective Reminding Test-Sum Recall (BSRT-SR)[8] assesses verbal learning and memory (L/M), and consists of the total words recalled across 6 selective reminding trials; and (3) Sustained Working Memory Test-Overall Mid-Day Index (SWMT-OMD)[9] assesses an executive and frontal-lobe function (E/F) component by requiring the participant to compare the spatial position of a stimulus with its position on a previous trial (n-back test), pressing one button if the spatial position was the same as that on the previous trial or a second button if it differed. For SWMT-OMD, a behavioral (task performance) and 2 electroencephalographic (task-related EEG [cortical activation] and resting EEG [alertness]) subindices are combined to yield an overall index indicating the degree of change from pre-treatment baseline for the midday test administration.[9] The secondary outcomes[3] were 7 neurocognitive and 2 sleepiness measures, the maintenance of wakefulness test (an objective test to assess participants' ability to remain awake) and the Epworth Sleepiness Scale (a questionnaire to assess subjective daytime sleepiness).

Each site had a blinded physician observer who assessed participant safety[3] throughout the study. The DCC monitored and reported safety data to the IRBs and Data and Safety Monitoring Board (DSMB). Stopping rules[3] were developed for early efficacy[10] in addition to safety (cardiovascular disease [CVD] and motor vehicle accidents [MVAs]); data were presented by blinded arm to the DSMB at each interim analysis (25%, 50%, and 75%).

Statistical Analyses

The protocol-specified primary comparison was the difference between slopes (active vs. sham) across time, but generalized estimating equations (GEE)[11] could only be applied to one of the 3 primary outcomes (PFN-TOTL), due to: (1) an inadvertent difference in difficulty of the BSRT-SR form versions between baseline and subsequent administrations and (2) the SWMT-OMD provided as a change from baseline score (Appendix Section 1B). Therefore, after review of the GEE results, it was decided that generalized linear models (GLM) for by-visit comparisons, generalized linear mixed models (GLMM) for repeated measures data, or parametric survival analyses for right-censored data be used to fit the primary outcomes for comparing means between study arms (Appendix Section 2B). Analyses for all 3 main outcomes were done with and without adjustment for baseline covariates. Post hoc CPAP adherence-adjusted and retention-adjusted primary outcome analyses are described in Appendix Sections 7-8. Post hoc primary outcome analyses were also performed restricted to CPAP-adherent individuals using the same methods described above (Appendix Section 7). Post hoc oxygen saturation analyses used GLM; sleepiness analyses used 2-sample t-tests or Spearman correlation coefficients (Appendix Section 2E-2G).

Comparison of AHI means between study arms by visits used 2-sample t-tests after Box-Cox transformation. CPAP adherence was analyzed as an outcome using a Kolmogorov-Smirnov 2-sample test,[12] χ^2 test, or permutation test (Appendix Sections 5A-5C). Agreement between blinded participant guesses and actual treatment assignment was estimated by a κ coefficient (Appendix Section 5D). Associations between sleepiness and CPAP adherence used Spearman correlation coefficients (Appendix Section 4A). Retention was analyzed as an outcome using a life-table method (Appendix Section 6A).

Following an a priori analysis plan, 7 secondary outcome neurocognitive variables were selected from an initial set of 12 via independent component analysis (ICA).[13] GLM or GLMM was used to regress each secondary outcome on study arm with adjustment for covariates (Appendix Section 3). Maintenance of wakefulness test analyses used a chop-lump test[14] due to a

high frequency of scores at the 20-min ceiling. Regression analyses for the Epworth Sleepiness Scale used GLM for an overdispersed binomial distribution. Safety analyses used GLM.

The DCC conducted all analyses (using SAS[15] and R[16]). Hypothesis testing was 2-tailed at a type I error rate of 3.07% for the primary neurocognitive analyses (due to interim tests) and a 5% type I error rate for the remaining analyses. Intention-to-treat parameters, verification of model assumptions, and treatment of missing data are described in the Appendix Sections 1C-1E.

RESULTS

Baseline

Of 1,516 participants enrolled, 1,105 were randomized. Three participants had an AHI < 10 (following PSG quality control), and 4 had inadvertent exposure to both treatment conditions. They were excluded from analyses, resulting in 1,098 randomized participants (556 active, 542 sham; Figure 1). Baseline participant characteristics revealed an obese, predominantly white, male, highly educated sample, and the sleep study data are consistent with those of untreated OSA patients; further characteristics are discussed in a separate publication on the baseline analyses conducted for this study.[17] Baseline data were similar between arms (Table 1); the only difference detected was that active participants were 1.4 years older on average.

Efficacy

Primary Neurocognitive Outcomes

For protocol-specified GEE analyses, no difference in slopes over time was detected for PFN-TOTL between arms (P = 0.8663) (Appendix Section 2A). Comparison of means (regression estimates) between arms revealed a difference for SWMT-OMD at the 2 month (2M) CPAP visit (active 0.035, sham -0.074, P = 0.0074; Table 2). No differences in means were detected between arms for SWMT-OMD at the 6 month (6M) CPAP visit, or for PFN-TOTL and BSRT-SR at either visit.

Effects of CPAP Adherence and Retention on Primary Outcome Analyses

CPAP adherence data (Appendix Section 5) for the participants' entire follow-up duration revealed a difference in mean nightly CPAP usage between arms (active 4.2, sham 3.4 h, P < 0.001). Adherence was also analyzed for various durations (night, week, month, and 2 months) prior to the 2M and 6M visits; differences in means were detected between arms for all durations at both visits (e.g., week prior to 2M and 6M: active 5.1, sham 4.1 h, P < 0.0001). Active participants adhered more by a standard criterion (≥ 4 h for > 70% of the nights) for all durations prior to both visits. A total of 55.3% of active participants correctly guessed their treatment assignment vs. 69.7% of sham participants ($\kappa = 0.25$, P < 0.0001). Participant retention at 6M differed between arms (active 79.7%, sham 74.4%, log-rank P = 0.0363; Appendix Section 6). Based on these findings, primary outcomes were adjusted for adherence and retention.

When primary neurocognitive analyses were restricted to CPAP-adherent individuals (mean nightly active or sham CPAP adherence ≥ 4 h for the 2 months prior to each neurocognitive testing visit), no differences in means were detected between arms for any of the primary outcomes at any visit (2M SWMT-OMD, estimated active mean minus sham mean = 0.088, P = 0.0892; Appendix Section 7). Restriction to the adherent population resulted in a smaller sample size (2M n = 511, 6M n = 413) and an imbalance for one baseline feature (mean IQ Verbal WASI was 2.5 units higher for sham than active at 6M, P = 0.0453) that was not present in the full population; however, the imbalance on baseline age that existed in the full population (Table 1) was not detectable in this subgroup (P ≥ 0.1366).

An analysis comparing baseline variables for the group of adherent individuals vs. non- adherent individuals at both the 2M and 6M time points revealed significant differences in a number of baseline variables. Adherent individuals were older on average (2M 4.8 y older, P < 0.0001; 6M 5.4 y older, P < 0.0001), were more likely to be white (2M/6M P < 0.0001) and married (2M P = 0.0474, 6M P = 0.0161), and also had higher WASI IQ scores on average (e.g., IQFull4WASI: 2M 5.1 points higher, 6M 4.5 points higher, P < 0.0001). Some differences in baseline polysomnographic variables also emerged. On average, the group of CPAP-adherent individuals at 2M and 6M had a lower sleep efficiency percentage at baseline (2M 1.9% lower, P = 0.0296; 6M 3.8% lower, P < 0.0001); and at 6M, adherers had a shorter total sleep time (15 min lower, P = 0.0011), longer sleep latency (4.2 min higher, P = 0.0063), longer REM latency (5.4 min higher, P = 0.0221), and a lower percentage of stage 3 sleep (0.67% lower, P = 0.0424).

We also performed analyses that adjusted for the confounding that could arise because participants selected their levels of adherence. Results for the primary outcomes remained unchanged when compared at each of 9 different levels of mean adherence (0, 1, 2, …, 8 hours per night), with adjustment for possible confounding using generalized propensity scores. These adjusted analyses detected a difference in means between arms for SWMT-OMD at 2M for 3 and 4 h of mean adherence per night (P ≤ 0.044, Appendix Section 7).

Retention-adjusted primary outcome analyses (Appendix Section 8) revealed the tendency to discontinue (drop or disqualification) from the study was associated with neurocognitive change from baseline for the 2M/6M BSRT-SR and for the 6M SWMT-OMD (P ≤ 0.0075); however, adjusting for these associations did not alter detection of treatment effects.

Effects of AHI, Oxygen Saturation, and Sleepiness on Primary Outcome Analyses

A significant difference was detected in AHI between active vs. sham CPAP groups at 2M (P < 0.0001) and 6M (P < 0.0001); no difference in AHI was detected between groups at baseline. Covariate-adjusted regression analyses detected a difference between arms in the 2M SWMT-OMD for only those participants with severe OSA at baseline (P = 0.0031) (Table 2). Additional analyses revealed that the only significant change in means between the 2M and 6M visits for the SWMT-OMD was for participants with severe OSA in the sham group (sham -0.150, P = 0.0132; Appendix Section 2D).

To assess whether baseline oxygen saturation may be correlated with the neurocognitive response to CPAP,[18] post hoc mean comparisons were made between the lower three $\%TSTO_2 < 85$ quartiles vs. the upper quartile separately by visit and arm. For the SWMT-OMD, those in the upper quartile (lower oxygen satura-

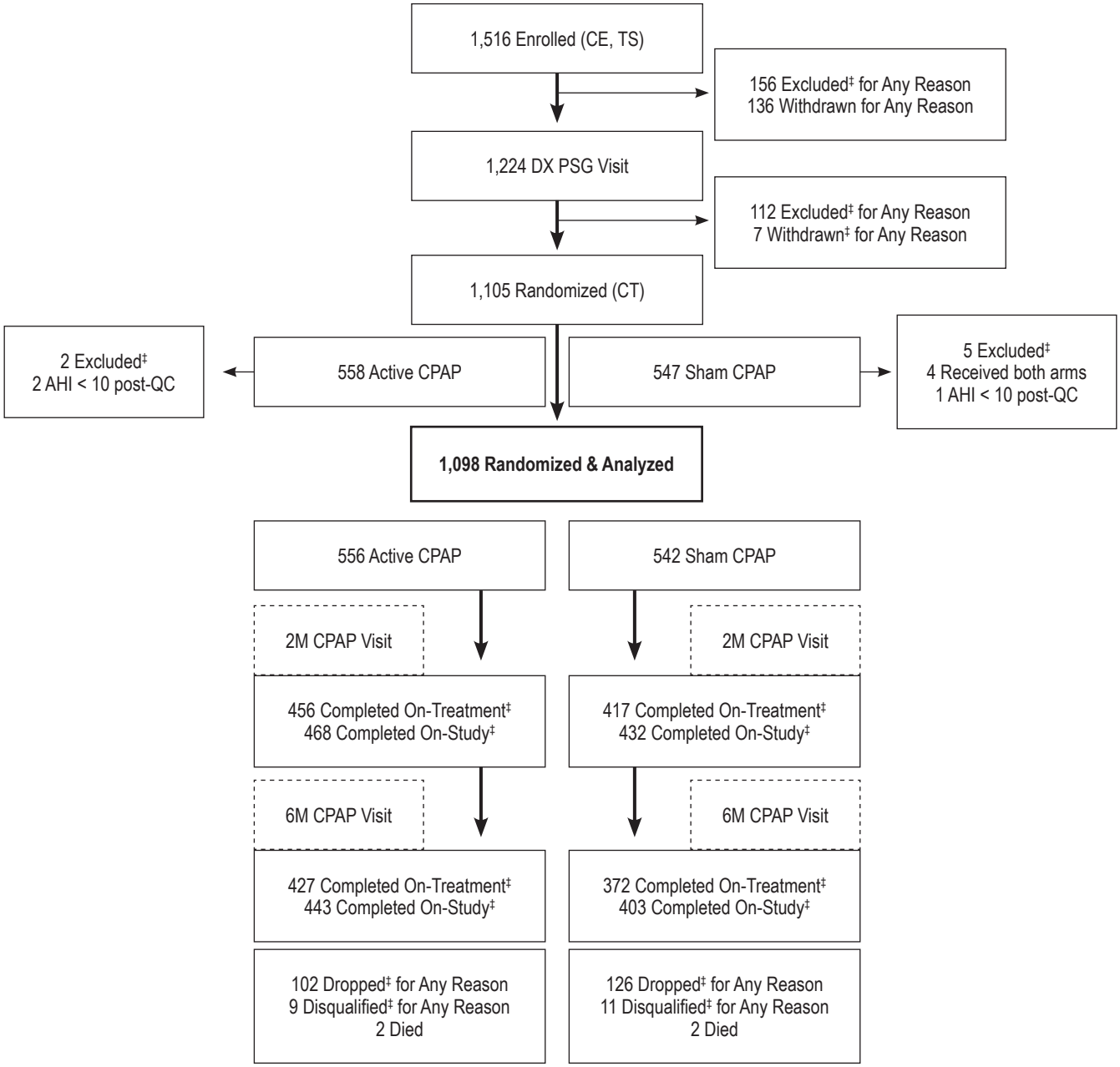

Figure 1—Participant flow diagram. Study visits included: (1) Clinical Evaluation (CE) included informed consent, baseline testing and screening, and a medical examination by a study physician; (2) Training Sessions 1 and 2 (TS) consisted of neurocognitive test training, screening, and administration of psychological tests; (3) Diagnostic Polysomnography (PSG) Visit (DX PSG Visit) involved an overnight diagnostic sleep study, questionnaires, maintenance of wakefulness test (MWT), and the neurocognitive test battery; (4) CPAP Titration Visit (CT) included administration of questionnaires and an overnight CPAP titration PSG study conducted for both active and sham CPAP group participants to determine the optimal CPAP pressure for those in the active CPAP group; (5) CPAP Set-up Visit provided the participant with the active or sham CPAP device following the CPAP titration visit; (6) Two Month Post-CPAP Follow-Up Visit (2M CPAP Visit) represented a follow-up overnight CPAP titration PSG study, with questionnaires, psychological tests, MWT, and neurocognitive test battery; (7) Four Month Post-CPAP Follow-Up Visit (4M CPAP Visit) consisted of questionnaires and a follow-up appointment with a study physician that included a physical examination and discussion of CPAP adherence, protocol compliance, safety issues, and medication changes; (8) Six Month Post-CPAP Follow-Up Visit (6M CPAP Visit) used the same protocol as the 2M-CPAP Visit; (9) Additional Follow-Up Visit allowed the participant to discuss any issues or problems; (10) Exit Interview gave the participant an opportunity to initiate other OSA treatment options. ‡Excluded: participant removed from study "pre-randomization" due to exclusion criteria (e.g., taking exclusionary medication); Withdrawn: participant quit study "pre-randomization" due to participant's choice (e.g., too busy); Dropped: participant quit study "post-randomization" due to a participant-initiated decision (e.g., did not wish to continue with protocol); Disqualified: participant removed from study "post-randomization" due to a physician-initiated decision based on medical/safety reasons (e.g., following SAE based on opinion of Physician-Observer). *All participants who dropped post-randomization were asked to continue with participant visits, even if they had discontinued therapy, based on our intention-to-treat study design. Participants who were disqualified for a medical/safety reason were asked to continue participant visits only after approval by the Site Director.* On-Treatment: participant completed visits on originally assigned treatment condition; On-Study: participant completed visits, but may or may not be on originally assigned treatment condition.

tion) performed better than those in the lower 3 quartiles (0.132 vs. 0.003, P = 0.0448; Appendix Section 2E) compared to baseline after 2 months on active CPAP. SWMT-OMD differences between quartiles were not detectable in 6M active participants or in 2M or 6M sham participants.

Active participants were significantly more alert than sham participants for the maintenance of wakefulness test-mean sleep latency (MWT-MSL) and Epworth Sleepiness Scale-Total Score (ESS-TS) at both visits (Table 3). Relative to sham, mean MWT-MSL scores only improved for those active participants with severe OSA (2M P = 0.0002; 6M P = 0.0002), and mean ESS-TS scores only improved for those active participants with moderate and severe OSA at each visit (2M P = 0.0236, P = 0.0005; 6M P = 0.0106, P = 0.0010). For active participants, greater CPAP adherence was associated with greater subjective alertness (ESS-TS; Appendix Section 4A). For subjectively sleepy participants (baseline ESS-TS > 10), average change from baseline differed between arms for 6M SWMT-OMD (active 0.150, sham 0.014, P = 0.0433; Appendix Section 2F) but not for 2M SWMT-OMD or the other primary outcomes at 2M or 6M. No differences between arms in mean change from baseline were observed for objectively sleepy participants (baseline MWT-MSL ≤ 14.5); but for this subgroup, a mild correlation between changes from baseline for the MWT-MSL and the 2M SWMT-OMD was detected in the active group (SCC = 0.2084, P = 0.0395; Appendix Section 2G).

Table 1—Baseline randomization factors, demographics, and sleep study data for APPLES participants randomized to active vs. sham CPAP†

	Active CPAP‡ Mean (SD) or Count (%)	Sham CPAP‡ Mean (SD) or Count (%)
Randomization Factors		
Sex		
Male (%)	363 (65.3)	356 (65.7)
Female (%)	193 (34.7)	186 (34.3)
Race		
White (%)	424 (76.3)	411 (75.8)
Not White (%)	132 (23.7)	131 (24.2)
OSA Severity		
Mild OSA (%)	78 (14.0)	71 (13.1)
Moderate OSA (%)	174 (31.3)	170 (31.4)
Severe OSA (%)	304 (54.7)	301 (55.5)
Demographics		
Age (y)	52.2 (12.2)*	50.8 (12.2)*
Married (%)	325 (58.5)	309 (57.0)
BMI (kg/m^2)	32.4 (7.3)	32.1 (7.0)
Highest Grade Level (y)	15.50 (2.6)	15.50 (2.6)
WASI Full-4 IQ	112.1 (12.7)	112.0 (13.3)
WASI Verbal IQ	110.0 (12.8)	110.0 (13.9)
WASI Performance IQ	111.6 (13.5)	111.4 (13.0)
Sleep Study		
Total Sleep Time (min)	375.4 (66.6)	378.3 (63.8)
Sleep Efficiency (%)	78.2 (13.3)	78.4 (12.2)
Sleep Latency (min)	18.8 (22.4)	19.0 (21.4)
REM Latency (min)	137.0 (83.6)	137.6 (82.9)
Stage 1 (% of TST)	18.8 (14.3)	18.9 (14.6)
Stage 2 (% of TST)	60.7 (13.3)	60.3 (13.8)
Stage 3 (% of TST)	2.4 (4.6)	2.6 (5.0)
Stage 4 (% of TST)	0.5 (2.0)	0.6 (2.0)
Stage REM (% of TST)	17.4 (7.2)	17.6 (6.9)
Apnea Hypopnea Index	39.7 (24.9)	40.6 (25.6)
Minimum O_2SAT – Sleep (%)	81.0 (7.6)	80.8 (8.5)
O_2SAT < 85% (% of TST)	2.2 (6.1)	2.3 (6.3)

†Hypothesis testing employed the χ^2 test for comparing groups on categorical outcomes, the t-test for approximately normally-distributed outcomes (or outcomes that could be Box-Cox transformed to an approximately normal distribution), and the Mann-Whitney-Wilcoxon summed ranks test for non-normal continuous or ordinal variables. Continuity correction was applied in χ^2 analyses for any tables where expected cell counts were ≤ 5. ‡Sample size is 1,098 (556 active, 542 sham) for all variables except Highest Grade Level (n = 1,079), WASI Verbal IQ (n = 1,091), and WASI Performance IQ (n = 1,090). Percentage values are column percentages within each factor. *P < 0.05 indicates statistical significance.

Secondary Neurocognitive Outcomes

The 7 variables selected using ICA were PFN-Reaction Time (reciprocal), Shifting Attention Test Discovery Condition-Number of Rule Changes, Psychomotor Vigilance Task (PVT)-Median Reaction Time (reciprocal), PVT-Mean Slowest 10% of Reaction Times (reciprocal), BSRT Delayed Recall-Total Recall, SWMT-Mid-Day Behavioral Index (SWMT-BMD), and SWMT-Mid-Day Activation Index (SWMT-AMD). Baseline covariate-adjusted regression models found active participants with severe OSA at 2M had better mean SWMT-BMD change scores from baseline (active 0.205, sham 0.011, P = 0.0031). Less attentional effort[9] during task performance compared to baseline (SWMT-AMD electrophysiologic score) was detected for active participants with mild OSA at 2M (active -0.050, sham 0.317, P = 0.0450). No differences in means between arms were observed for any other secondary outcomes (Appendix Section 3).

Safety

Incidence proportions for participants with ≥ 1 post-randomization serious adverse events were CVD: active 0.00719, sham 0.01107, P = 0.504; MVA: no SAEs; and deaths: active 0.00360, sham 0.00369, P = 0.9797 (Appendix Section 9).

DISCUSSION

Limitations in the research on OSA and neurocognitive function include inconsistent findings, small sample sizes, non-comprehensive test batteries, inadequate control groups, and short treatment durations.[19-35] APPLES was designed to address these limitations by assessing the sham-controlled, long-term efficacy of CPAP therapy on neurocognitive function in a study with comprehensive tests of major neurocognitive domains and

Table 2—Comparisons between participants randomized to active vs. sham CPAP on primary neurocognitive outcomes: mean estimates from regression models without and with covariate adjustment†

Visits/OSA Severity	Active CPAP: Mean Estimate (95% CI LB – UB)	Sham CPAP: Mean Estimate (95% CI LB – UB)	P Value
Pathfinder Number Test Total Time (PFN-TOTL)‡			
DX (Active n = 554; Sham n = 542)	23.32 (22.88 – 23.78)	23.08 (22.64 – 23.54)	0.4538
2M (Active n = 453; Sham n = 418)	23.56 (23.05 – 24.10)	22.92 (22.41 – 23.45)	0.0860
6M (Active n = 442; Sham n = 401)	23.48 (22.98 – 24.00)	23.01 (22.51 – 23.54)	0.2103
COVARIATE-Adjusted			
2M (n = 868)			
Mild OSA	23.11 (22.59 – 23.66)	23.06 (22.43 – 23.73)	0.9039
Moderate OSA	23.34 (22.93 – 23.77)	23.24 (22.87 – 23.63)	0.7123
Severe OSA	23.08 (22.75 – 23.42)	22.9 (22.64 – 23.22)	0.4121
6M (n = 838)			
Mild OSA	23.12 (22.58 – 23.69)	22.97 (22.30 – 23.69)	0.7389
Moderate OSA	23.35 (22.91 – 23.81)	23.16 (22.72 – 23.61)	0.5280
Severe OSA	23.09 (22.73 – 23.47)	22.84 (22.51 – 23.18)	0.3003
Buschke Selective Reminding Test Sum Recall (BSRT-SR)			
DX (Active n = 556; Sham n = 541)	49.72 (48.95 – 50.48)	49.86 (49.09 – 50.64)	0.7936
2M (Active n = 453; Sham n = 421)	52.32 (51.50 – 53.13)	51.95 (51.10 – 52.80)	0.5444
6M (Active n = 442; Sham n = 402)	54.09 (53.26 – 54.91)	54.28 (53.41 – 55.13)	0.7569
COVARIATE-Adjusted			
2M (n = 870)			
Mild OSA	53.69 (52.14 – 55.24)	52.99 (51.15 – 54.83)	0.5659
Moderate OSA	53.38 (52.31 – 54.46)	52.73 (51.63 – 53.83)	0.4004
Severe OSA	52.60 (51.82 – 53.38)	52.35 (51.55 – 53.15)	0.6591
6M (n = 838)			
Mild OSA	54.20 (52.65 – 55.75)	55.98 (54.27 – 57.70)	0.1320
Moderate OSA	54.20 (53.11 – 55.28)	54.83 (53.74 – 55.92)	0.4212
Severe OSA	55.39 (54.63 – 56.14)	54.90 (54.11 – 55.70)	0.3764
Sustained Working Memory Test Overall Mid-Day Index (SWMT-OMD)			
2M (Active n = 437; Sham n = 394)	0.035 (-0.019 – 0.090)	-0.074 (-0.133 – -0.015)	0.0074*
6M (Active n = 426; Sham n = 374)	0.072 (0.012 – 0.132)	0.018 (-0.046 – 0.082)	0.2254
COVARIATE-Adjusted			
2M (n = 828)			
Mild OSA	-0.017 (-0.152 – 0.119)	0.011 (-0.135 – 0.157)	0.7834
Moderate OSA	0.016 (-0.087 – 0.120)	-0.032 (-0.128 – 0.064)	0.4950
Severe OSA	0.054 (-0.017 – 0.125)	-0.112 (-0.197 – -0.028)	0.0031*
6M (n = 796)			
Mild OSA	0.023 (-0.132 – 0.177)	-0.046 (-0.216 – 0.123)	0.5515
Moderate OSA	0.017 (-0.086 – 0.121)	0.008 (-0.108 – 0.125)	0.9101
Severe OSA	0.113 (0.031 – 0.195)	0.039 (-0.046 – 0.124)	0.2176

DX, Diagnostic Polysomnography Visit; 2M, Two-Month Post-CPAP Follow-Up Visit; 6M, Six-Month Post-CPAP Follow-Up Visit. †Analysis details included in Appendix Section 2B. ‡PFN-TOTL data were reciprocal transformed for analysis and back-transformed for reporting. *P < 0.0307 indicates statistical significance. None of the primary neurocognitive analyses (designated as the 6 primary analyses performed at 2M and 6M unadjusted for covariates) were significant after adjustment for multiple comparisons (Appendix Section 2C).

adequate statistical power. Using these study design parameters, we showed a difference between active vs. sham CPAP for only the E/F variable at 2 months.

Once analyses were conducted by OSA severity and adjusted for covariates, we detected slight improvement in the active arm for both the primary and two of the secondary E/F vari-

ables in participants with an AHI > 30 (severe OSA) at the 2M visit. Dividing patients into quartiles by baseline oxygenation also showed short-term improvement in the active arm at the 2M visit for the primary E/F variable. These results suggest disease severity may be important for detecting improvement in neurocognitive outcomes. As measures of disease severity, both AHI[30,36] and oxygen saturation have been previously implicated in the etiology of the OSA-associated neurocognitive dysfunction. Although some studies on OSA[36] and hypoxemic patients[37] failed to find a relationship between measures of oxygen saturation and neurocognitive function, others,[38] including the large-scale Sleep Heart Health Study,[18] reported that OSA patients with decreased oxygen saturation were more cognitively impaired compared to those without significant desaturations. Additionally, baseline analyses of the APPLES population found that severity of oxygen desaturation was weakly associated with worse neurocognitive performance on some measures of intelligence, attention, and processing speed.[17]

CPAP has been demonstrated to improve OSA-related sleepiness.[39] We found that active participants were less sleepy, whether measured by an objective (MWT-MSL) or subjective (ESS-TS) measure, and participants with more severe OSA benefited the most from active CPAP. In a subgroup of those who were sleepy at baseline, change from baseline in the E/F measure was significantly different on average between arms for subjectively sleepy individuals at 6 months and was correlated with change in objective sleepiness at 2 months, suggesting sleepiness may be associated with one domain of OSA-related neurocognition.

Table 3—Measures of objective and subjective sleepiness: comparison of means by visit between participants randomized to active vs. sham CPAP†

	Active CPAP Mean (SD)	Sham CPAP Mean (SD)	P Value
MWT Mean Sleep Latency (objective sleepiness)			
DX (n = 1,086; Active n = 551; Sham n = 535)	17.13 (3.86)	16.95 (4.13)	0.6540
Mild OSA (n = 147)	17.51 (3.71)	17.62 (3.38)	0.9778
Moderate OSA (n = 340)	17.74 (3.50)	17.76 (3.68)	0.8314
Severe OSA (n = 599)	16.68 (4.05)	16.35 (4.43)	0.5018
2M (n = 853; Active n = 445; Sham n = 408)	17.96 (3.40)	17.27 (3.89)	0.0052*
Mild OSA (n = 108)	17.52 (3.60)	18.21 (2.94)	0.2476
Moderate OSA (n = 253)	17.91 (3.39)	18.14 (2.93)	0.7520
Severe OSA (n = 492)	18.10 (3.35)	16.63 (4.34)	0.0002*
6M (n = 827; Active n = 432; Sham n = 395)	18.11 (3.27)	17.34 (3.82)	0.0022*
Mild OSA (n = 110)	17.77 (4.00)	17.89 (3.27)	0.7630
Moderate OSA (n = 246)	17.90 (3.41)	18.18 (3.27)	0.5170
Severe OSA (n = 471)	18.30 (2.98)	16.78 (4.10)	0.0002*
ESS Total Score (subjective sleepiness)			
DX (n = 1,098; Active n = 556; Sham n = 542)	10.07 (4.26)	10.09 (4.39)	0.9291
Mild OSA (n = 149)	10.10 (4.55)	9.73 (4.43)	0.6152
Moderate OSA (n = 344)	9.57 (4.13)	9.75 (4.56)	0.7040
Severe OSA (n = 605)	10.35 (4.24)	10.37 (4.28)	0.9537
2M (n = 875; Active n = 453; Sham n = 422)	7.86 (4.20)	8.89 (4.31)	0.0004*
Mild OSA (n = 111)	8.59 (4.31)	7.90 (4.01)	0.3886
Moderate OSA (n = 261)	7.25 (3.89)	8.39 (4.29)	0.0236*
Severe OSA (n = 503)	8.00 (4.31)	9.34 (4.34)	0.0005*
6M (n = 846; Active n = 443; Sham n = 403)	7.39 (4.21)	8.41 (4.18)	0.0005*
Mild OSA (n = 113)	8.37 (4.64)	7.64 (3.98)	0.3796
Moderate OSA (n = 250)	7.07 (3.87)	8.43 (4.55)	0.0106*
Severe OSA (n = 483)	7.31 (4.25)	8.56 (4.02)	0.0010*

†The MWT was administered at 10:00, 12:00, 14:00, and 16:00 at the Diagnostic Polysomnography (DX), Two-Month Post-CPAP Follow-Up (2M), and Six-Month Post-CPAP Follow-Up (6M) visits. The mean sleep latency was calculated using the 4 trials from a given visit, and required that at least 3 of the 4 visits trials were performed and validated. The ESS was administered the evening before the PSG at the DX, CPAP, 2M, 4M, and 6M Visits. *P < 0.05 indicates statistical significance.

To address whether CPAP may only improve cognition in CPAP-compliant individuals, we repeated the primary outcome analyses restricted to a CPAP-adherent group. That subgroup analysis no longer detected a difference in means between arms for any of the primary outcomes at any visit. These analyses are difficult to interpret due to a smaller sample size, a difference in mean baseline IQ Verbal WASI between sham and active CPAP in this self-selected subpopulation, and differences in several baseline features between adherent and non-adherent individuals. Interestingly, baseline features associated with better adherence included increased age, higher IQ, white ethnicity, being married, and poorer sleep quality (e.g., decreased sleep efficiency, longer sleep onset, longer REM onset). When we performed an adjustment for potential baseline confounders between CPAP adherence and 1NC outcomes, the study's primary findings remained unchanged, although we recognize that additional analyses remain to be performed to explore other methods of adjustment for variable adherence and retention (Appendix Section 7E).

The detection of CPAP effects for only the primary E/F variable suggests this test is a more sensitive measure for subtle neurocognitive changes in that it combines a cognitive task with simultaneous EEG measures of brain function. However, the fact that these effects could only be detected at 2 months, that there was some evidence for worsening in the sham arm at 2 months, that circadian confounding may have been present (Appendix Section 1B), and that effects of CPAP were minor compared to effects of caffeine or diphenhydramine[40-42] on this measure in other studies must be considered in interpreting the significance of this finding. Further, given the number of statistical tests conducted, these findings may reflect type 1 statistical error (Appendix Section 2C).

There are limitations related to the study sample. Although participants with severe OSA were included, those who had the lowest oxygen saturation, significant sleepiness including a history of sleepiness-related accidents, or major cardiac comorbidities were excluded from participation. Participants also willingly deferred effective treatment for up to 6 months in the sham arm; a majority of these participants were recruited from advertisements rather than clinically referred for OSA; and

participants had lower CPAP adherence than expected despite close follow-up to troubleshoot and encourage adherence in our participants. A majority of sham participants correctly guessed their treatment assignment. These factors collectively may have resulted in a sample with relatively lower susceptibility to the neurocognitive effects of OSA and a subsequent reduced response to treatment.

In summary, active CPAP improved the primary measure of E/F at 2 months, and for those participants with severe OSA, improved both the primary and two secondary measures of E/F at the same time point of the study. There is evidence that deficits in neurobehavioral function vary significantly between individuals, are stable within individuals, and may involve a trait-like vulnerability to impairment from sleep loss.[43] The cognitive reserve theory may also be relevant for our findings; individual differences in how the brain processes tasks may allow some to cope with greater insult by using preexisting cognitive processes or by enlisting compensatory processes before performance is detrimentally impacted.[44] While it is possible that our intelligent population (WASI IQ) may have had less neurocognitive impairment due to OSA because they had more cognitive reserve, resulting in their ability to maintain performance, adjusting for WASI IQ in the models did not change the results. It is also possible that the lengthy list of baseline covariates we tested is not properly aligned with more complex neurocognitive traits; perhaps neurocognitive testing incorporating advanced electroencephalographic and imaging technology will be necessary to identify potential changes in neurocognitive outcomes in OSA patients. We believe this study supports the theory that OSA is a multifaceted disorder with many comorbidities and outcomes; we believe that the mixed results from prior studies and the limited effect of CPAP on E/F measures of neurocognition in this study suggest the existence of a complex OSA-neurocognitive relationship, and that clinicians should consider disease severity, sleepiness, and individual differences including treatment adherence in managing their patients with CPAP.

ACKNOWLEDGMENTS

APPLES was funded by contract 5UO1-HL-068060 from the National Heart, Lung and Blood Institute. The APPLES pilot studies were supported by grants from the American Academy of Sleep Medicine and the Sleep Medicine Education and Research Foundation to Stanford University and by the National Institute of Neurological Disorders and Stroke (N44-NS-002394) to SAM Technology.

In addition, APPLES investigators gratefully recognize the vital input and support of Dr. Sylvan Green who died before the results of this trial were analyzed, but was instrumental in its design and conduct.

Administrative Core

Clete A. Kushida, MD, PhD; Deborah A. Nichols, MS; Eileen B. Leary, BA, RPSGT; Pamela R. Hyde, MA; Tyson H. Holmes, PhD; Daniel A. Bloch, PhD; William C. Dement, MD, PhD

Data Coordinating Center

Daniel A. Bloch, PhD; Tyson H. Holmes, PhD; Deborah A. Nichols, MS; Rik Jadrnicek, Microflow, Ric Miller, Microflow, Usman Aijaz, MS; Aamir Farooq, PhD; Darryl Thomander, PhD; Chia-Yu Cardell, RPSGT; Emily Kees, Michael E. Sorel, MPH; Oscar Carrillo, RPSGT; Tami Crabtree, MS; Booil Jo, PhD; Ray Balise, PhD; Tracy Kuo, PhD

Clinical Coordinating Center

Clete A. Kushida, MD, PhD, William C. Dement, MD, PhD, Pamela R. Hyde, MA, Rhonda M. Wong, BA, Pete Silva, Max Hirshkowitz, PhD, Alan Gevins, DSc, Gary Kay, PhD, Linda K. McEvoy, PhD, Cynthia S. Chan, BS, Sylvan Green, MD

Clinical Centers

Stanford University

Christian Guilleminault, MD; Eileen B. Leary, BA, RPSGT; David Claman, MD; Stephen Brooks, MD; Julianne Blythe, PA-C, RPSGT; Jennifer Blair, BA; Pam Simi, Ronelle Broussard, BA; Emily Greenberg, MPH; Bethany Franklin, MS; Amirah Khouzam, MA; Sanjana Behari Black, BS, RPSGT; Viola Arias, RPSGT; Romelyn Delos Santos, BS; Tara Tanaka, PhD

University of Arizona

Stuart F. Quan, MD; James L. Goodwin, PhD; Wei Shen, MD; Phillip Eichling, MD; Rohit Budhiraja, MD; Charles Wynstra, MBA; Cathy Ward, Colleen Dunn, BS; Terry Smith, BS; Dane Holderman, Michael Robinson, BS; Osmara Molina, BS; Aaron Ostrovsky, Jesus Wences, Sean Priefert, Julia Rogers, BS; Megan Ruiter, BS; Leslie Crosby, BS, RN

St. Mary Medical Center

Richard D. Simon Jr., MD; Kevin Hurlburt, RPSGT; Michael Bernstein, MD; Timothy Davidson, MD; Jeannine Orock-Takele, RPSGT; Shelly Rubin, MA; Phillip Smith, RPSGT; Erica Roth, RPSGT; Julie Flaa, RPSGT; Jennifer Blair, BA; Jennifer Schwartz, BA; Anna Simon, BA; Amber Randall, BA

St. Luke's Hospital

James K. Walsh, PhD, Paula K. Schweitzer, PhD, Anup Katyal, MD, Rhody Eisenstein, MD, Stephen Feren, MD, Nancy Cline, Dena Robertson, RN, Sheri Compton, RN, Susan Greene, Kara Griffin, MS, Janine Hall, PhD

Brigham and Women's Hospital

Daniel J. Gottlieb, MD, MPH, David P. White, MD, Denise Clarke, BSc, RPSGT, Kevin Moore, BA, Grace Brown, BA, Paige Hardy, MS, Kerry Eudy, PhD, Lawrence Epstein, MD, Sanjay Patel, MD

**Sleep HealthCenters for the use of their clinical facilities to conduct this research*

Consultant Teams

Methodology Team: Daniel A. Bloch, PhD, Sylvan Green, MD, Tyson H. Holmes, PhD, Maurice M. Ohayon, MD, D Sc, David White, MD, Terry Young, PhD

Sleep-Disordered Breathing Protocol Team: Christian Guilleminault, MD, Stuart Quan, MD, David White, MD

EEG/Neurocognitive Function Team: Jed Black, MD, Alan Gevins, DSc, Max Hirshkowitz, PhD, Gary Kay, PhD, Tracy Kuo, PhD

Mood and Sleepiness Assessment Team: Ruth Benca, MD, PhD, William C. Dement, MD, PhD, Karl Doghramji, MD, Tracy Kuo, PhD, James K. Walsh, PhD

Quality of Life Assessment Team: W. Ward Flemons, MD, Robert M. Kaplan, PhD

APPLES Secondary Analysis-Neurocognitive (ASA-NC) Team: Dean Beebe, PhD, Robert Heaton, PhD, Joel Kramer, PsyD, Ronald Lazar, PhD, David Loewenstein, PhD, Frederick Schmitt, PhD

National Heart, Lung, and Blood Institute (NHLBI)

Michael J. Twery, PhD, Gail G. Weinmann, MD, Colin O. Wu, PhD

Data and Safety Monitoring Board (DSMB)

Seven year term: Richard J. Martin, MD (Chair), David F. Dinges, PhD, Charles F. Emery, PhD, Susan M. Harding MD, John M. Lachin, ScD, Phyllis C. Zee, MD, PhD

Other term: Xihong Lin, PhD (2 yrs), Thomas H. Murray, PhD (1 yr)

DISCLOSURE STATEMENT

This study was funded by Respironics, Inc. Dr. Kushida received research support through Stanford University from ResMed, Pacific Medico Co., Ltd., Merck & Co., Cephalon, Ventus Medical, Jazz Pharmaceuticals, and Respironics. Dr. Walsh receives research support from Pfizer, Merck & Co., Somnus, Vanda Pharmaceuticals, Neurogen, Sanofi-Aventis, Ventus Medical, Respironics, Apnex, and Jazz Pharmaceuticals. He has consulted for Sanofi-Aventis, Respironics, Transcept, Neurogen, Glaxo-SmithKline, Eli Lilly, Merck & Co., Kingsdown, Vanda Pharmaceuticals, Ventus Medical, Vivus Inc., and Somnus Therapeutics, Inc. Dr. Simon has consulted for Asante Communications and has received sponsorship fees from World Class CME. Dr. White is the chief medical officer for Philips Respironics. Dr. Schweitzer has received research support from Apnex Medical, Merck Sharpe & Dohme, Vanda Pharmaceuticals, and Ventus Medical. She also serves on the speaker bureau for Somaxon Pharmaceuticals. Dr. Hirshkowitz serves on the speaker bureau for Cephalon and Somaxon Pharmaceuticals. Dr. Gevins is employed by Technology, Inc. Dr. Kay is the president of a contract research organization; clients in the last 12 months: Allergan, Arena, Factor Nutrition, Helicon, Merck & Co., Pfizer, Shire, Vivus Inc., and Watson. The other authors have indicated no financial conflicts of interest.

REFERENCES

1. Li C, Ford ES, Zhao G, Croft JB, Balluz LS, Mokdad AH. Prevalence of self-reported clinically diagnosed sleep apnea according to obesity status in men and women: National Health and Nutrition Examination Survey, 2005-2006. Prev Med;51:18-23.
2. Sullivan CE, Issa FG, Berthon-Jones M, Eves L. Reversal of obstructive sleep apnoea by continuous positive airway pressure applied through the nares. Lancet 1981;1:862-5.
3. Kushida CA, Nichols DA, Quan SF, et al. The Apnea Positive Pressure Long-term Efficacy Study (APPLES): rationale, design, methods, and procedures. J Clin Sleep Med 2006;2:288-300.
4. American Academy of Sleep Medicine Task Force. Sleep-related breathing disorders in adults: recommendations for syndrome definition and measurement techniques in clinical research. The Report of an American Academy of Sleep Medicine Task Force. Sleep 1999;22:667-89.
5. Krieger J, Kurtz D, Petiau C, Sforza E, Trautmann D. Long-term compliance with CPAP therapy in obstructive sleep apnea patients and in snorers. Sleep 1996;19(9 Suppl):S136-43.
6. McArdle N, Devereux G, Heidarnejad H, Engleman HM, Mackay TW, Douglas NJ. Long-term use of CPAP therapy for sleep apnea/hypopnea syndrome. Am J Respir Crit Care Med 1999;159(4 Pt 1):1108-14.
7. Farre R, Hernandez L, Montserrat JM, Rotger M, Ballester E, Navajas D. Sham continuous positive airway pressure for placebo-controlled studies in sleep apnoea. Lancet 1999;353:1154.
8. Hannay J, ed. Experimental techniques in human neuropsychology. New York: Oxford University Press, 1986.
9. Gevins A, Smith ME, McEvoy LK, et al. A cognitive and neurophysiological test of change from an individual's baseline. Clin Neurophysiol 2011;122:114-20.
10. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979;35:549-56.
11. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13-22.
12. Daniel WW. Applied Nonparametric Statistics. 2nd ed. Boston, MA: PWS-Kent Publishing Company, 1990.
13. Stone JV. Independent component analysis: an introduction. Trends Cogn Sci 2002;6:59-64.
14. Follmann D, Fay MP, Proschan M. Chop-lump tests for vaccine trials. Biometrics 2009;65:885-93.
15. SAS v. 9.2, SAS Institute Inc. Cary, NC.
16. R Development Core Team. R: A language and environment for statistical computing. 2009:ISBN 3-900051-07-0, URL http://www.R-project.org.
17. Quan SF, Chan CS, Dement WC, et al. The association between obstructive sleep apnea and neurocognitive performance--the Apnea Positive Pressure Long-term Efficacy Study (APPLES). Sleep 2011;34:303-14B.
18. Quan SF, Wright R, Baldwin CM, et al. Obstructive sleep apnea-hypopnea and neurocognitive functioning in the Sleep Heart Health Study. Sleep Med 2006;7:498-507.
19. Canessa N, Castronovo V, Cappa SF, et al. Obstructive sleep apnea: brain structural changes and neurocognitive function before and after treatment. Am J Respir Crit Care Med 2011;183:1419-26.
20. Lau EY, Eskes GA, Morrison DL, Rajda M, Spurr KF. Executive function in patients with obstructive sleep apnea treated with continuous positive airway pressure. J Int Neuropsychol Soc 2010;16:1077-88.
21. Sforza E, Roche F, Thomas-Anterion C, et al. Cognitive function and sleep related breathing disorders in a healthy elderly population: the SYNAPSE study. Sleep 2010;33:515-21.
22. Aloia MS, Ilniczky N, Di Dio P, Perlis ML, Greenblatt DW, Giles DE. Neuropsychological changes and treatment compliance in older adults with sleep apnea. J Psychosom Res 2003;54:71-6.
23. Aloia MS, Sweet LH, Jerskey BA, Zimmerman M, Arnedt JT, Millman RP. Treatment effects on brain activity during a working memory task in obstructive sleep apnea. J Sleep Res 2009;18:404-10.
24. Ancoli-Israel S, Palmer BW, Cooke JR, et al. Cognitive effects of treating obstructive sleep apnea in Alzheimer's disease: a randomized controlled study. J Am Geriatr Soc 2008;56:2076-81.
25. Bardwell WA, Ancoli-Israel S, Berry CC, Dimsdale JE. Neuropsychological effects of one-week continuous positive airway pressure treatment in patients with obstructive sleep apnea: a placebo-controlled study. Psychosom Med 2001;63:579-84.
26. Lim W, Bardwell WA, Loredo JS, et al. Neuropsychological effects of 2-week continuous positive airway pressure treatment and supplemental oxygen in patients with obstructive sleep apnea: a randomized placebo-controlled study. J Clin Sleep Med 2007;3:380-6.
27. Twigg GL, Papaioannou I, Jackson M, et al. Obstructive sleep apnea syndrome is associated with deficits in verbal but not visual memory. Am J Respir Crit Care Med 2010;182:98-103.
28. Barbe F, Mayoralas LR, Duran J, et al. Treatment with continuous positive airway pressure is not effective in patients with sleep apnea but no daytime sleepiness. A randomized, controlled trial. Ann Intern Med 2001;134:1015-23.
29. Naegele B, Pepin JL, Levy P, Bonnet C, Pellat J, Feuerstein C. Cognitive executive dysfunction in patients with obstructive sleep apnea syndrome (OSAS) after CPAP treatment. Sleep 1998;21:392-7.
30. Naegele B, Thouvard V, Pepin JL, et al. Deficits of cognitive executive functions in patients with sleep apnea syndrome. Sleep 1995;18:43-52.

31. Kim HC, Young T, Matthews CG, Weber SM, Woodward AR, Palta M. Sleep-disordered breathing and neuropsychological deficits. A population-based study. Am J Respir Crit Care Med 1997;156:1813-9.
32. Redline S, Strauss ME, Adams N, et al. Neuropsychological function in mild sleep-disordered breathing. Sleep 1997;20:160-7.
33. Bedard MA, Montplaisir J, Richer F, Rouleau I, Malo J. Obstructive sleep apnea syndrome: pathogenesis of neuropsychological deficits. J Clin Exp Neuropsychol 1991;13:950-64.
34. Cheshire K, Engleman H, Deary I, Shapiro C, Douglas NJ. Factors impairing daytime performance in patients with sleep apnea/hypopnea syndrome. Arch Intern Med 1992;152:538-41.
35. Ingram F, Henke KG, Levin HS, Ingram PT, Kuna ST. Sleep apnea and vigilance performance in a community-dwelling older sample. Sleep 1994;17:248-52.
36. Greenberg GD, Watson RK, Deptula D. Neuropsychological dysfunction in sleep apnea. Sleep 1987;10:254-62.
37. Fix AJ, Golden CJ, Daughton D, Kass I, Bell CW. Neuropsychological deficits among patients with chronic obstructive pulmonary disease. Int J Neurosci 1982;16:99-105.
38. Findley LJ, Barth JT, Powers DC, Wilhoit SC, Boyd DG, Suratt PM. Cognitive impairment in patients with obstructive sleep apnea and associated hypoxemia. Chest 1986;90:686-90.
39. Jenkinson C, Davies RJ, Mullins R, Stradling JR. Comparison of therapeutic and subtherapeutic nasal continuous positive airway pressure for obstructive sleep apnoea: a randomised prospective parallel trial. Lancet 1999;353:2100-5.
40. Gevins A, Smith ME, McEvoy LK. Tracking the cognitive pharmacodynamics of psychoactive substances with combinations of behavioral and neurophysiological measures. Neuropsychopharmacology 2002;26:27-39.
41. McEvoy LK, Smith ME, Fordyce M, Gevins A. Characterizing impaired functional alertness from diphenhydramine in the elderly with performance and neurophysiologic measures. Sleep 2006;29:957-66.
42. Gevins A, Ilan AB, Jiang A, Sam-Vargas L, Baum C, Chan CS. Combined neuropsychological and neurophysiological assessment of drug effects on groups and individuals. J Psychopharmacol 2011;25:1062-75.
43. Van Dongen HP, Baynard MD, Maislin G, Dinges DF. Systematic interindividual differences in neurobehavioral impairment from sleep loss: evidence of trait-like differential vulnerability. Sleep 2004;27:423-33.
44. Stern Y. Cognitive reserve. Neuropsychologia 2009;47:2015-28.

APPENDIX

Table of Contents

Table S1—Summary of neurocognitive outcome data for pilot studies

		Difference from Baseline			
Test: Neurocognitive	**Baseline**	**Active CPAP**	**Sham CPAP**	**Effect Size**	**Group P**
SWMT - Performance	0.315 ± 0.032	0.002 ± 0.030	-0.018 ± 0.032	0.62	0.38
SWMT - Electrophysiologic	50.10 ± 3.75	2.992 ± 7.378	-3.724 ± 5.090‡	1.32	0.03*
SWMT - Composite Index	**	1.250 ± 3.327	-1.714 ± 2.928	1.01	0.08
Trails Making A Test, TMA (median RT-msec)	683.6 ± 108.2	-62.8 ± 73.59‡	-88.8 ± 127.0	0.20	1.00
Buschke Selective Reminding Test, BSRT (total recall)	103.3 ± 14.0	15.4 ± 11.48‡	11.5 ± 14.58‡	0.26	0.56§

**Since the SWMT composite index is a measure of the difference from baseline, there is no baseline value to report; RT, reaction time; ‡P < 0.05 for active Baseline vs. Post-CPAP values or sham Baseline vs. Post-CPAP values; *P < 0.05 for the active vs. sham difference scores (Post-CPAP minus Baseline); §P < 0.05 for active vs. sham Post-CPAP values; Effect Size = absolute difference between active vs. sham groups difference scores / standard deviation of sham group difference scores.

SECTION 1. APPLES STUDY DESIGN

1A. Sample Size Calculations

Two pilot studies[1] were completed at Stanford University with a total of 16 participants (14 men and 2 women, aged 28-65 years). Eight participants were assigned, in random order, to active CPAP and 8 to sham CPAP. These pilot studies demonstrated the feasibility of the methods that were employed in APPLES and provided preliminary data used in our sample size calculations.

Sample size was calculated to permit detection of treatment effects at least as large as those estimated from the two pilot studies (n = 16) with 90% power and a type I error rate of 5%. The APPLES sample size was based on pilot study results for the Pathfinder Number Test because this test required the largest sample size (Table S1) among the 3 primary outcome measures. Allowing for 3 interim analyses and a 20% dropout (estimated based on our clinical research experience and 2 studies measuring long-term CPAP adherence)[2,3] resulted in a randomization target of 1,100 total participants.

The following are additional justifications as to why 1,100 participants are necessary for this study (*from APPLES Protocol Section 6.6*):

1. *Large sample sizes are needed for neurocognitive outcomes in CPAP-treated OSA subjects.*
   Although the effect sizes for impairment in various cognitive domains reported by Engleman and colleagues[4] ranged from ≤ 0.3 to > 3.0, most studies found effect sizes < 0.3. Although the sum of the two pilot studies consisted of a limited sample size of eight subjects in each treatment arm, we found a range of effect sizes (0.20 to 2.46) similar to those found by Engleman and colleagues in their review. Smaller effect sizes require larger sample sizes to achieve statistical significance. We estimate an effect size of 0.2 for the Pathfinder Number Test. The effect size of 0.2 translates to the clinically significant difference of 26 msec in reaction time between the Active and Sham CPAP groups for this test. An effect size ≥ 0.2 also translates to clinically significant differences between the groups for the other two primary outcome measures.
2. *We are examining neurocognitive outcomes in response to CPAP therapy for a wide spectrum of OSA severity.*
   The effect sizes previously reported were typically related to patients with a limited severity range of OSA; the more severe the case of OSA, the greater the neurocognitive impairment.[5,6] Since our study will include subjects varying over the entire range of OSA severity, we need a larger sample size than would be indicated by the prior studies.
3. *Prior studies had small sample sizes and showed conflicting results.*
   The majority of case-control or randomized controlled studies evaluating neurocognitive function and OSA had sample sizes < 50 OSA subjects. The conflicting results of these studies could be due to the following: a) low sample sizes, b) tests in any one study did not cover a range of neurocognitive domains, and c) lack of multiple measures within each neurocognitive domain. Our study will avoid these methodological limitations through a large sample size and multiple measures within several neurocognitive domains.
4. *Secondary neurocognitive outcome measures will also be explored.*
   Based on prior smaller studies, CPAP treatment was shown to improve various domains of neurocognitive function in a clinically important way. Treatments will be compared statistically for these secondary neurocognitive outcome measures.

Pilot Studies – Results (from APPLES Protocol Section 3.3.2)

The main results from the pilot studies are summarized in Table S1. There was a wide variability in the therapeutic effect sizes for changes in neurocognitive function, ranging from small (0.01) to large (1.32). For the SWMT, we focused our analysis on the third test interval, which occurred at 2:30 pm. The effect of active vs. sham CPAP therapy was examined for a number of behavioral and EEG variables independently. A summary behavioral measure from the task improved in the active CPAP group whereas the sham group showed a small decrease on the same measure, resulting in a treatment effect size of 0.62 (P = 0.38). Similarly, the active CPAP group showed a decrease in an electrophysiologic variable associated with drowsiness, whereas the sham group showed an increase in the same variable, resulting in a treatment effect size of 1.32 (P = 0.03).

In addition to examination of the neurophysiological and behavioral variables in isolation, we also used a composite index. This index can serve as a summary measure for the degree of change in each patient following treatment. The index was weighted so that positive index values reflect relatively greater alertness in the post-treatment condition, negative values reflect relatively lower post-treatment alertness and zero reflects no change. On average, the active group showed improved alertness on this measure whereas the sham group showed decreased alertness, resulting in an effect size of 1.01 (P = 0.08). Five of the 8 subjects in the active CPAP group had positive scores, indicative of improved alertness, whereas 6 of the 8 subjects in the sham CPAP group had negative scores. Examination of the individual subject data suggests that the direction of change indicated on the SWMT composite index is in good agreement with the data from the other measures.

Other measures of neurocognitive function were also used to assess changes in attention and psychomotor function, learning and memory, as well as executive and frontal-lobe function. With respect to attention and psychomotor function, the effect sizes ranged from 0.01 to 1.02. The active group showed trends toward greater improvement compared to the sham group. For measures of learning and memory, the effect sizes ranged from 0.26 to 0.40, with an effect size of 0.26 for the BSRT; there was a significant difference for the active group between their baseline and post-CPAP values. For measures of executive and frontal-lobe function, the effect sizes ranged from 0.14 to 1.32.

1B. Primary Neurocognitive Outcome Analyses

The per-protocol intention-to-treat analyses[7] specified that all primary efficacy outcomes be regressed on study arm, days since randomization, and their interaction using generalized estimating equations (GEE)[8] to account for the repeated measures on participants over time; the primary comparison was the difference between slopes (active vs. sham) across time.

Upon presenting these initial analyses to the SC it was determined that the GEE method outlined in the protocol could not be applied across the Baseline, 2M and 6M visits for all three primary outcomes. SC decided that generalized regression models (generalized linear models [GLM] or generalized linear mixed models [GLMM]) be alternatively fit to the primary outcomes.

For CogScreen Pathfinder Number-Total Time (PFN-TOTL), repeated measure mean comparisons were estimated by GLMM.

For the Buschke Selective Reminding Test-Sum Recall (BSRT-SR), a difference was identified in the difficulty of the form versions[9] between baseline and the 2 or 6 month administrations; therefore, the Steering Committee (SC) voted that comparisons could not be made across the three visits. Instead, SC specified that comparisons be conducted separately for each post-randomization visit using GLM. For covariate-adjusted analyses, the baseline BSRT-SR score was included as a covariate.

Sustained Working Memory Test-Overall Mid-day Index (SWMT-OMD) was provided as a change-from-baseline score. SC voted that comparisons could not be made across the three visits based on the structure of this variable. Instead, comparison was of mean change-from-baseline score by visit, as estimated by a GLM fit to the dataset for 2M and 6M. The change-from-baseline score was formulated to compare the mid-day measurement at each follow-up visit (2M, 6M) against the combination of the morning, mid-day, and afternoon measurements at baseline, which advances the possibility that change scores may be confounded with diurnal variation.

1C. Intention-to-Treat Parameters

The protocol specified that analyses be conducted in accordance with the intention-to-treat principle.[10,11] On this basis, all participants who dropped (due to a participant-initiated decision) or were disqualified (due to a physician-initiated decision based on medical/safety reasons) were invited to continue attending study visits and provide protocol-specified data, even if they discontinued their originally assigned therapy. As a result, an individual analyzed as active may not have used CPAP at all or an individual analyzed as sham either may not have used CPAP (sham or active) at all or used active CPAP for a portion of the intervention period. All analyses were performed strictly based on the participants' original randomization assignments, with the exception of seven participants (3 had an AHI < 10 and were excluded after PSG quality control, and 4 had inadvertent exposure to both treatment conditions as a result of staff error rather than participant choice; the decisions to exclude these participants were made by SC). Quantities of participants On-Treatment (completed visits on originally assigned treatment condition) vs. On-Study (completed visits, but may or may not have been on originally assigned treatment condition) are reported in Figure 1 of the manuscript.

Another aspect of the intention-to-treat principle regards inclusion of individuals who were randomized but only completed a baseline visit with no post-randomization follow-up visits. For the primary neurocognitive outcomes, the protocol specified that all three visits were to be used together in a longitudinal regression analysis with GEE. This analysis included participants who had only a baseline visit, and was the analysis employed for PFN-TOTL. For BSRT-SR, analyses were run separately by visit due to differences in forms between visits (see Section 1B). Here, participants who only had a baseline visit were included in the means comparison between arms at baseline. For SWMT-OMD, the data were provided as a change score. As a result, participants who only had a baseline visit were excluded from this analysis (see Section 1B). The retention-adjusted analyses were formulated as a change-from-baseline variable for all three primary neurocognitive findings. When allowance was made for potentially informative dropout via selection modeling, results for the primary neurocognitive outcomes remain unchanged from the results reported in the main paper without this adjustment. The secondary neurocognitive analysis plan specified that all three visits were to be used together in a longitudinal mixed-model regression (GLMM). This was done for CogScreen Shifting Attention Test Discovery Condition- Rule Changes Completed Dichotomized (SAT-DIRUL), CogScreen Pathfinder Number-Reaction Time (PFN-RTC), Psychomotor Vigilance Task-Mean Slowest 10% of Reaction Times (PVT-MSRT), and PVT Median RT (PVT-MDRT); so that participants who only had a baseline visit were included in these analyses. As with BSRT-SR, analyses of BSRT Delayed Recall (BSRT-DR) were performed by visit for each of the three visits. The SWMT-Activation Index: Mid-day (SWMT-AMD)

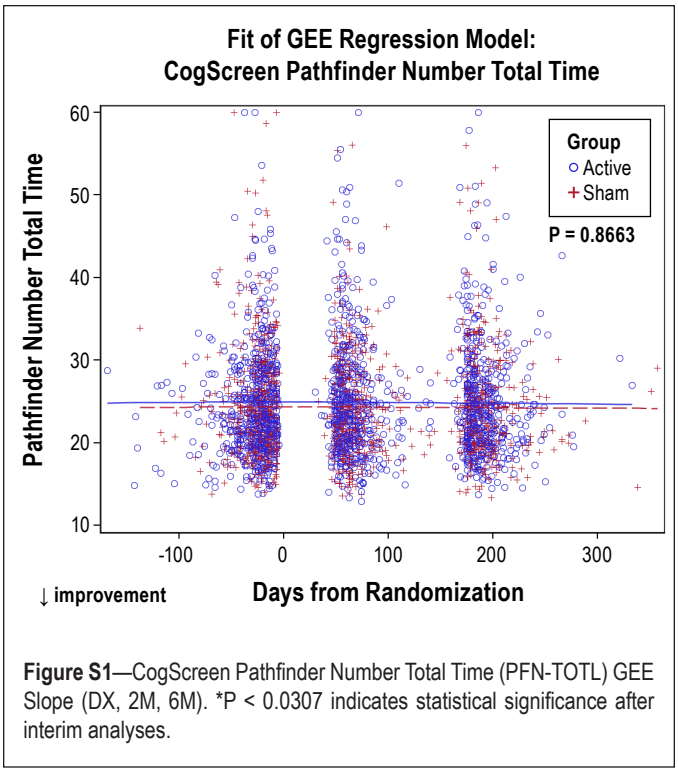

Figure S1—CogScreen Pathfinder Number Total Time (PFN-TOTL) GEE Slope (DX, 2M, 6M). *P < 0.0307 indicates statistical significance after interim analyses.

and SWMT-Behavioral Index: Mid-day (SWMT-BMD) were provided as change from baseline scores; so analyses of these two secondary outcomes excluded those participants who only had a baseline visit.

1D. Assessing Model Assumptions

For all GEE, GLMM and GLM analyses for both the primary and secondary neurocognitive analyses, we checked variance and link assumptions.[12] Residuals were plotted against fitted values and against model covariates to ensure that a given model was not misspecified. This procedure also provided a final check on data quality to confirm no outliers existed in these data. Influence diagnostics were performed as needed to assess model fit. In some cases, polynomial terms (up to cubic) for continuous covariates were added to improve fit.

For the primary neurocognitive parametric survival model fit to PFN-TOTL, model fit was assessed via simulating data from the fitted model and comparing observed data versus simulated values. For GLMM, we employed a random intercept for each participant and assessed if random effects were approximately normally distributed. GLMM fitting employed adaptive gaussian quadrature. For GLMM analyses of the secondary neurocognitive variables SAT-DIRUL, PFN-RTC, PVT-MSRT, and PVT-MDRT, data were centered and scaled to aid algorithm convergence.

1E. Treatment of Missing Data

GEE (Section 2A) assumed data were missing completely at random (MCAR). GLM and GLMM (Sections 2 and 3) assumed that data were missing at random (MAR). MAR and MCAR[13] are both types of missingness that assume that data are missing for reasons unrelated to the outcome that would have been observed.

We addressed the possibility that missingness depends upon a person's outcome through the retention-adjusted analyses (Section 8). Those analyses provide some evidence for informative missingness in that change from baseline for some of the primary neurocognitive outcomes are correlated with tendency to discontinue. However, we found that adjusting for this through the use of Heckman-type selection models did not change the primary efficacy findings.

No imputation was performed except for the Kolmogorov-Smirnov two-sample test analysis of adherence as outcome (Section 5A), where one version imputed missing values to zeros before calculating mean per person. Imputing missing to zero did not change findings from the Kolmogorov-Smirnov two-sample test.

In tables, figures, and text, reported sample sizes that don't sum to the entire randomized sample size of 1,098, this disparity was due to missing data in outcomes and/or covariates. See Section 6 on participant retention for additional details.

SECTION 2. RESULTS – PRIMARY NEUROCOGNITIVE DATA

2A. Per-Protocol GEE Regression Analyses for Primary Neurocognitive Outcomes

The per-Protocol GEE analysis for the PFN-TOTL variable is presented in Figure S1. This outcome was regressed on study arm, days since randomization, and interaction using GEE. We tested the hypothesis that the slope over time (DX, 2M, 6M) differed between study arms (P = 0.8663).

No GEE models testing slope over time were fit for BSRT-SR or SWMT-OMD (see Section 1B).

2B. GLM, GLMM, and Parametric Survival Analyses

In general, GLM were used for by-visit comparisons and GLMM were used to model repeated measures data. All means reported from GLM and GLMM are least-squares means centered at the mean values of all continuous covariates and at observed marginal frequencies of categorical variables.[14] Parametric survival analyses were conducted on PFN-TOTL for by-visit comparisons since these data were right censored at 60. Assessment of model assumptions was addressed in Section 1D and treatment of missing data was reviewed in Section 1E.

For unadjusted analyses a parametric survival analysis was run for PFN-TOTL and GLM analyses were run by visit (2M and 6M) for BSRT-SR and SWMT-OMD.

For covariate adjusted analyses, parametric GLMM analyses were run for repeated measurements (baseline, 2M and 6M) of PFN-TOTL and GLM analyses were run by visit (2M and 6M) for BSRT-SR and SWMT-OMD. PFN-TOTL data were reciprocal transformed for analysis and back-transformed for reporting. Covariate-adjusted analyses included the randomization factors. For all outcomes, covariates were OSA severity, sex, race, $\%TSTO_2 < 85$, age < 60 years, WASI verbal IQ and performance IQ. A pre-randomization baseline was also included as a covariate for BSRT-SR and PFN-TOTL, and months since randomization was also included for PFN-TOTL. Group by OSA severity interactions were included, allowing the difference in active vs. sham means to change among levels of OSA severity.

Table S2—Adjustment for multiple comparisons at final analyses

Visit	Test	Raw P Value	Adjusted P Value	Significant after Adjustment for Multiple Comparisons?
2M-CPAP Visit	PFN-TOTL	0.0860	0.4300	No
	BSRT-SR	0.5444	1.0000	No
	SWMT-OMD	0.0074*	0.0444*	No
6M-CPAP Visit	PFN-TOTL	0.2103	0.8412	No
	BSRT-SR	0.7569	1.0000	No
	SWMT-OMD	0.2254	0.8412	No

*P < 0.0307 indicates statistical significance for raw P values.

Table S3—SWMT-OMD mean (covariate adjusted using GLM) comparison of mean estimate of difference (6M – 2M) for participants randomized to active or sham CPAP

SWMT Overall Mid-day	OSA Severity	Mean Estimate (6M Minus 2M) (95% CI LB – UB)	P Value
Active CPAP (n = 455)	Mild	-0.043 (-0.249 – 0.162)	0.6790
	Moderate	0.003 (-0.142 – 0.147)	0.9726
	Severe	-0.056 (-0.167 – 0.054)	0.3168
Sham CPAP (n = 409)	Mild	0.043 (-0.181 – 0.267)	0.7047
	Moderate	-0.040 (-0.192 – 0.112)	0.6057
	Severe	-0.150 (-0.269 – -0.031)	0.0132*

*P < 0.05 indicates statistical significance.

2C. Adjustment for Multiple Comparisons at Final Analyses after Multiple Interim Analyses

For the purpose of adjusting for multiplicity, the tests run by primary neurocognitive outcome and visit (2M and 6M) without adjustment for covariates were utilized. Adjustments for multiple comparisons were limited to the 2M and 6M visits for the three primary neurocognitive outcomes, for those analyses without adjustment for covariates and without stratification. These six primary neurocognitive hypothesis tests are presented in Table S2 with and without adjustment for multiple comparisons at final. O'Brien-Fleming spending across three interim analyses left 3.07% Type-I Error for the final analysis.[15] Correction for multiple comparisons at final analyses employed sequential Bonferroni adjustment.[16]

None of the six primary neurocognitive analyses were significant after these adjustments were made.

2D. GLM by OSA Severity between 2M and 6M Visits within Arms

GLM analyses were run for SWMT-OMD with covariates to determine whether there was a significant difference in the SWMT-OMD at 2M vs. 6M (6M Minus 2M) when compared within each OSA severity level and study arm (Table S3). In addition to study arm and OSA severity, covariates were sex, race, $\%TSTO_2 < 85$, age < 60 years, WASI Verbal IQ, and WASI Performance IQ. Confidence interval lower bounds (CI LB) and upper bounds (UB) are provided for each estimated mean.

2E. GLM with $\%TSTO_2 < 85$ Quartiles

GLM analyses stratified by quartiles of $\%TSTO_2 < 85$, study arm, and visit are presented in Table S4 (without adjustment for any other covariates). Estimates of the means for the neurocognitive (NC) outcomes are compared between the lower three $\%TSTO_2 < 85\%$ quartiles vs. the upper quartile (most hypoxic) within visits and study arms. Estimates are least squares means.[14] Quartiles were estimated by first pooling the data across both study arms. Quartile analyses were designed based on work by Quan and colleagues.[17]

The results for the SWMT-OMD are discussed in the main text; however, the significant findings for the PFN-TOTL and BSRT-SR (shown below) are not described since the primary analyses did not show differences between arms across the visits.

2F. Neurocognitive Change Scores for Participants with Baseline ESS > 10 or MWT ≤ 14.5

These sub-analyses were conducted to determine the association of clinically significant subjective and objective sleepiness on our primary outcomes. Two-sample t-tests were performed for participants with a baseline Epworth Sleepiness Scale-Total Score (ESS-TS) > 10 (subjectively sleepy participants; Table S5); this ESS-TS score is indicative of clinically significant sleepiness. Separate analyses were also run for participants with a baseline MWT-Mean Sleep Latency (MWT-MSL) score ≤ 14.5 (objectively sleepy participants). This threshold was selected because it is 1 SD below the MWT-MSL for a population of normal individuals tested for a 20-minute MWT trial duration.[18] SWMT-OMD is already formulated as a change-from-baseline score for the 2M and 6M visits. For BSRT-DR and PFN-TOTL, change-from-baseline scores were calculated for both 2M and 6M (2M Minus DX and 6M Minus DX, respectively).

2G. Correlation Coefficients for Participants with Baseline MWT ≤ 14.5

Analyses were run for a subgroup of objectively sleepy participants. To evaluate the correlation of the change from baseline MWT-MSL score and the change from baseline primary neurocognitive score at both 2M and 6M by study arm, Spearman correlation coefficients and P values were obtained (Table S6).

SECTION 3. RESULTS – SECONDARY NEUROCOGNITIVE DATA

3A. Selection of 12 Secondary Neurocognitive Outcomes for Dimension Reduction

Based on a recommendation by the APPLES Data and Safety Monitoring Board (DSMB), the APPLES Team utilized an independent team of neurocognitive experts to assist them in creating an *a priori S*econdary Neurocognitive Analysis Plan. Following multiple conference calls and the dissemination of materials related to the APPLES neurocognitive test battery, including the psychometric properties (normative data, test-retest reliability, and trends including potential practice effects) for each outcome, a summary of the literature, and the APPLES Methods Paper,[7] the team of neurocognitive experts provided specific recommendations to the APPLES Team.

Table S4—Primary neurocognitive outcomes (adjusted for oxygen saturation using GLM) comparison of mean estimates between % $TSTO_2$ < 85 quartiles by study arm and visit

	CPAP Study Arm	NC Mean Estimate (95% CI LB – UB) Lower 3 Quartiles for %$TSTO_2$ < 85	NC Mean Estimate (95% CI LB – UB) Upper Quartile for %$TSTO_2$ < 85	P Value
CogScreen Pathfinder Number Total Time				
DX	Active CPAP	24.05 (23.46 – 24.64)	26.80 (25.59 – 28.02)	< 0.0001*
	Sham CPAP	24.49 (23.83 – 25.15)	24.36 (23.36 – 25.36)	0.9572
2M	Active CPAP	24.77 (23.99 – 25.54)	26.78 (25.36 – 28.20)	0.0043*
	Sham CPAP	24.41 (23.70 – 25.12)	23.70 (22.67 – 24.74)	0.5405
6M	Active CPAP	24.19 (23.52 – 24.87)	27.40 (25.85 – 28.95)	< 0.0001*
	Sham CPAP	24.40 (23.67 – 25.13)	24.35 (23.01 – 25.69)	0.8879
BSRT Sum Recall				
DX	Active CPAP	49.93 (49.05 – 50.81)	49.07 (47.52 – 50.62)	0.3357
	Sham CPAP	50.12 (49.23 – 51.02)	49.09 (47.60 – 50.58)	0.2545
2M	Active CPAP	52.87 (52.00 – 53.75)	50.68 (48.92 – 52.43)	0.0208*
	Sham CPAP	52.10 (51.07 – 53.13)	51.52 (49.90 – 53.14)	0.5561
6M	Active CPAP	54.47 (53.56 – 55.38)	52.95 (51.20 – 54.69)	0.1108
	Sham CPAP	54.31 (53.30 – 55.32)	54.18 (52.63 – 55.73)	0.8928
SWMT Overall Mid-day				
2M	Active CPAP	0.003 (-0.061 – 0.066)	0.132 (0.023 – 0.242)	0.0448*
	Sham CPAP	-0.079 (-0.146 – -0.013)	-0.057 (-0.173 – 0.059)	0.7411
6M	Active CPAP	0.070 (0.001 – 0.139)	0.079 (-0.040– 0.198)	0.9010
	Sham CPAP	0.005 (-0.069 – 0.078)	0.058 (-0.070 – 0.187)	0.4785

*P < 0.05 indicates statistical significance.

Table S5—For participants with a baseline ESS-TS > 10 or MWT-MSL ≤ 14.5, comparison of neurocognitive change from baseline scores between study arms by visit

		Active CPAP Mean (SD) Change-from-Baseline	Sham CPAP Mean (SD) Change-from-Baseline	P Value
For Participants with a Baseline ESS Total Score > 10				
Pathfinder Number Total Time	2M (Active n = 198; Sham n = 189)	0.20 (5.70)	-0.55 (4.33)	0.1055
	6M (Active n = 194; Sham n = 188)	-0.12 (5.67)	0.39 (6.02)	0.7511
BSRT Sum Recall	2M (Active n = 199; Sham n = 190)	2.31 (6.94)	2.58 (7.08)	0.7015
	6M (Active n = 196; Sham n = 187)	4.38 (6.92)	5.25 (6.75)	0.2149
SWMT Overall Mid-day	2M (Active n = 195; Sham n = 179)	0.074 (0.626)	-0.031 (0.595)	0.0957
	6M (Active n = 187; Sham n = 175)	0.150 (0.661)	0.014 (0.613)	0.0433*
For Participants with a Baseline MWT Mean Sleep Latency ≤ 14.5				
Pathfinder Number Total Time	2M (Active n = 102; Sham n = 88)	-0.25 (6.52)	-0.72(4.46)	0.5271
	6M (Active n = 103; Sham n = 84)	-0.32 (5.95)	0.10 (5.10)	0.2952
BSRT Sum Recall	2M (Active n = 102; Sham n = 88)	3.49 (7.65)	2.03 (6.64)	0.1660
	6M (Active n = 103; Sham n = 83)	4.56 (7.24)	4.46 (7.33)	0.7559
SWMT Overall Mid-day	2M (Active n = 99; Sham n = 83)	0.083 (0.658)	0.033 (0.694)	0.6193
	6M (Active n = 97; Sham n = 79)	0.089 (0.713)	0.099 (0.720)	0.9331

*P < 0.05 indicates statistical significance.

Table S6—For participants with MWT-MSL ≤ 14.5, correlation between change in MWT-MSL vs. change in primary neurocognitive outcome by visit and study arm

	Active CPAP	P Value	Sham CPAP	P Value
CogScreen Pathfinder Number – Total Time				
2M: Spearman Correlation Coefficient (Δ MWT-MSL vs. Δ PFN-TOTL)	0.0322, n = 101	0.7494	0.1197, n = 85	0.2754
6M: Spearman Correlation Coefficient (Δ MWT-MSL vs. Δ PFN-TOTL)	-0.1629, n = 101	0.1035	0.1239, n = 83	0.2643
BSRT – Sum Recall				
2M: Spearman Correlation Coefficient (Δ MWT-MSL vs. Δ BSRT-SR)	-0.0894, n = 101	0.3740	0.0109, n = 85	0.9214
6M: Spearman Correlation Coefficient (Δ MWT-MSL vs. Δ BSRT-SR)	-0.1356, n = 101	0.1764	0.1108, n = 82	0.3216
SWMT – Mid-day Overall Index				
2M: Spearman Correlation Coefficient (Δ MWT-MSL vs. Δ SWMT-OMD)	0.2084, n = 98	0.0395*	0.0774, n = 80	0.4948
6M: Spearman Correlation Coefficient (Δ MWT-MSL vs. Δ SWMT-OMD)	0.1598, n = 95	0.1219	0.1015, n = 78	0.3766

*P < 0.05 indicates statistical significance.

Twelve variables were identified across the three neurocognitive domains of attention and psychomotor function (A/P), learning and memory (L/M), and executive and frontal-lobe function (E/F): 1) Psychomotor Vigilance Task-Median Reaction Time (PVT-MDRT); 2) PVT-Mean Slowest 10% of Reaction Times (PVT-MSRT); 3) PFN-Reaction Time (PFN-RTC); 4) CogScreen Symbol Digit Coding-Correct Responses (SDC-CORR); 5) CogScreen Shifting Attention Task Instruction Condition-Thruput (SAT-INPUT); 6) BSRT-Summary Score (BSRT-MSUM): Mean of BSRT-SR, Long-term Storage (LTS), Long-term Retrieval (LTR), and Consistent Long-term Retrieval (CLTR); 7) BSRT Delayed Recall-Total Recall (BSRTDR-TR); 8) Paced Auditory Serial Addition Test-total Correct (PASAT-TC); 9) CogScreen Shifting Attention Task Discovery Condition-Rule Shifts Completed (SAT-DIRUL); 10) CogScreen Pathfinder Combined-Total Time (PFC-TOTL); 11) SWMT-Activation Index: Mid-day (SWMT-AMD); and 12) SWMT-Behavioral Index: Mid-day (SWMT-BMD) (Table S7).

The plan specified that these 12 variables be shortened to a short list of approximately 4-6 variables which best preserve the information structure of all 12 using a statistical dimensionality reduction method.

Table S7—Twelve secondary neurocognitive variables for independent components analysis

Attention and Psychomotor Function	PVT-Median Reaction Time
	PVT-Mean Slowest 10% of Reaction Times
	PN-Reaction Time
	SDC-Correct Responses
	SAT-Instruction Condition -Thruput
Learning and Memory	BSRT Summary Score: Mean of Sum Recall, LTS, LTR, and CLTR
	BSRT Delayed Recall-Total Recall
Executive and Frontal-Lobe Function	PASAT-Total Correct
	SAT-Discovery Condition - Rule Shifts Completed
	PFC-Total Time
	SWMT-Activation Index Mid-day
	SWMT-Behavioral Index Mid-day

3B. Selection of a Statistical Dimension Reduction Method

The APPLES *a priori* Secondary Neurocognitive Analysis Plan specified that the method of Krzanowski[19] be used to reduce our 12 secondary neurocognitive outcomes to a set of 4 to 6. Upon beginning that work using the follow-on paper by Wang and Gehan[20] a subtle, but important math error was detected in the published method. This error was traced back to an error made in the first paper in the series.[21] The APPLES Data Coordinating Center (DCC) was reluctant to use a method that was specified incorrectly in the literature and for which no proposed correction has undergone formal peer review.

Based on this finding, Independent Component Analysis (ICA) was employed instead of Krzanowski's method. ICA has the "goal of decomposing measured signals or variables into a set of underlying variables,"[22] which is exactly what was required per the APPLES Secondary Neurocognitive Analysis Plan. The decision to change the method for dimension reduction was approved by the SC.

We selected those secondary neurocognitive outcomes that met the following criterion: If an ICA component was very highly correlated with one and only one of the original 12 outcomes, and had low correlation with all other outcomes, evidence suggested that outcome provided a separable source of non-redundant information.

3C. Covariate Adjusted Regression Models for Secondary Neurocognitive Outcomes

Covariate adjusted regression models were fit for the 7 secondary neurocognitive outcomes identified by ICA. GLMM were utilized to account for the repeated measures for CogScreen (PFN and SAT-D) and PVT outcomes (DX, 2M, 6M), while GLM was run by visit (2M and 6M) for BSRT and SWMT outcomes (Table S8). The covariates included in this analysis were those designated in the secondary analysis plan as being the most likely to explain variation in these outcomes. Covariate-adjusted analyses included the randomization factors. In addition to study arm, covariates were: OSA severity, sex, race, $\%TSTO_2 < 85$, age < 60 years, WASI Verbal IQ, and WASI Performance IQ. A pre-randomization baseline was also

Table S8—Comparisons of means between participants randomized to active vs. sham CPAP on secondary neurocognitive outcomes: estimated means from regression models with covariate adjustment

		Active CPAP Mean Estimate (95% CI LB – UB)	Sham CPAP Mean Estimate (95% CI LB – UB)	P Value
Pathfinder Number – Reaction Time				
2M (n = 850)	Mild OSA	0.811 (0.785 – 0.839)	0.801 (0.774 – 0.830)	0.5606
	Moderate OSA	0.831 (0.811 – 0.852)	0.825 (0.806 – 0.845)	0.6487
	Severe OSA	0.818 (0.802 – 0.834)	0.812 (0.797 – 0.828)	0.5667
6M (n = 822)	Mild OSA	0.811 (0.784 – 0.839)	0.795 (0.767 – 0.826)	0.3972
	Moderate OSA	0.830 (0.808 – 0.853)	0.819 (0.799 – 0.841)	0.3973
	Severe OSA	0.817 (0.800 – 0.836)	0.806 (0.790 – 0.823)	0.3055
Shifting Attention Test Discovery Condition – Number of Rule Changes (Dichotomized)				
2M (n = 846)	Mild OSA	0.931 (0.885 – 0.977)	0.929 (0.873 – 0.985)	0.9518
	Moderate OSA	0.936 (0.904 – 0.968)	0.951 (0.924 – 0.979)	0.4108
	Severe OSA	0.952 (0.931 – 0.972)	0.942 (0.918 – 0.967)	0.4528
6M (n = 813)	Mild OSA	0.897 (0.827 – 0.966)	0.907 (0.832 – 0.982)	0.8391
	Moderate OSA	0.903 (0.853 – 0.953)	0.935 (0.896 – 0.975)	0.2771
	Severe OSA	0.927 (0.894 – 0.959)	0.924 (0.888 – 0.960)	0.8961
BSRT Delayed Recall – Total Recall				
2M (n = 870)	Mild OSA	8.54 (7.99 – 9.10)	8.20 (7.53 – 8.87)	0.4262
	Moderate OSA	8.49 (8.13 – 8.85)	8.22 (7.82 – 8.62)	0.3161
	Severe OSA	8.48 (8.20 – 8.76)	8.21 (7.92 – 8.51)	0.1835
6M (n = 838)	Mild OSA	9.01 (8.47 – 9.54)	9.44 (8.92 – 9.97)	0.2462
	Moderate OSA	8.56 (8.14 – 8.98)	8.91 (8.54 – 9.28)	0.2069
	Severe OSA	8.87 (8.58 – 9.16)	8.75 (8.48 – 9.01)	0.5235
SWMT – Mid-day Behavioral Index				
2M (n = 843)	Mild OSA	0.180 (0.006 – 0.355)	0.104 (-0.074 – 0.283)	0.5419
	Moderate OSA	0.137 (0.035 – 0.238)	0.126 (0.007 – 0.245)	0.8900
	Severe OSA	0.205 (0.117 – 0.294)	-0.011 (-0.128 – 0.106)	0.0031*
6M (n = 815)	Mild OSA	0.143 (-0.072 – 0.357)	0.116 (-0.123 – 0.356)	0.8703
	Moderate OSA	0.194 (0.062 – 0.325)	0.314 (0.191 – 0.437)	0.1838
	Severe OSA	0.321 (0.212 – 0.430)	0.173 (0.052 – 0.295)	0.0739
SWMT – Mid-day Activation Index				
2M (n = 815)	Mild OSA	-0.050 (-0.268 – 0.169)	0.317 (0.031 – 0.603)	0.0450*
	Moderate OSA	0.262 (0.084 – 0.440)	0.170 (0.006 – 0.334)	0.4512
	Severe OSA	-0.003 (-0.109 – 0.103)	0.033 (-0.093 – 0.159)	0.6672
6M (n = 787)	Mild OSA	0.157 (-0.089 – 0.403)	0.118 (-0.117 – 0.353)	0.8197
	Moderate OSA	0.016 (-0.131 – 0.162)	0.014 (-0.188 – 0.216)	0.9890
	Severe OSA	0.058 (-0.068 – 0.185)	0.123 (-0.016 – 0.262)	0.5029
PVT – Median Reaction Time				
2M (n = 851)	Mild OSA	245.31 (230.94 – 260.58)	253.89 (238.02 – 270.82)	0.3699
	Moderate OSA	248.68 (237.96 – 259.89)	248.94 (237.99 – 260.40)	0.9673
	Severe OSA	243.25 (235.36 – 251.41)	247.55 (238.79 – 256.62)	0.3426
6M (n = 820)	Mild OSA	245.23 (230.20 – 261.23)	254.05 (237.26 – 272.03)	0.3901
	Moderate OSA	248.60 (236.74 – 261.04)	249.10 (236.95 – 261.86)	0.9464
	Severe OSA	243.17 (234.05 – 252.64)	247.70 (237.46 – 258.38)	0.4372
PVT – Mean Slowest 10% of Reaction Times				
2M (n = 851)	Mild OSA	403.00 (375.99 – 431.95)	402.32 (376.77 – 429.62)	0.9656
	Moderate OSA	412.44 (390.91 – 435.16)	407.84 (387.67 – 429.06)	0.6765
	Severe OSA	400.57 (384.51 – 417.30)	406.11 (387.05 – 426.07)	0.5288
6M (n = 820)	Mild OSA	396.87 (370.79 – 424.79)	401.28 (375.51 – 428.82)	0.7807
	Moderate OSA	406.17 (383.96 – 429.66)	406.78 (385.63 – 429.09)	0.9603
	Severe OSA	394.48 (377.66 – 412.05)	405.04 (384.90 – 426.24)	0.3075

*P < 0.05 indicates statistical significance.

Table S9—Correlation between change in ESS with CPAP adherence by visit and study arm

	Active CPAP	P Value	Sham CPAP	P Value
Change in ESS Total Score – (2M Minus DX)				
Spearman Correlation Coefficient (Δ ESS-TS vs. CPAP Adherence)	-0.20865	< 0.0001*	-0.02394	0.6285
Change in ESS Total Score – (6M Minus DX)				
Spearman Correlation Coefficient (Δ ESS-TS vs. CPAP Adherence)	-0.18161	0.0003*	-0.08282	0.1137

*P < 0.05 indicates statistical significance.

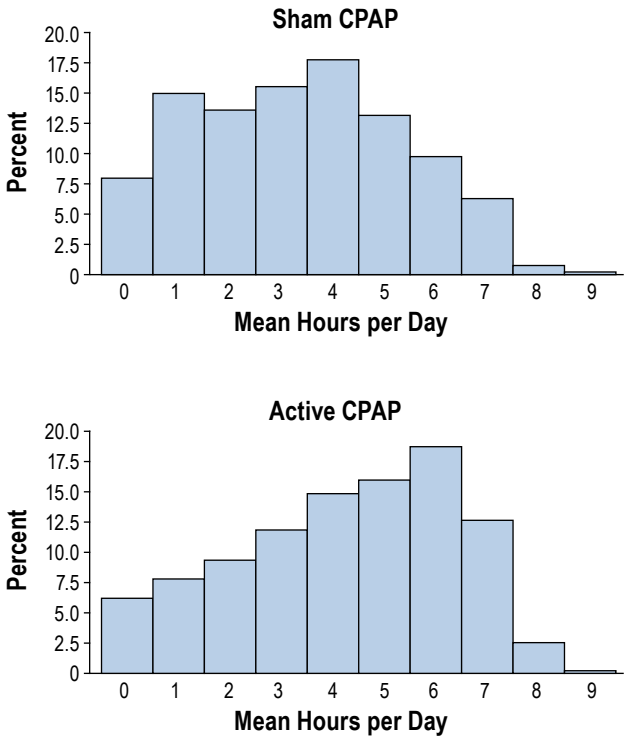

Figure S2—Frequency histograms of mean hours of nightly CPAP per participant (entire follow-up duration) between study arms. Active CPAP mean = 4.2 hours. Sham CPAP mean = 3.4 hours. P < 0.001.

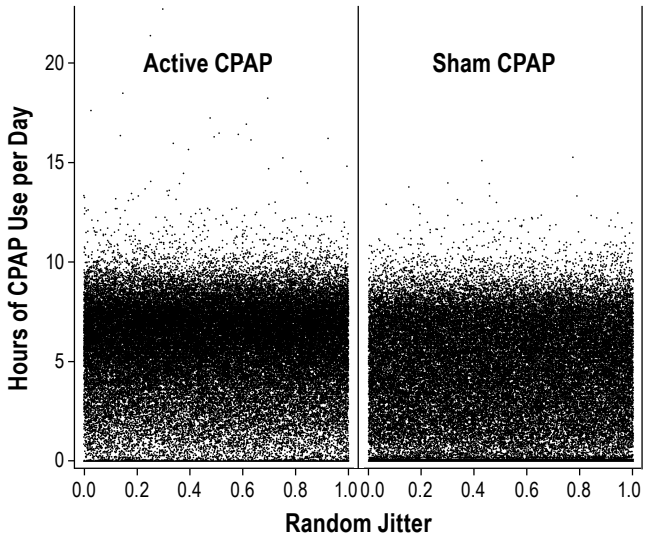

Figure S3—Plot of 24-hour CPAP usage (entire follow-up duration) per participant by study arm. Active CPAP usage = 76,770 days. Sham CPAP usage = 69,807 days.

included as a covariate for the BSRT and PFN analyses. Months since randomization was also included as a covariate for the repeated measures analyses for PFN and PVT. Group by OSA severity interactions were included in the regression models, allowing a difference in active vs. sham means for each level of OSA severity. SAT-D was formulated as a dichotomized variable (≤ 2 vs. ≥ 3) based on a 5th percentile cut-off for studies performed for pilots, based on recommendations from the developer of this test. PFN and PVT data were reciprocal transformed for analysis and back-transformed for reporting. Estimates from the models are provided for each study arm, visit, and OSA severity level.

SECTION 4. RESULTS – SECONDARY SLEEPINESS DATA

4A. Correlation Coefficients for Change in ESS-TS vs. CPAP Adherence

Spearman Correlation Coefficients were obtained to evaluate the correlation of the change in ESS-TS from baseline (for 2M and 6M) with CPAP adherence (Table S9). Mean hours of adherence for the 2 months prior to the neurocognitive visit was used as the CPAP adherence variable. The number of days on the SmartCard was the denominator for this variable.

SECTION 5. RESULTS – CPAP ADHERENCE

5A. Mean Hours of Nightly Usage – Entire Study Duration

Figure S2 presents the frequency distribution of mean hours of nightly CPAP usage per participant by study arm. All of the CPAP adherence data for the duration of a patient's follow-up were used to calculate his/her mean. The P value is for the comparison of distributions between arms via a Kolmogorov-Smirnov two-sample test.[23] Figure S3 plots the 24-hour CPAP usage values by study arm for the entire follow-up duration for the 1,098 randomized participants. The horizontal axis is a random jitter (i.e., each observation was paired with a number from a uniform distribution on the interval 0 to 1) of these data. In the active CPAP arm, the greatest frequency of usages is between 5 and 7 hours. Also notice the higher density of zero and near-zero usage for sham. For all adherence analyses presented in this section, missing data were assumed to be non-informative. We allowed for missingness to be informative in an analysis not

Table S10—Comparison of mean hours of nightly CPAP between study arms (for various durations prior to the 2M and 6M post-CPAP visits)

	CPAP Study Arm	Sample Size	Mean Hours of Nightly Adherence (SD)	P Value
2M Post-CPAP Visit				
Night prior to Visit	Active	372	5.45 (2.57)	< 0.0001*
	Sham	337	4.59 (2.73)	
Week prior to Visit	Active	394	5.11 (2.14)	< 0.0001*
	Sham	366	4.10 (2.27)	
Month prior to Visit	Active	425	4.78 (2.09)	< 0.0001*
	Sham	399	3.80 (2.16)	
2 Months prior to Visit	Active	436	4.75 (2.02)	< 0.0001*
	Sham	412	3.97 (2.05)	
6M Post-CPAP Visit				
Night prior to Visit	Active	305	5.77 (2.29)	< 0.0001*
	Sham	285	4.34 (2.79)	
Week prior to Visit	Active	351	5.11 (2.15)	< 0.0001*
	Sham	320	4.06 (2.36)	
Month prior to Visit	Active	387	4.73 (2.13)	< 0.0001*
	Sham	351	3.54 (2.26)	
2 Months prior to Visit	Active	396	4.68 (2.10)	< 0.0001*
	Sham	366	3.40 (2.20)	

*P < 0.05 indicates statistical significance.

shown (missing data imputed to zero usage), but this did not change the findings.

5B. Mean Hours of Nightly Usage – Various Durations Prior to the 2M and 6M Visits

Table S10 compares mean hours of nightly CPAP usage between the study arms for various durations (1 night, 1 week, 1 month, and 2 months) prior to the 2M- and 6M-CPAP Visits using permutation testing. Four different durations were utilized to thoroughly describe CPAP adherence prior to the neurocognitive visits and to select the most informative variable for CPAP adherence-adjusted analyses.

Mean hours of adherence were longest for the night prior to a neurocognitive visit, decreasing as the duration was lengthened to 1 week and 1 month prior to a visit. Mean hours of nightly adherence seemed to stabilize over 1 and 2 month durations.

5C. ≥ 4 hours for > 70% of the Time – Various Durations Prior to the 2M and 6M Visits

A chi square analysis was run to compare between study arms the number of participants with ≥ 4 hours of CPAP use for > 70% of the nights for each of the given durations (1 night, 1 week, 1 month, and 2 months) prior to the 2M- and 6M-CPAP Visits (Table S11). The percentages are the number of participants divided by the sample size for each row.

Four different durations were utilized to thoroughly describe CPAP adherence prior to the neurocognitive visits. The number of participants who met the adherence criterion was the greatest for the night prior to a neurocognitive visit, decreasing as the duration was lengthened to 1 week, 1 month, and 2 months prior to a visit.

Table S11—Comparison of the number of participants with ≥ 4 hours of CPAP use for > 70% of the duration (for various durations prior to the 2m and 6m post-CPAP visits) between study arms

	CPAP Study Arm	Sample Size	Number of Participants with ≥ 4 h for > 70% of Duration (%)	P Value
2M Post-CPAP Visit				
Night prior to Visit	Active	464	280 (60.34)	0.0002*
	Sham	431	207 (48.03)	
Week prior to Visit	Active	464	257 (55.39)	< 0.0001*
	Sham	431	165 (38.28)	
Month prior to Visit	Active	464	212 (45.69)	< 0.0001*
	Sham	431	115 (26.68)	
2 Months prior to Visit	Active	464	184 (39.66)	< 0.0001*
	Sham	431	108 (25.06)	
6M Post-CPAP Visit				
Night prior to Visit	Active	443	249 (56.21)	< 0.0001*
	Sham	402	160 (39.80)	
Week prior to Visit	Active	443	219 (49.44)	< 0.0001*
	Sham	402	133 (33.08)	
Month prior to Visit	Active	443	174 (39.28)	< 0.0001*
	Sham	402	89 (22.14)	
2 Months prior to Visit	Active	443	188 (42.44)	< 0.0001*
	Sham	402	90 (22.39)	

*P < 0.05 indicates statistical significance.

5D. Participant Treatment Group Guesses by Arm

Prior to unblinding participants to their assigned treatment group condition, participants were asked to guess to which study arm they believed they had been assigned (Figure S4). A κ coefficient was used to estimate the degree of chance-adjusted agreement between participant guesses and arm assignment. A total of 69.67% of sham CPAP participants correctly guessed their treatment assignment vs. 55.28% of active CPAP participants ($\kappa = 0.25$, $P < 0.0001$). A κ coefficient of 0.25 is suggestive of relatively poor agreement.[24]

SECTION 6. RESULTS – PARTICIPANT RETENTION

6A. Life-Table Retention Curves

Figure S5 presents results of a life-table analysis of retention. Retention curves are provided by study arm. Analysis employed 25-day intervals and retention was measured from the time of the Diagnostic Visit to the last neurocognitive visit date. The P value presented is for the log-rank test comparing the retention curves between study arms.

SECTION 7. RESULTS – ADJUSTING PRIMARY NEUROCOGNITIVE ANALYSES FOR CPAP ADHERENCE

7A. Varied Adherence

Participants were randomly assigned to the sham *vs*. active CPAP conditions. Each participant was then encouraged to adhere to his/her assigned treatment. According to the APPLES Protocol:

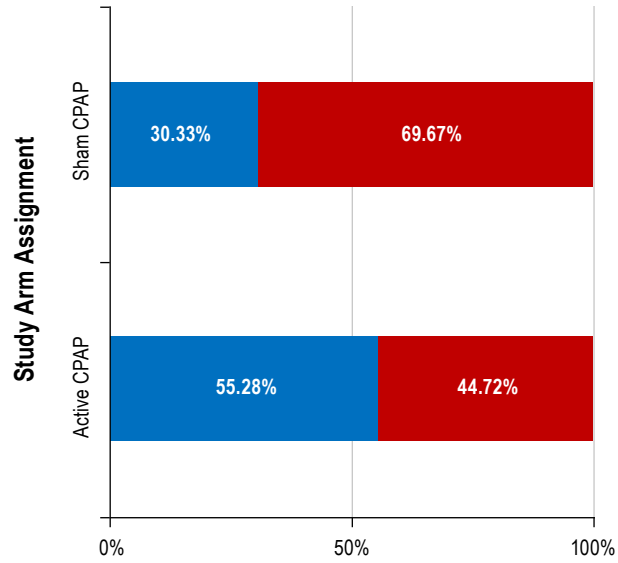

Figure S4—Estimation of chance-adjusted agreement between participant guesses and study arm assignment.

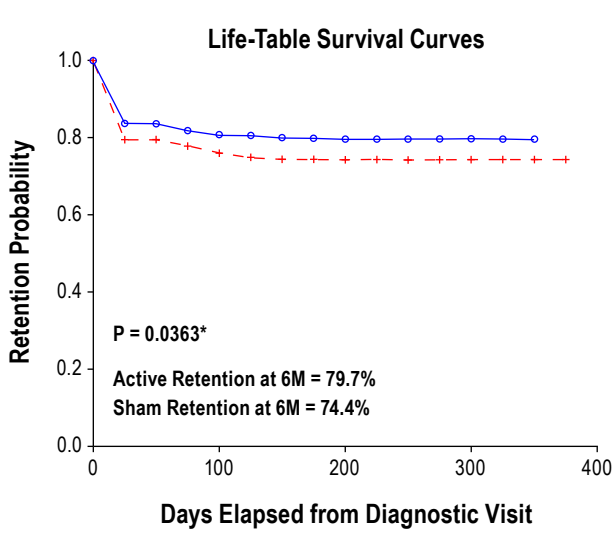

Figure S5—Comparison of participant retention between study arms. *P < 0.05 indicates statistical significance.

"Each APPLES participant will be followed closely by the assigned staff member. All compliance issues will be brought to the attention of the CC Coordinator. It may be necessary for the CC Coordinator to contact a non-blinded study physician in the event of a difficult CPAP compliance problem."

Despite these efforts, substantial variation in adherence was observed in both study arms (Figure S6). Reduction in adherence was most pronounced for participants in the sham arm by the 6M visit.

7B. Adherent Subgroup Analysis

Consider a subpopulation restricted to just those "adherent" individuals who use their assigned device for at least 4 hours per night on average in the two months prior to the visit (2M and 6M). An analysis comparing baseline variables for the group of adherent individuals vs. non-adherent individuals at both the 2M and 6M time points revealed significant differences in a number of baseline variables (Tables S12 and S13). Adherent individuals appear to be older on average (2M 4.8 yrs higher, $P < 0.0001$; 6M 5.4 yrs higher, $P < 0.0001$), are more likely to be White (2M/6M $P < 0.0001$) and married (2M $P = 0.0474$, 6M $P = 0.0161$), and have higher WASI IQ scores on average (e.g., IQFull4WASI: 2M 5.1 points higher, $P < 0.0001$; 6M 4.5 points higher, $P < 0.0001$). Some differences in baseline polysomnographic variables also emerged. On average, the group of CPAP-adherent individuals at 2M and 6M have a lower sleep efficiency percentage at baseline (2M 1.9% lower, $P = 0.0296$; 6M 3.8% lower, $P < 0.0001$); and at 6M, adherers had a shorter total sleep time (15 minutes lower, $P = 0.0011$), longer sleep latency (4.2 minutes higher, $P = 0.0063$), longer REM latency (5.4 minutes higher, $P = 0.0221$), and a lower percentage of stage 3 sleep (0.67% lower, $P = 0.0424$).

In the adherent subpopulation, means of the baseline variables of Table 1 in the manuscript and means of the 1NC outcomes were compared between the sham and active conditions, by post-randomization visit (Table S14). Mean scores are approximately 2.5 units lower at 6M ($P = 0.0453$) on the IQ Verbal WASI for those on active compared to those on sham.

7C. Dose Response

The APPLES SC wished to know if variation in adherence could be responsible for variation in the primary neurocognitive (1NC) outcomes. In particular it was thought that a dose-response relationship may exist between adherence and 1NC outcomes. As demonstrated in section 7b, a potential difficulty with such an assessment is that each participant can self-select his/her level of adherence. Self-selection opens the possibility that participants who adhere more are different on other traits from those who adhere less (Table S12). If some of these traits drive variation in adherence and in neurocognitive performance, then *confounding* may be present. Namely, a detected association between adherence and a 1NC outcome may actually be due in whole or in part to one or more other factors—confounders. Unless analysis adjusts for any such confounders effectively, then variation in a 1NC outcome could be wrongly attributed to variation in CPAP adherence.

7D. Search for Confounders

Various methods have been developed in the statistical literature for adherence adjustment in the presence of possible confounders. Given that CPAP adherence was captured on a continuous scale in APPLES, the generalized propensity method of Imbens[25,26] seems well-suited for this purpose. This method allows construction of a dose-response curve between adherence to the active condition and a 1NC outcome within each study arm while balancing on observed potential baseline confounders. Mean response is then compared between study arms at points along these curves to assess the effects of sham *vs*. active CPAP as a function of dose because adjustment for

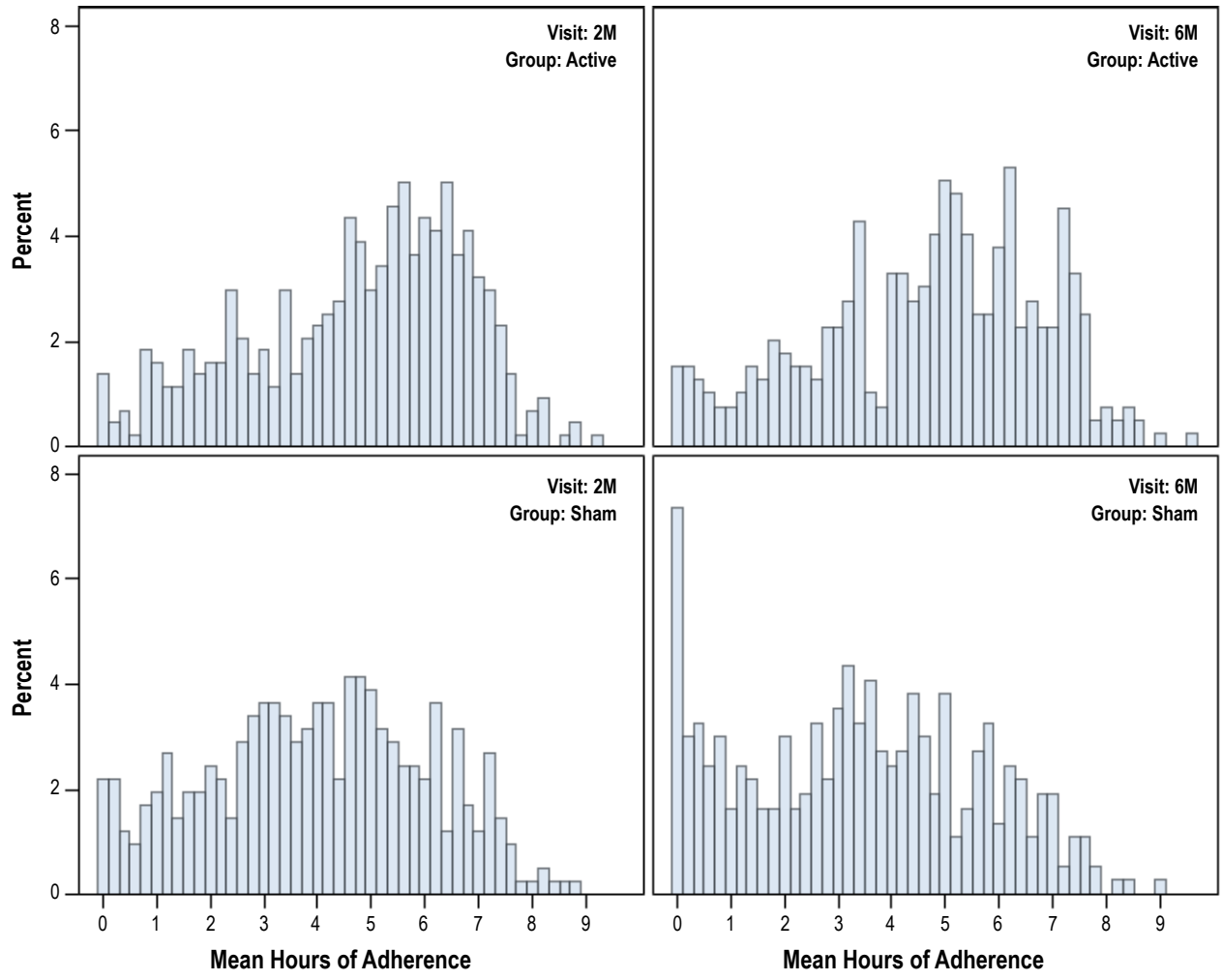

Figure S6—Frequency histograms, by visit and study arm – mean hours of adherence for 2 months prior to the visit.

the same set of confounders has been performed in both arms and randomization should ensure that treatment assignment is independent of a person's baseline features.

Before proceeding to that modeling exercise, a list of possible confounders was first identified. APPLES' investigators compiled a comprehensive list of possible confounders that were captured in the database (i.e., variables possibly causally related to both adherence and 1NC outcome). These 102 variables are listed in Table S15. Development of this list erred on the side of including too many rather than too few candidates to avoid missing any true confounders that had been observed.

7E. Adherence Adjustment

The generalized propensity score method was applied, closely following section 7.4 of Hirano and Imbens.[26] Estimation of generalized propensity scores for adherence to the active condition employed the variables of Table S15, was performed separately for each visit (2M and 6M), and used the sample from the active arm, with variable selection via the lasso and coefficient estimation via least squares. The resultant estimated dose-response curves are shown in Figure S7.

Difference between active and sham in mean dose-response was compared at nine levels of adherence (0, 1,…8 hours), as summarized in Table S16. Table S16 reveals a difference in means between study arms at the six-month visit for Overall Midday at 3 and 4 hours of adherence. The fact that differences are detected only at intermediate levels of adherence may be in part a statistical artifact, in that error in estimates of a fitted mean are wider toward the lower and upper ends of the extent of the regressor,[31] which here is adherence. Adherence was employed as a regressor in the second of three stages of the method of Hirano and Imbens.[26]

There is the possibility that the difference detected for SWMT Overall Midday was due to sham worsening. Table 2 in the manuscript provides estimates at 2M for active mean [CI] of 0.035 [-0.019 to 0.090] and for sham mean [CI] of -0.074 [-0.133 to -0.015]), where the confidence bounds on the mean for sham indicate a significant decline from baseline for sham. Also from Table 2 in the manuscript, estimates at 6M for active mean [CI] are 0.072 [0.012 to 0.132] and for sham mean [CI] are 0.018 [-0.046 to 0.082]. Adherence dropped strongly between 2M and 6M for sham. To enhance

Table S12—Comparison of interval-scale and ratio-scale baseline variables' means and 1NC outcomes' means between adherent and non-adherent subpopulations, by post-randomization visit

Visit	Outcome	Adherent Mean Minus Nonadherent Mean	Std. Error	P value
2M	Age	4.7887	0.82379	< 0.0001
2M	BMI	-0.4435	0.4976	0.3730
2M	TSTPSG	-8.3809	4.5397	0.0652
2M	SleepEffPSG	-1.9118	0.8774	0.0296
2M	SOfLOPSG	2.3961	1.4612	0.1014
2M	PerTSTS3PSG	-0.3231	0.3246	0.3199
2M	PerTSTS4PSG	-0.0032	0.1384	0.9812
2M	PerTSTREMPSG	-0.7182	0.4840	0.1382
2M	RDITSTPSG	1.6371	1.7464	0.3488
2M	MinimumSPO_2QC	0.0657	0.5808	0.9100
2M	PerTSTS1PSG	-0.5054	1.0147	0.6186
2M	PerTSTS2PSG	1.5757	0.9661	0.1033
2M	PerSPO_2lt85TST	-0.2984	0.4578	0.5147
2M	HighestGradeHP	0.3259	0.1785	0.0683
2M	IQFull4WASI	5.1066	0.8885	< 0.0001
2M	IQPerfWASI	5.5818	0.9085	< 0.0001
2M	IQVerbalWASI	3.5284	0.9158	0.0001
2M	SOREMfSOPSG	11.1299	5.8202	0.0562
6M	Age	5.4205	0.8442	< 0.0001
6M	BMI	-0.2526	0.5092	0.6200
6M	TSTPSG	-15.254	4.6673	0.0011
6M	SleepEffPSG	-3.8157	0.9091	< 0.0001
6M	SOfLOPSG	4.1901	1.5284	0.0063
6M	PerTSTS3PSG	-0.6696	0.3294	0.0424
6M	PerTSTS4PSG	-0.1462	0.1437	0.3094
6M	PerTSTREMPSG	-0.6733	0.5013	0.1796
6M	RDITSTPSG	1.8909	1.7970	0.2930
6M	MinimumSPO_2QC	-0.5384	0.6059	0.3745
6M	PerTSTS1PSG	0.8400	1.0321	0.4160
6M	PerTSTS2PSG	0.6842	0.9948	0.4918
6M	PerSPO_2lt85TST	0.2022	0.4336	0.6411
6M	HighestGradeHP	0.0544	0.1889	0.7735
6M	IQFull4WASI	4.4960	0.9065	< 0.0001
6M	IQVerbalWASI	2.7468	0.9381	0.0035
6M	IQPerfWASI	5.3655	0.9314	< 0.0001
6M	SOREMfSOPSG	13.6840	5.9651	0.0221

Table S13—Comparison of nominal-scale baseline variables percentages between adherent and non-adherent subpopulations, by post-randomization visit

Visit	Table	% Adherent – % Nonadherent	P value
2M	Female	2.28	0.4953
2M	OSASeverityPostQC[a]		0.5231
2M	White	11.66	< 0.0001
2M	MarriedHP	6.85	0.0474
6M	Female	0.49	0.8867
6M	OSASeverityPostQC		0.5988
6M	White	11.78	< 0.0001
6M	MarriedHP	8.58	0.0161

[a]This factor had three levels and so percentages not reported here.

comparability between 2M and 6M, direct standardization[32] was used to provide an overall estimate per arm at 6M wherein each of the nine adherence-adjusted means (Table S16) were weighted according to the observed (see footnote A following appendix) frequency of participants of 0, 1,…8 hours of adherence at 2M.

This adjustment at 6M resulted in an estimated active mean [CI] of 0.098 [-0.035 to 0.231] and estimated sham mean [CI] of -0.002 [-0.010 to0.097]. Direct-standardization -adjusted and unadjusted point estimates of the mean for active indicate possible improvement at 6M from baseline. With direct standardization, the difference between sham and active means is nearly two-fold larger than without this standardization; although the confidence interval for the direct-standardization estimates of means for sham and active each include a mean change score of zero. The estimate for the sham mean has become negative, which agrees with the finding at 2M for sham worsening. However, we do not have evidence at 6M for a statistically significant decline from baseline, based on confidence intervals, so the possibility of sham worsening to completely explain our findings remains an open question. The confidence interval on the difference in direct-standardization means (active mean – sham mean) is [-0.056, 0.256], which includes a difference in means of zero.

7F. Future Work

We recognize that the extension of propensity methods to non-binary exposure variables has been an active area of research. Further analyses which adjust for adherence could certainly be conducted on the APPLES data that make use of other generalized propensity approaches, such as those of Imai and Van Dyk (2004).[33] Moreover, combined adjustment for adherence dose-response and retention merits exploration. These topics are being addressed in a separate manuscript in preparation.

SECTION 8. RESULTS – ADJUSTING PRIMARY NEUROCOGNITIVE ANALYSES FOR PARTICIPANT RETENTION

8A. Model Specification

A Heckman-type selection model was employed.[34] Let Δ be change from baseline on the neurocognitive outcome and D be the (latent) measure of the tendency to discontinue follow-up. Both outcomes are continuous. For person i,

$\Delta_i = \boldsymbol{X_i\beta} + E_{1i}$

$D_i = \boldsymbol{Z_i\gamma} + E_{2i}$

where $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$ are the variables associated with their respective outcomes, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of regression coefficients, and the $\{E_{1i}, E_{2i}\}$ follow a bivariate normal distribution of mean $\{0, 0\}$ and correlation parameter ρ. Δ_i is only observed when $D_i > 0$, That is, change scores on neurocognitive outcomes are only observed when the tendency to discontinue follow up crosses a threshold, typically set arbitrarily to zero as here. Denote the observed change scores by $\widetilde{\Delta}$. The APPLES Steering

Table S14—For adherent subpopulation, comparison of mean baseline characteristics and mean 1NC outcomes between sham and active arms, by post-randomization visit

Outcome	Analysis	Transformed	Visit	Active Mean Minus Sham Mean	Std. Error	P value
PFNTOTL	Parametric Survival	1/x	2M	0.0001	0.0009	0.9015
PFNTOTL	Parametric Survival	1/x	6M	0.0010	0.0010	0.3096
SWMTOverall	t-test	None	2M	0.0877	0.0515	0.0892
SWMTOverall	t-test	None	6M	0.0742	0.0629	0.2389
SumRecall	GLM	None	2M	-0.0043	0.0539	0.9364
SumRecall	GLM	None	6M	-0.0504	0.0659	0.4447
Age	t-test	None	2M	-0.8494	1.0550	0.4211
BMI	t-test	None	2M	0.6676	0.6428	0.2995
TSTPSG	t-test	None	2M	5.1766	5.7672	0.3698
SleepEffPS	t-test	None	2M	1.2351	1.1305	0.2751
SOfLOPSG	t-test	None	2M	-1.1520	1.9940	0.5637
PerTSTS3PSG	t-test	None	2M	0.6381	0.3943	0.1062
PerTSTS4PSG	t-test	None	2M	-0.2269	0.1578	0.1512
PerTSTREMP	t-test	None	2M	0.1442	0.6302	0.8191
RDITSTPSG	t-test	None	2M	1.3978	2.2296	0.5310
MinimumSPO_2	t-test	None	2M	-0.2071	0.7063	0.7695
PerTSTS1PSG	t-test	None	2M	-0.2004	1.2742	0.8751
PerTSTS2PSG	t-test	None	2M	-0.3620	1.2253	0.7678
PerSPO_2lt8	t-test	None	2M	0.5438	0.5808	0.3495
HighestGrade	t-test	None	2M	-0.2704	0.2297	0.2396
IQFull4WAS	t-test	None	2M	-1.1790	1.1016	0.2850
IQVerbalWASI	t-test	None	2M	-1.7519	1.1276	0.1209
IQPerfWASI	t-test	None	2M	-0.2129	1.1248	0.8500
SOREMfSOPSG	t-test	None	2M	-2.4781	7.5635	0.7433
Age	t-test	None	6M	-1.7605	1.1803	0.1366
BMI	t-test	None	6M	0.8159	0.7392	0.2704
IQFull4WASI	t-test	None	6M	-2.0680	1.2299	0.0934
IQVerbalWASI	t-test	None	6M	-2.5434	1.2668	0.0453
IQPerfWASI	t-test	None	6M	-1.0756	1.2746	0.3992
TSTPSG	t-test	None	6M	9.6719	6.7719	0.1540
SleepEffPSG	t-test	None	6M	2.5318	1.3543	0.0623
SOfLOPSG	t-test	None	6M	-3.0530	2.3431	0.1933
PerTSTS3PSG	t-test	None	6M	0.6751	0.4116	0.1017
PerTSTS4PSG	t-test	None	6M	-0.2642	0.1746	0.1310
PerTSTREMP	t-test	None	6M	0.1378	0.7173	0.8477
RDITSTPSG	t-test	None	6M	0.2463	2.5434	0.9229
MinimumSPO_2	t-test	None	6M	0.3983	0.8973	0.6574
PerTSTS1PSG	t-test	None	6M	0.2823	1.4750	0.8483
PerTSTS2PSG	t-test	None	6M	-0.8258	1.4527	0.5700
PerSPO_2lt8	t-test	None	6M	0.7776	0.6687	0.2456
HighestGrade	t-test	None	6M	-0.2245	0.2535	0.3763
SOREMfSOPSG	t-test	None	6M	-9.0680	8.5408	0.2890
Female[a]	Chi-square	None	2M	0.35		0.9348
OSASeverity[b]	Chi-square	None	2M			0.8962
White	Chi-square	None	2M	2.37		0.4868
MarriedHP	Chi-square	None	2M	1.09		0.8040
Female	Chi-square	None	6M	5.98		0.2227
OSASeverity	Chi-square	None	6M			0.9088
White	Chi-square	None	6M	2.99		0.4339
MarriedHP	Chi-square	None	6M	1.17		0.8128

[a]Differences in percentages are reported for nominal-scale variables. [b]This factor had three levels and so is not reported here.

Table S15—List of candidate confounders

Variable Category	Number of Variables	Variables
Demographics	11	Age, BMI, Married, WASI Full-4 IQ, WASI Verbal IQ, WASI Performance IQ, Highest Grade Level, MMSE Total Score, Ethnicity, Study Arm, Site
Health Variables	33	Caffeine Servings/Wk, Alcohol Servings/Wk, Current Smoker, CV History, AM Headaches, Dry Mouth/Throat, Bruxism, Nasal Congestion, Hypertension, Asthma, COPD, GERD, Chronic Pain Syndrome, Thyroid Disease, Diabetes, Eczema, Anemia, 5 Year Weight Gain > 20 Lbs, Allergic Rhinitis, Depression, Anxiety, Rhinoplasty, Cancer, Smoker, Claustrophobia, Neck Circumference, Nose Exam, Oral/Throat Exam, Coughing/Wheezing, Shortness of Breath, Pain in Joints/Muscles/Back, Leg Cramps/Jerks, Need to Go to Bathroom
Sleep Variables	38	AHI TST, AHI NREM, AHI REM, O_2 Sat < 85%TST, Avg SpO_2 NREM, Avg SpO_2 REM, Min SpO_2, Hrs Sleep/Night, Snore Duration, TIB, TST, Sleep Efficiency, SO after LO, %TSTS3, %TSTS4, %REM, Arousal Index, PLM Index, OA Index, CA Index, MA Index, Hypopnea Index, Avg SpO_2 Wake, Desaturation Index, Number of Awakenings, Naps/Wk, Difficulty Rising, EDS, Trouble Falling Asleep, Difficulty Falling Back to Sleep at Night, Difficulty Falling Back to Sleep in AM, Pain Affects Sleep, Worry About Sleep, Unrested During Day, Not Enough Sleep, Noisy Surroundings, MEQ Total Score, MEQ Category
Neurocognitive Outcomes	3	PVT Median RT, PVT Mean Slowest 10% of RTs, PASAT Total Correct
Mood Outcomes	9	HAM-D Total Score, POMS TMD, POMS Factor F, POMS Factor T, POMS Factor D, POMS Factor A, POMS Factor C, POMS Factor V, BDI Total Score
Sleepiness Outcomes	3	MWT Mean Sleep Latency, ESS Total Score, SSS Mean Score
Quality of Life Outcomes	5	SAQLI Total Score, SAQLI Domain A Mean, SAQLI Domain B Mean, SAQLI Domain C Mean, SAQLI Domain D Mean

Table S16—Estimated difference in means (Diff) and its estimated standard error (SE) of each 1NC outcome between arms (active minus sham) by mean hours of adherence per night, adjusted for confounders via generalized propensity scores

	Hours of Adherence																	
	0		1		2		3		4		5		6		7		8	
Outcome	Diff (SE)	P	Diff (SE)	P	Diff (SE)	P	Diff (SE)	P	Diff (SE)	P	Diff (SE)	P	Diff (SE)	P	Diff (SE)	P	Diff (SE)	P
PFNTOL																		
2M	2.111 (2.342)	0.367	1.261 (1.468)	0.390	0.671 (0.906)	0.459	0.342 (0.667)	0.608	0.273 (0.618)	0.658	0.464 (0.599)	0.439	0.916 (0.624)	0.142	1.628 (0.955)	0.088	2.600 (1.851)	0.160
6M	-0.092 (1.914)	0.961	-0.178 (1.344)	0.895	-0.237 (0.947)	0.803	-0.203 (0.734)	0.783	-0.057 (0.648)	0.930	0.134 (0.620)	0.829	0.262 (0.672)	0.697	0.272 (1.009)	0.787	0.202 (1.811)	0.911
Sum Recall																		
2M	-0.752 (3.37)	0.823	0.265 (2.24)	0.906	0.780 (1.54)	0.612	0.846 (1.22)	0.487	0.636 (1.08)	0.556	0.352 (1.01)	0.728	0.091 (1.07)	0.932	-0.220 (1.57)	0.889	-0.762 (2.80)	0.786
6M	0.960 (3.81)	0.801	0.589 (2.60)	0.821	0.223 (1.71)	0.896	-0.029 (1.21)	0.981	-0.097 (1.05)	0.927	0.007 (1.02)	0.994	0.196 (1.15)	0.864	0.359 (1.79)	0.841	0.417 (3.13)	0.894
Overall Midday																		
2M	0.071 (0.23)	0.756	0.100 (0.16)	0.526	0.136 (0.11)	0.210	0.163 (0.09)	0.044	0.156 (0.07)	0.023	0.110 (0.06)	0.078	0.050 (0.07)	0.470	0.008 (0.11)	0.946	-0.004 (0.21)	0.985
6M	0.144 (0.26)	0.583	0.134 (0.19)	0.486	0.121 (0.14)	0.374	0.107 (0.10)	0.282	0.095 (0.08)	0.256	0.088 (0.08)	0.282	0.086 (0.10)	0.375	0.087 (0.15)	0.563	0.089 (0.25)	0.729

SE was obtained via a standard bootstrap. Tests of significance were made by assuming the ratio of Diff/SE approximately follows a standard normal distribution under the null hypothesis of no difference in means between arms.

Committee (SC) identified the following variables for the $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ (Table S17).

Probit modeling was employed because whether or not a person discontinued was observed instead of D (i.e., D is latent). Joint estimation of parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and ρ was via maximum likelihood. For analysis at the two-month visit (2M), a participant was scored as having discontinued by two months if they provided no data on any of the three neurocognitive outcomes at 2M or the six-month visit (6M). For analysis at 6M, a participant was scored as having discontinued by 6M if they provided no data on any of the three neurocognitive outcomes at 6M, regardless of whether the three neurocognitive outcomes were provided at 2M or not. The sample size for each analysis was 1,098 minus only those cases where a participant was missing that particular neurocognitive outcome or one of its covariates (i.e., missing data not due to discontinuation from the study). These sample sizes were PFN Total 2M at 1,043, PFN Total

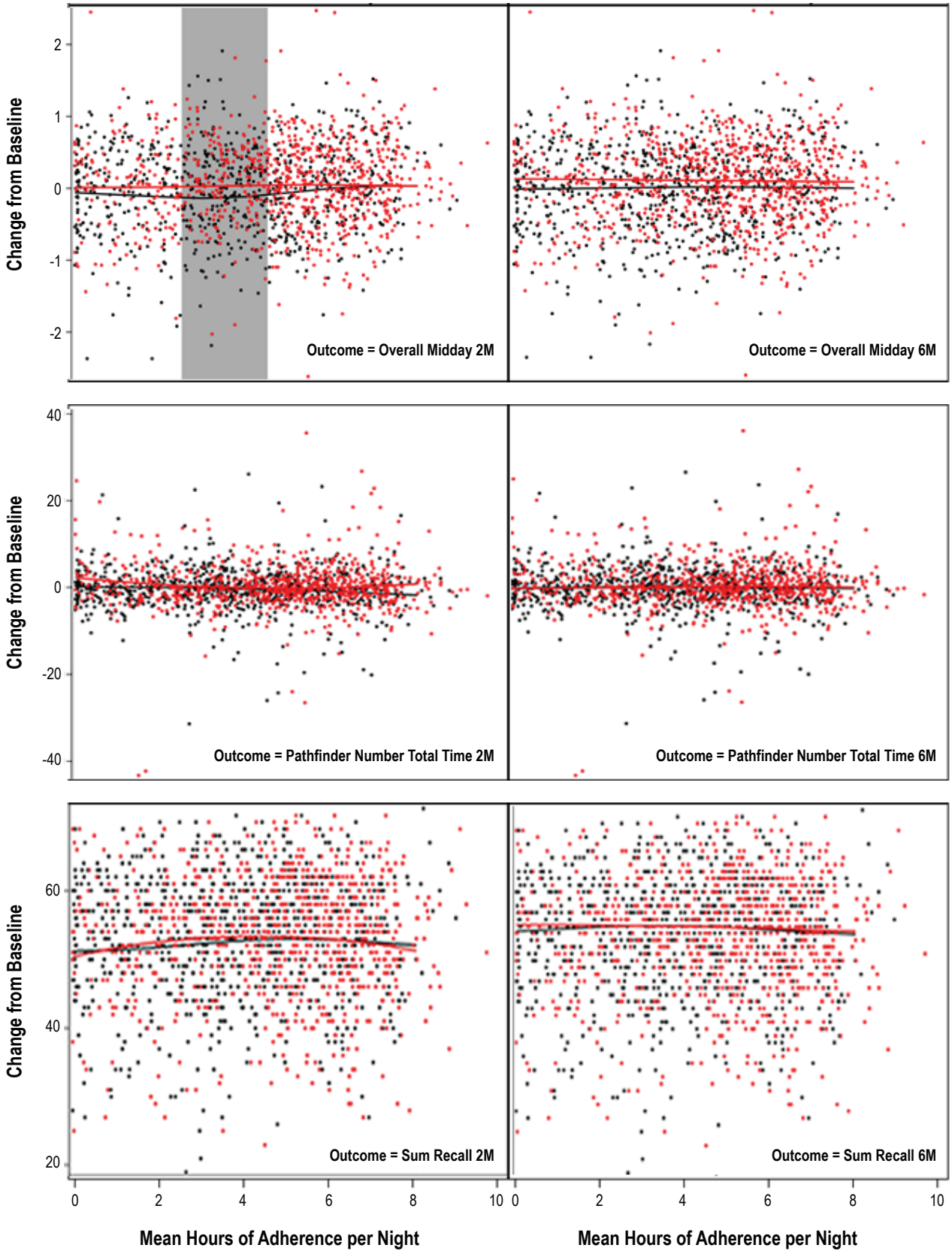

Figure S7—CPAP adherence-adjusted primary neurocognitive outcomes. Estimates (lines) of mean values for the three primary neurocognitive outcomes at 2M and 6M as a function of mean hours of CPAP adherence per night, with generalized propensity score adjustment for candidate confounders. Points are observed data with active CPAP (red) and sham CPAP (black); gray vertical bar marks average adherence levels where attained significance levels are $P < 0.05$. Differences detected only at intermediate mean hours of adherence could be due, at least in part, to the fact that confidence intervals on a regression line are wider toward its upper and lower ends.[31] Adherence was employed as a regressor in the second of three stages of the method of Hirano and Imbens.[26]

6M at 1,061, Sum Recall 2M at 1,046, Sum Recall 6M at 1,063, Overall Midday 2M at 1,006 and Overall Midday at 1,024.

8B. Assessing Model Assumptions

i. Bivariate normal distribution

Because the bivariate normality assumption is untestable, the model of section 1 was run for different transformations ($\log \widetilde{\Delta}$, $\sqrt{\widetilde{\Delta}}$ and $\widetilde{\Delta}^{3/2}$ [see footnote B following appendix]) (cf. ref[35]) of the observed change scores $\widetilde{\Delta}$. Results are summarized in Table S18 for baseline to 2M and baseline to 6M.

Overall, these results in combination with those for the untransformed outcome (Table S22) indicate that findings with regard to treatment effects are robust to assumptions about the shape of the distribution of the change outcome (conditional on the $\boldsymbol{X}_i$). The one possible exception is for PFN Total at 2M. For this outcome and visit, a more definitive analysis could explore application of methods which explicitly relax assumptions about the distributions of E_1 and E_2 (refs in [36]).

ii. Collinearity

Correlations among the variables listed in Table S17 were examined.[31] None were found to be highly correlated with each other, with all estimated correlations less than 0.74 (Tables S19 and S20).

iii. Exclusion Restriction

To help distinguish the processes that govern discontinuation versus neurocognitive performance, it is desirable to have covariates (possible "instruments") associated with the tendency to discontinue follow-up that are not associated with change in neurocognitive outcome.[37] Table S21 reveals that possible instruments were identified for all models fit except PFN Total at two months. Negative coefficients on the indicator variable for active arm suggest that sham condition caused dropout. Those participants with higher quality of life, higher intelligence, older age and better oxygen saturation status at baseline were less likely to discontinue; and these variables may serve as instruments as well.

8C. Results

Selection modeling results are given in Table S22. Correlations between the tendency to discontinue and neurocognitive

Table S17—Covariates proposed by the SC as possibly associated with each of the two outcomes

Neurocognitive Change Δ	Tendency to Discontinue D
Age < 60 (binary)	Age (years)
Gender	Gender
WASI Performance	WASI Performance
WASI Verbal	WASI Verbal
Moderate OSA (binary)	Apnea Hypopnea Index
Severe OSA (binary)	Avg SpO_2 NREM
% SpO_2 < 85	% SpO_2 < 85
Caucasian Race (binary)	Body Mass Index
Highest Education Level	Marital Status
	Minimum SpO_2
	SAQLI Total Score

Table S18—Sensitivity analyses by outcome and visit across three transformations

		$\log \widetilde{\Delta}$		$\sqrt{\widetilde{\Delta}}$		$\widetilde{\Delta}^{3/2}$	
Change Score	**Visit**	$\hat{\rho}$	$\widehat{Tx}$	$\hat{\rho}$	$\widehat{Tx}$	$\hat{\rho}$	$\widehat{Tx}$
PFN Total	2M	< 0.0001	0.7689	< 0.0001	0.1905	0.2563	0.0419
Sum Recall	2M	0.2483	0.2657	0.0742	0.2592	< 0.0001	0.1747
Overall Midday	2M	0.6656	0.0047	0.5104	0.0046	< 0.0001	0.0018
PFN Total	6M	< 0.0001	0.1418	< 0.0001	0.9764	0.1282	0.4757
Sum Recall	6M	< 0.0001	0.1484	< 0.0001	0.2055	0.1266	0.6212
Overall Midday	6M	< 0.0001	0.2268	< 0.0001	0.5196	0.9296	0.2964

$\widehat{Tx}$ denotes the estimated treatment effect (active mean minus sham mean) and $\hat{\rho}$ is the estimated correlation between the tendency to discontinue D and neurocognitive change from baseline Δ. Values in tables are the P-values associated with each estimate.

Table S19—Estimated Pearson correlations, point-biserial correlations and phi coefficients among covariates in model for neurocognitive change from baseline Δ as outcome

Variable	Active	White	AgeLT60	PerSpO_2lt85TST	HighestGradeHP	IQPerfWASI	IQVerbalWASI	BasePFN
Active	1.00	0.01	-0.06	-0.01	0.00	0.01	0.00	0.02
White	0.01	1.00	-0.18	-0.03	0.12	0.30	0.33	0.00
AgeLT60	-0.06	-0.18	1.00	0.03	-0.08	-0.07	-0.13	-0.36
Per SpO_2lt85TST	-0.01	-0.03	0.03	1.00	-0.09	-0.08	-0.11	0.02
HighestGradeHP	0.00	0.12	-0.08	-0.09	1.00	0.27	0.46	-0.08
IQPerfWASI	0.01	0.30	-0.07	-0.08	0.27	1.00	0.52	-0.25
IQVerbalWASI	0.00	0.33	-0.13	-0.11	0.46	0.52	1.00	-0.16
BasePFN	0.02	0.00	-0.36	0.02	-0.08	-0.25	-0.16	1.00

Table S20—Estimated Pearson correlations, point-biserial correlations and phi coefficients among covariates in model for discontinuation of follow-up as outcome

Variable	Active	Age	$AvgSpO_2$ REM	BMI	SAQLI Total Score	ESSTotal Score	IQPerf WASI	IQVerbal WASI	$MinSpO_2$ QC	$PerSpO_2$ lt85TST	RDITST PSG
Active	1.00	0.06	0.03	0.02	0.04	0.00	0.01	0.00	0.02	-0.01	-0.02
Age	0.06	1.00	-0.21	-0.16	0.16	-0.08	0.05	0.16	-0.08	-0.03	0.00
$AvgSpO_2$REM	0.03	-0.21	1.00	-0.42	-0.05	-0.03	0.08	0.11	0.74	-0.65	-0.49
BMI	0.02	-0.16	-0.42	1.00	-0.12	0.11	-0.18	-0.19	-0.40	0.32	0.38
SAQLITotalScore	0.04	0.16	-0.05	-0.12	1.00	-0.26	0.10	0.04	-0.04	0.01	-0.01
ESSTotal Score	0.00	-0.08	-0.03	0.11	-0.26	1.00	0.01	-0.03	-0.07	0.10	0.10
IQ Perf WASI	0.01	0.05	0.08	-0.18	0.10	0.01	1.00	0.52	0.11	-0.08	-0.07
IQ Verbal WASI	0.00	0.16	0.11	-0.19	0.04	-0.03	0.52	1.00	0.14	-0.11	-0.14
$MinSpO_2$QC	0.02	-0.08	0.74	-0.40	-0.04	-0.07	0.11	0.14	1.00	-0.60	-0.55
$PerSpO_2$lt85TST	-0.01	-0.03	-0.65	0.32	0.01	0.10	-0.08	-0.11	-0.60	1.00	0.48
RDI TST PSG	-0.02	0.00	-0.49	0.38	-0.01	0.10	-0.07	-0.14	-0.55	0.48	1.00

Table S21—Possible instrumental variables, estimated regression coefficients, and associated P values for discontinuation of follow-up as outcome

PFN Total		Sum Recall		Overall Midday	
2M	6M	2M	6M	2M	6M
None	Active (-0.19, 0.0274)	SAQLI (-0.10, 0.0367)	Active (-0.20, 0.0180)	SAQLI (-0.10, 0.0367)	Active (-0.19, 0.0306)
	$\%SpO_2 < 85$ (0.12, 0.0454)		$\%SpO_2 < 85$ (0.12, 0.0275)	WASI Perf (-0.15, 0.0059)	Age (-0.11, 0.0230)
			SAQLI (-0.10, 0.0160)	Age (-0.19, 0.0005)	
			WASI Perf (-0.12, 0.0244)		

change from baseline (conditional on the covariates $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$) were statistically significant for Sum Recall, at two months and six months, and for Overall Midday at six months. The negative sign of the correlation for Sum Recall by two months suggests that participants who are doing worse neurocognitively have greater tendency to leave during this early phase of follow-up. This situation may change during late follow-up. The positive signs on correlation coefficients by six months indicate that participants who do worse neurocognitively are less likely to discontinue by the end of six months of follow-up. The results by six months are stronger evidence in two regards. (1) Significant correlations were identified for two primary neurocognitive outcomes (Sum Recall and Overall Midday) and perhaps a third (PFN Total, P = 0.0549) while only one correlation was significant by two months (Sum Recall). (2) Estimated correlations are larger in absolute value by six months compared to two months.

8D. Conclusions

i. Results are generally robust to transformations on the neurocognitive outcome, no evidence of collinearity among the covariates of Table S17 were identified, and possible instruments were detected for the completion outcome. Taken altogether, the assumptions underlying application of a Heckman-type selection model appear to have been satisfied. One possible exception might be PFN Total at 2M, for which detection of a treatment effect did vary with transformation and for which no possible instruments were detected.

ii. Different factors (possible instruments) may govern dropout (Table S21). Among these, the sham condition appears to have been a cause of dropout by six months, as evidenced for all three primary outcomes. Differential dropout between arms was also identified via life-table and competing risks analyses, as reported in the main paper.

iii. Completion status appears to be associated with change from baseline ($\hat{\rho}$ of Table S22) after adjusting for covariates. In particular, evidence from Sum Recall suggests those who do worse neurocognitively during the first two months are more likely to leave the study early; but, by the end of follow-up, evidence from two to perhaps all three neurocognitive outcomes suggests those who are doing worse neurocognitively are less likely to leave the study. Evidence is stronger for the latter finding.

iv. Taking these results together, by six months the sham condition appears to cause some amount of discontinuation; however, beyond that effect, those who are doing worse neurocognitively are less inclined to discontinue.

v. When allowance is made for the potentially informative dropout via selection modeling, statistical detectabilities of treatment effects on primary outcomes remain unchanged ($\widehat{Tx}$ of Table S22) compared to the results reported in the main paper without this adjustment.

SECTION 9. RESULTS – SAFETY

All Serious Adverse Events (SAEs) and Adverse Events (AEs) were categorized into one of 17 body systems/event categories by the DCC Medical Director. Analyses were performed on all post-randomization SAEs and AEs and tabulated to report incidence proportions. Multiple events for an individual subject were recorded and defined as a single On-Study incidence. All safety analyses used GLM. The Poisson distribution was used to model rare events (incidences less

Table S22—Estimate of correlation coefficient $\hat{\rho}$ between tendency to discontinue follow-up and change in neurocognitive outcome

Change Score	Visit	$\hat{\rho}$ (P value)	$\widehat{Tx}$ (P value)
PFN Total	2M	0.23 (0.1399)	0.61 (0.0601)
PFN Total	6M	0.25 (0.0549)	0.22 (0.5088)
Sum Recall	2M	-0.39 (0.0075)	0.52 (0.2380)
Sum Recall	6M	0.70 (< 0.0001)	-0.44 (0.3339)
Overall Midday	2M	-0.21 (0.3500)	0.12 (0.0047)
Overall Midday	6M	0.80 (< 0.0001)	-0.01 (0.9130)

Estimate of difference in means $\widehat{Tx}$ (active arm minus sham arm) for change in neurocognitive outcome.

Table 23—Post-randomization serious adverse event incidence proportions (cardiovascular, MVA, and deaths) comparison of quantity of participants with at least one event between study arms

Event Category	CPAP Study Arm	Number of Participants with ≥1 Event	Incidence Proportion†	P Value
SAE Only				
Cardiovascular	Active	4	0.00719	0.5044
	Sham	6	0.01107	
MVA	Active	0	0	n/a
	Sham	0	0	
Death	Active	2	0.00360	0.9797
	Sham	2	0.00369	

†Sample sizes: Active CPAP = 556 Ps, Sham CPAP = 542 Ps.

than 10%). For non-rare events, the binomial distribution was employed to account for the greater dependence of the variance on the finite population size. Table S23 provides comparisons of incidence proportions between study arms made for all SAEs in the Cardiovascular, motor vehicle accident (MVA), or Death event categories. These three body system/event categories were deemed the most import to examine by the APPLES Steering Committee and Data and Safety Monitoring Board (DSMB). Table S24 provides comparisons of incidence rates between study arms for all safety events (SAE+AE) in all body system/event categories.

FOOTNOTE A

We conditioned on the observed frequencies. A more thorough analysis would incorporate the sampling error in the estimated frequencies from the sample at 2M. This would not alter conclusions here because reported conditional confidence intervals include zero.

FOOTNOTE B

The transformations were actually more complicated than this. A shift constant was added to each variable to make all values positive before logarithmic, square-root or 3/2 power transformation.

Table 24—Post-randomization safety event incidence rates (for all categories) comparison of quantity of participants with at least one event between study arms

Event Category	CPAP Study Arm	Number of Participants with ≥ 1 Event	Incidence Rate per Participant	P Value
SAE + AEs				
Cardiovascular	Active	31	0.0558	0.8733
	Sham	29	0.0535	
MVA	Active	10	0.0180	0.7822
	Sham	11	0.0203	
Death	Active	2	0.0036	0.9797
	Sham	2	0.0037	
Dermatological	Active	102	0.1835	0.0011*
	Sham	61	0.1126	
Endocrinological	Active	7	0.0126	0.2337
	Sham	3	0.0055	
GI/Digestive	Active	37	0.0666	0.7104
	Sham	33	0.0609	
General	Active	53	0.0953	0.1825
	Sham	39	0.0720	
Genitourinary	Active	13	0.0234	0.6564
	Sham	15	0.0277	
Head, Eyes, Ears, Nose, and Throat	Active	208	0.3741	0.0020*
	Sham	155	0.2860	
Hematologic/ Lymphatic	Active	3	0.0054	0.9751
	Sham	3	0.0055	
Musculoskeletal	Active	54	0.0971	0.7164
	Sham	49	0.0904	
Near-miss MVA	Active	7	0.0126	0.4380
	Sham	10	0.0185	
Neurological	Active	36	0.0648	0.7041
	Sham	32	0.0590	
Other Accident	Active	21	0.0378	0.2780
	Sham	28	0.0517	
Psychiatric	Active	42	0.0755	0.0703
	Sham	59	0.1089	
Respiratory	Active	136	0.2446	0.1377
	Sham	154	0.2841	
Work-related Accident	Active	4	0.0072	0.2236
	Sham	1	0.0019	

*P < 0.05 indicates statistical significance.

SECTION 10. REFERENCES

1. Kushida CA, Kuo T, McEvoy L, Gevins A, Guilleminault C, Dement WC. Apnea Positive Pressure Long-Term Efficacy Study (APPLES): Preliminary Studies. Sleep 2004;27(Supplement):A181-182.
2. Krieger J, Kurtz D, Petiau C, Sforza E, Trautmann D. Long-term compliance with CPAP therapy in obstructive sleep apnea patients and in snorers. Sleep Nov 1996;19(9 Suppl):S136-143.
3. McArdle N, Devereux G, Heidarnejad H, Engleman HM, Mackay TW, Douglas NJ. Long-term use of CPAP therapy for sleep apnea/hypopnea syndrome. Am J Respir Crit Care Med 1999;159(4 Pt 1):1108-14.
4. Engleman HM, Kingshott RN, Martin SE, Douglas NJ. Cognitive function in the sleep apnea/hypopnea syndrome (SAHS). Sleep 2000;23 Suppl 4:S102-108.

5. Bedard MA, Montplaisir J, Richer F, Rouleau I, Malo J. Obstructive sleep apnea syndrome: pathogenesis of neuropsychological deficits. J Clin Exp Neuropsychol 1991;13:950-64.
6. Kim HC, Young T, Matthews CG, Weber SM, Woodward AR, Palta M. Sleep-disordered breathing and neuropsychological deficits. A population-based study. Am J Respir Crit Care Med 1997;156:1813-9.
7. Kushida CA, Nichols DA, Quan SF, et al. The Apnea Positive Pressure Long-term Efficacy Study (APPLES): rationale, design, methods, and procedures. J Clin Sleep Med 2006;2:288-300.
8. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13-22.
9. Hannay HJ, Levin HS. Selective reminding test: an examination of the equivalence of four forms. J Clin Exp Neuropsychol 1985;7:251-63.
10. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. Br J Cancer 1976;34:585-612.
11. Lachin JM. Statistical considerations in the intent-to-treat principle. Control Clin Trials 2000;21:167-89.
12. McCullagh P, Nelder JA. Generalized Linear Models. 2nd ed. London: Chapman & Hall, Inc.; 1991.
13. Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc; 2002.
14. Milliken GA, Johnson DE. Analysis of Messy Data: Designed Experiments. Vol 1. Boca Raton, FL: Chapman & Hall/CRC; 1992.
15. Lan K, DeMets D. Discrete sequential boundaries for clinical trials. Biometrika. 1983;70:659-63.
16. Holm S. A simple sequentially rejective multiple test procedure. Scand Stat Theory Appl 1979;6:65-70.
17. Quan SF, Wright R, Baldwin CM, et al. Obstructive sleep apnea-hypopnea and neurocognitive functioning in the Sleep Heart Health Study. Sleep Med 2006;7:498-507.
18. Doghramji K, Mitler MM, Sangal RB, et al. A normative study of the maintenance of wakefulness test (MWT). Electroencephalogr Clin Neurophysiol 1997;103:554-62.
19. Krzanowski WJ. A stopping rule for structure-preserving variable selection. Stat Comput 1996;6:51-6.
20. Wang A, Gehan EA. Gene selection for microarray data analysis using principal component analysis. Stat Med 2005;24:2069-87.
21. Krzanowski WJ. Cross-validation in principal component analysis. Biometrics 1987;43:575-84.
22. Stone JV. Independent component analysis: an introduction. Trends Cogn Sci 2002;6:59-64.
23. Daniel WW. Applied Nonparametric Statistics. 2nd ed. Boston, MA: PWS-Kent Publishing Company; 1990.
24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
25. Apnea Positive Pressure Long-Term Efficacy Study (APPLES) Manual of Operations: NHLBI-APPLES. 2003.
26. Hirano K, Imbens GW. The propensity score with continuous treatments. In: Gelman A, Meng X-L, eds. Applied Bayesian Modeling and Causal Inference from an Incomplete-Data Perspective. Hoboken, N.J.: John Wiley & Sons, Inc.; 2004:73-84.
27. Rubin DB. Estimating causal effects from large data sets using propensity scores. Ann Intern Med 1997;127(8 Pt 2):757-63.
28. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 1996;58:267-88.
29. Tomer A. The structure of cognitive speed measures in old and young adults. Multivariate Behav Res 1993;28:1-24.
30. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc 1984;79:516-24.
31. Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. Applied Linear Statistical Models. 4th ed. New York: WCB McGraw-Hill; 1996.
32. Fleiss JL. Statistical Methods for Rates and Proportions. 3rd ed. New Jersey: John Wiley & Sons Inc; 2003.
33. Imai K, Van Dyk DA. Causal inference with general treatment regimes: Generalizing the propensity score. J Am Stat Assoc 2004;99:854-66.
34. Heckman JJ. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement 1976;5:120-37.
35. Little R, Rubin D. Statistical analysis with missing data. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc.; 2002.
36. Puhani PA. The Heckman correction for sample selection and its critique. J Econ Surv 2000;14:53-68.
37. Heckman JJ, Vytlacil E. Policy-relevant treatment effects. Am Econ Rev 2001;91:107-11.