Supplementary Online Content

Redline S, Cook K, Chervin RD, et al; Pediatric Adenotonsillectomy Trial for Snoring (PATS) Study Team. Adenotonsillectomy for snoring and mild sleep apnea in children: a randomized clinical trial. *JAMA*. doi:10.1001/jama.2023.22114

Supplemental Study Information

Study Investigators and Institutions

Investigator	**Institution**
Raouf Amin, MD	Cincinnati Children's Hospital Medical Center, Cincinnati, OH
Jessie Bakker, Ph.D.	Brigham and Women's Hospital, Harvard Medical School, Boston, MA
Cristina Baldassari, MD	Children's Hospital of the King's Daughters, Norfolk, VA
Caron AC Clark, Ph.D.	University of Nebraska-Lincoln, NE
Kaitlyn Cook, Ph.D.	Harvard Pilgrim Health Center, Harvard Medical School, Boston, MA
Lisa Elden, MD	Children's Hospital of Philadelphia, Philadelphia
Susan Furth, MD	Children's Hospital of Philadelphia, Philadelphia
Susan Garetz, MD	University of Michigan Health System, Ann Arbor, MI
Fauziya Hassan, MD	University of Michigan Health System, Ann Arbor, MI
Sally Ibrahim, MD	University Hospitals-Cleveland Medical Center, Cleveland, OH
Stacey Ishman, MD	Cincinnati Children's Hospital Medical Center, Cincinnati, OH
Erin M Kirkham, MD, MPH	University of Michigan Health System, Ann Arbor, MI
Dongdong Li, Ph.D.	Harvard Pilgrim Health Center, Harvard Medical School, Boston, MA
Carole Marcus, MBChB	Children's Hospital of Philadelphia, Philadelphia, PA
Ron Mitchell, MD	University of Texas Southwestern Medical Center, Dallas, TX
Kamal Naqvi, MD	University of Texas Southwestern Medical Center, Dallas, TX
Todd Otteson, MD	University Hospitals-Cleveland Medical Center, Cleveland, OH
Judith Owens, MD	Boston Children's Hospital, Harvard University, Boston MA
Jerilynn Radcliffe, Ph.D.	Children's Hospital of Philadelphia, Philadelphia, PA
Susan Redline, MD, Ph.D.	Brigham and Women's Hospital, Harvard Medical School, Boston, MA
Carol Rosen, MD	University Hospitals-Cleveland Medical Center, Cleveland, OH

Jay Shah, MD	University Hospitals-Cleveland Medical Center, Cleveland, OH
Ignacio Tapia, MD	Children's Hospital of Philadelphia, Philadelphia, PA
H Gerry Taylor, Ph.D.	Abigail Wexner Research Institute at Nationwide Children's Hospital and The Ohio State University, Columbus, OH
Rui Wang, Ph.D.	Harvard Pilgrim Health Center and Brigham and Women's Hospital, Harvard Medical School, and Harvard TH Chan School of Public Health, Boston, MA
Lisa Young, MD	Children's Hospital of Philadelphia, Philadelphia, PA
Jay Zopf, MD	University Hospitals-Cleveland Medical Center, Cleveland, OH

All the authors vouch for the completeness and accuracy of the data and the fidelity of the study to the protocol. The study did not receive commercial support.

Recruitment Sites

Participants were recruited from the following clinical sites: Boston Children's Hospital, Boston MA; Children's Hospital of Philadelphia, Philadelphia, PA; Children's Hospital of the King's Daughters, Norfolk, VA; Cincinnati Children's Hospital Medical Center, Cincinnati, OH; Rainbow Babies and Children's Hospital, University Hospitals-Cleveland Medical Center, Cleveland, OH; University of Michigan Health System, Ann Arbor, MI; University of Texas Southwestern Medical Center, Dallas, TX. Note that the Boston Children's Hospital site entered the study late and was closed early due to low enrollment. The one child recruited from that site was not included in analyses of the primary or secondary endpoints but was included in the reporting of adverse events.

Study Entities

Data Coordinating Center: Brigham and Women's Hospital and Harvard Pilgrim Health Center, Harvard Medical School, Boston, MA- Susan Redline, Rui Wang, Co-Directors

Clinical Coordinating Center: Children's Hospital of Philadelphia, Philadelphia; PA- Carole Marcus, Director 2016-2017 (deceased); Co-Directors: Susan Furth 2017-2023, Lisa Young, 2019-2023

Sleep Reading Center: Brigham and Women's Hospital, Harvard Medical School, Boston, MA-Susan Redline, Director; Daniel Mobley, Manager

Neurobehavioral Quality Control Subcommittee: Jerilynn Radcliffe; H Gerry Taylor, Co-Directors

Surgical Subcommittee: Susan Garetz, MD, Director

Independent Medical Monitor: Heidi Connolly, MD, University of Rochester, Rochester, NY

NIH: National Heart Lung and Blood Institute

Data and Safety Monitoring Board

Meyer Kattan, MD, Chair, Columbia University Medical Center, New York, NY
Kristen H. Archbold, PhD, RN, College of Nursing, University of Tennessee Health Science Center, Germantown, TN
James S. Kemp, MD, St. Louis Children's Hospital, Saint Louis, MO
Tonya S. King, PhD, Penn State University College of Medicine, Hershey, PA
Cynthia D. Morris, PhD, Oregon Health Sciences University, Portland, OR
David E. Tunkel, MD, Johns Hopkins University, Baltimore, MD
Kathryn L. Weise, MD, MA, Cleveland Clinic, Cleveland, OH

Supplemental Methods

Recruitment

Children were recruited from several sources, including Ear Nose & Throat (ENT) clinics, sleep clinics, and sleep laboratories. Initial eligibility criteria such as age range, referring diagnosis, and comorbidities were identified by review of the electronic health record and other clinical records as permitted by a partial Health Insurance Portability and Accountability (HIPAA) waiver. Prior to scheduling a screening visit, referring physicians were contacted to confirm appropriateness of study participation for each child and to agree to allow study personnel to contact the child/caregiver. Following consent, a standardized set of screening questions were administered to further assess eligibility criteria. Polysomnography studies (PSGs) and ENT evaluations performed as part of routine clinical care were obtained and reviewed centrally to confirm eligibility criteria. Children who met initial study eligibility criteria who had not had an overnight PSG within 60 days of consent underwent a research PSG. Children who met initial study eligibility criteria who had not had an ENT evaluation within 90 days of consent were referred to an otolaryngologist to ensure appropriateness for participation in the trial.

Ethics

Written informed consent was obtained from caregivers and assent obtained from children 7 years or older from all sites other than King's Daughters, for which assent was obtained per local regulations for children 8 years and older. The study was approved by a single institutional review board at Children's Hospital of Philadelphia and at the local boards at each participating site.

Study Eligibility

Inclusion criteria were:

1. Aged $\geq$3 and <13 years at the date of consent. (Note, initial criteria identified an upper age limit of < 10 years, but was changed to 12.9 years on Nov 21, 2016 based on interest in expanding the applicability of findings to older children in whom growing prevalence of SDB was clinically recognized, while inclusive of an age range for which pediatric scoring rules were applicable.)
2. Mild SDB, as defined as meeting all of the following criteria:
    a. Caregiver report of habitual snoring that occurs most of the night on at least three nights per week, and was present for at least three months (on average occurring > three nights per week or more half of sleep time) *and*
    b. A centrally scored PSG confirming an obstructive apnea index (OAI) <1/hour *and* obstructive AHI <3/hour *and* no SpO_2 desaturation < 90% in conjunction with obstructive events. (Note, initial upper AHI limit was <2/hour, but was changed to

3/hour on Nov 21, 2016 based on review of the impact of changes in the AASM 2012 scoring rules that decreased amplitude reductions of hypopneas from 50% to 30%, which increased the number of hypopneas that are identified).

3. Tonsillar hypertrophy obstructing at least 25% oropharyngeal obstruction (Brodsky scale≥ 2)
4. Deemed to be a candidate for adenotonsillectory (AT) on clinical evaluation by ENT; including no technical issues that would be a contraindication for surgery such as submucus cleft palate
5. Primary indication for AT was nocturnal obstructive symptoms (that is, not recurrent infections or other indications)

Exclusion criteria were:

1. Previous tonsillectomy
2. Recurrent tonsillitis that merited prompt AT per the American Academy of Otolaryngology-Head and Neck Surgery Clinical Practice Guidelines; that is, ≥7 episodes/year in the past year; ≥5 episodes/year over the past two years or ≥3 episodes/year over the past three years
3. Severe obesity (BMI z-score ≥3)
4. Failure to thrive, defined as either height or weight being below the 5th percentile for age and gender
5. Severe chronic health condition/s that might hamper participation or confound key variables under study. These conditions included, but were not limited to:
    a. Severe cardiopulmonary disorders such as cystic fibrosis, congenital heart disease
    b. Bleeding disorders
    c. Sickle cell disease
    d. Epilepsy requiring medication
    e. Significant arrhythmia noted on PSG including: Non-sustained ventricular tachycardia, atrial fibrillation, second degree atrioventricular block, sustained bradycardia, or sustained tachycardia
    f. Other severe chronic health problems such as diabetes, narcolepsy, poorly controlled asthma
    g. Intellectual deficit or assigned to a self-contained classroom for all academic subjects
    h. Known genetic, craniofacial, neurological or psychiatric conditions likely to affect the airway, cognition or behavior
    i. Current use of psychotropic medication (other than medications for attention deficit hyperactivity disorder [ADHD]), hypnotics, melatonin, antihypertensives, hypoglycemic agents including insulin, anticonvulsants, anticoagulants, or growth hormone
    j. A diagnosis of autism spectrum disorder
    k. History of severe developmental disability or an ABAS-3 score <60
    l. Children/caregivers planning to move out of the area within the year
    m. Children/caregivers who do not speak English well enough to complete the neurobehavioral measures, as validated versions in other languages are not available for all of the measures.

Overview of Protocol

At baseline, 6-month and 12-month examinations, children were studied at a pediatric research center at a time when they were free of acute illness (after sleeping at home following their typical

schedule). Assessments included clinical evaluation, anthropometry, neurobehavioral testing, and distribution of wrist actigraphy devices for in-home use for 7 days. At baseline, morning urine and blood samples were collected. Neurobehavioral testing was conducted by staff blinded to treatment, trained and supervised by board-certified psychologists. Assessments began 0800-0900 in an area suitable for pediatric neurocognitive assessments. Children were encouraged to follow their usual bedtime routine the night prior to testing. The order of administration for child assessments were (1) 9-Hole Pegboard Dexterity Test; (2) GNG Test; and (3) Child Report version of PedsQL (as age appropriate). Between visits, caregivers were contacted every 1-2 months to ascertain any change in health status (identifying adverse events, health care utilization) and medications, and to reinforce study participation. After March 2020, the 6-month examinations were simplified to address the challenges of in-person testing during the COVID-19 pandemic. Those 6-month follow-up visits were permitted to be conducted remotely, focusing on collection of the BRIEF primary outcome.

Data Collected

The following lists pre-specified primary and secondary endpoints. For this primary outcome paper, we focused on the trial's two primary endpoints (BRIEF GEC and GNG CPT d-prime) and 22 selected secondary outcomes available for the entire sample. This paper does not report on teacher reported outcomes or actigraphy, which were only available for a subset and were impacted by the COVID-19 pandemic, or the Connors' behavioural assessment, which was only available for children ages 6 and older. Health care utilization will be reported in a separate paper given the unique complexity of that outcome.

Endpoints reported in this paper are bolded. Note that the primary analysis includes all PATS participants, save one child who was the only subject randomized in from Boston Children's Hospital.

The co-primary outcomes are:

- **BRIEF global composite score;**
- **GNG continuous performance measure, d-prime (d')**

The secondary outcomes are:

Objective performance testing:

- **Fine motor coordination: NIH Toolbox 9-Hole Pegboard Dexterity Test**.

Behavioral scales:

- **Executive function: BRIEF and BRIEF-P subscales from caregiver reports (here reporting the 5 subscales that overlap BRIEF and BRIEF-P)** and teacher reports.
- **Behavior: Child Behavior Checklist– Parent Version (CBCL)** and Teacher Report Form (TRF) summary score and subscale scores
- Attention: Conners 3rd Edition Short Form (Conners 3) (both the caregiver and teacher versions).
- **Sleepiness: Epworth Sleepiness Scale (mESS; modified for children) total score**
- **SDB symptoms burden: Pediatric Sleep Questionnaire Sleep Related Breathing Disorder (SRBD) subscale**

Quality of life:

- **Generic: Pediatric Quality of Life Inventory (PedsQL) total score**
- **Disease-specific: Obstructive Sleep Apnea-18 (OSA-18) total score**

Physical examination:

- Anthropometrics: **height, weight**, waist circumference, hip circumference, neck circumference.
- **Systolic and Diastolic Blood pressure (BP) and heart rate**

Healthcare utilization:

- Monthly caregiver interviews, including inquiries for adverse events (AEs), changes in health status, changes in medications, and healthcare visits.
- Electronic medical record (EMR) surveillance for hospitalizations, emergency room visits, medical or surgical procedures, consultation visits, medication prescriptions.

Polysomnography (PSG)

All children underwent full-night PSG by study-certified technicians using a standardized protocol and following the AASM Manual for the Scoring of Sleep and Associated Events Version 2.2 pediatric standards at least a week before the baseline assessments. Scoring was performed according to the AASM pediatric criteria by certified technologists blinded to all other study data at a central sleep reading center (Brigham and Women's Hospital). The Apnea-Hypopnea Index (AHI) was defined as the sum of all obstructive and mixed apneas, plus hypopneas associated with a 30% reduction in airflow and either a > 3% desaturation or electroencephalographic arousal, divided by hours of total sleep time.

PSG equipment that currently existed at each site was used, which included an interface to an end-tidal CO_2 monitor. To the extent possible, ancillary equipment and sensors were standardized across sites by requiring the use of a PATS-specific montage with defined sampling rates and digital specifications. Children were monitored by a certified sleep technician trained in pediatric PSG under the supervision of a lead technician who was certified at central training or another

designee approved by the trial's Sleep Reading Center Chief Polysomnologist. All PSG data were edited, scored and summarized at the Brigham and Women's Hospital Sleep Reading Center using well-developed quality assurance approaches[1]. Intra- and inter-class correlations for key PSG parameters was monitored over time, and generally exceeded 0.90.

The PATS montage consisted of:

- 6 EEG sites: F_3-M_2, F_4-M_1, C_3-M_2, C_4-M_1, O_1-M_2, O_2-M_1
- Ground and reference electrodes: CZ , Fpz
- Bilateral electrooculograms (EOG): E1-M2, E2-M2
- Submental electromyography (EMG): LChin, RChin, CChin
- ECG with standard 3-lead precordial placement: ECG1, ECG2, ECG3
- Airflow via oronasal thermocouple
- Nasal pressure flow via nasal cannula
- Snoring vibration via tracheal (piezo) sensor
- Respiratory effort via chest and abdominal wall inductive plethysmography
- Capnography waveform and numeric values (End-tidal CO2)
- Pulse oximetry plethysmograph waveforms and numeric values (2 sec averaging mode preferred, with an acceptable maximum of 3 sec).
- Leg movements via bilateral EMG: LLeg1-LLeg2, RLeg1-RLeg2 (gold cups or single-use cups preferred, with piezoelectric sensors as acceptable alternatives).
- Body position

Progression to an AHI ≥3 and AHI ≥5 at 12 month follow-up was pre-specified as a secondary analysis and included in the False Discovery Rate adjustment for multiple comparisons. Analyses of other measures of sleep architecture from polysomnography (sleep stage percentages, arousal index, continuously measured AHI) were considered exploratory and were not included in multiple comparison adjustments.

Measurements

Physical Examination: The PI or his/her designee reviewed the child's medical and sleep history and performed a brief physical examination, including standardized assessment of tonsillar size (Brodsky scale), evaluation of the oropharynx using Friedman and Mallampati scales[2], and identified any abnormalities on heart, lung, neurological and ears, nose and throat assessments.

Blood Pressure: After a 5-minute period of seated rest, systolic and diastolic BP and pulse were measured using cuffs measured to be appropriate for their arm circumference. Pre-school children were, as appropriate, seated on their caregiver's lap. Older children were asked to sit in a chair supporting the back and feet with legs uncrossed. Three recordings were taken, with a 1-minute interval between each. During this time, the child's arm was lifted and held for a period of 15

seconds. The average of the three measurements was calculated and converted to age and height adjusted percentiles[3].

Urinary cotinine: Cotinine was assayed in duplicate at the Translational Core Laboratory at Children's Hospital of Philadelphia from frozen urine samples using the Abnova Cotinine ELISA kit and reported as the average of the two assays.

Height and weight: Standing height was measured to the nearest mm (0.1 cm) with the child in stocking feet, using a wall-mounted stadiometer. The child was positioned so that their heels, buttocks, back and head were touching the backboard of the stadiometer and positioned with the Frankfurt plane parallel. Weight was measured to the nearest 0.1 kg using a calibrated digital scale, with the child in stocking feet and with no blue jeans. Height and weight were made in triplicate and averaged, and converted into BMI-, age-, and sex-adjusted percentiles and z-scores (http://www.cdc.gov/growthcharts/).

Go/No-go Continuous Performance Task (GNG CPT): The is a computer-based attention test developed for longitudinal studies of heterogenous samples of children ages 3 to 12 years. After a practice session, children were presented with stimuli that consisted of different colored cartoon fish and grey-colored sharks and asked to 'catch' the fish by pressing a key on a single-button response pad as quickly as possible. The primary test outcome from this test is d prime for the Continuous Performance Test (CPT) trial block, a signal detectability parameter that assesses the child's ability to correctly identify targets corrected for their response bias (higher is better). The second task block was used to assess sustained attention. Trial display time and inter-trial variability was adjusted in test versions designed for children below 5 years, 5 through 6 years inclusive, and older than 7 years of age[4]. Details of the psychometric properties of this test are reported in Clark et al.[5] and indicate that the GNG/CPT task has limited floor or ceiling effects, was sensitive to development, and correlates with parent-reported executive function and externalizing behaviour. Specifically, the internal reliability of the GNG/CPT was high (coefficient omega = .85) and internal validity of the GNG as an attention test is supported by a correlation of r=.72 between the task's two latent factors (sustained attention and response inhibition). The GNG Test is programmed to include practice presentations. Additional teaching on the GNG Test involved showing the child how to press the button and practice pressing the button first to command and then when the RA pretends that a fish has appeared.

The NIH Toolbox 9-Hole Pegboard Dexterity Test was used to assess fine motor control. [6] Scores for this test are the times needed for the child to put each of 9 pegs into the pegboard and then take them out using each hand, converted to age-adjusted scale scores (higher is worse). Scores

for each hand were averaged to obtain a total score. In children, this measure is reported to have a high test-retest reliability (r's= 0.81 for the dominant hand and 0.79 for the non-dominant hand) and high inter-rater agreement (r's> 0.99).[7]

Evaluation Instruments

Behavior Rating Inventory of Executive Function (BRIEF): Caregivers completed age-appropriate versions of the BRIEF (BRIEF-Preschool Edition for ages 2-4 and age 5 years if in preschool, or the BRIEF 2nd Edition for ages 6-18 and age 5 years in kindergarten, PAR Inc, Lutz, Florida). [8,9] These instruments survey behaviours associated with executive functioning (ability to self-regulate, pay attention and organize in "real-world" situations) and provide clinically relevant information on a wide range of behaviours, have good convergent and divergent validity, high test-retest reliability including in preschool children (total score r = 0.90), and high internal consistency (alpha values 0.80 - 0.98) [10]. The primary outcome was the BRIEF Global Composite Executive Score (GEC), which provides summary information on meta-cognition and self-regulation (higher score equates to lower function).

The Child Behavior Checklist (CBCL): The CBCL is a caregiver-completed assessments of child behaviour. The CBCL/1.5-5 (ASEBA, Burlington, VT) was administered to children 5 years or younger, while the CBCL/6-18 was administered to children 6 years or older. The total score combines the internalizing and externalizing scores. Higher T-scores indicate worse functioning, and a T-score of 65 or higher is considered clinically abnormal.

Epworth Sleepiness Scale modified for children (mESS): The mESS is an 8-item questionnaire with scores that range from 0 to 24. Higher scores indicating greater daytime sleepiness. Elevated scores were defined as ≥ 10.

Pediatric Sleep Questionnaire (PSQ)-- Sleep-Related Breathing Disorder Scale (PSQ): The PSQ-SRBD is a 22-item questionnaire that captures symptoms of SDB. It includes 3 subscales: snoring, daytime sleepiness, and hyperactive behaviors/inattention and can be used to characterize SDB symptom burden. The PSQ-SRBD is commonly used to assess SDB risk in pediatric patients, but is also increasingly being utilized to assess symptom burden[11]. Elevated PSQ-SRDB scores were defined as scores ≥0.33.

The OSA-18 is an 18-item disease-specific quality-of-life questionnaire assessing symptoms in domains of sleep disturbance (range 4-28), physical suffering (range 4-28), emotional distress (range 3-21), daytime problems (range 3-21) and caregiver concerns (range 4-28). Total scores

range from 18 to 126 with higher scores indicating worse quality of life. Scores ≥ 60 are considered to indicate a moderate or more severe impact on disease-specific quality of life.

The Pediatric Quality of Life Inventory (PedsQL): The PedsQL is a generic measure of global quality of life comprised of 23 items across four Generic Core Scales: Physical Functioning (8 items); Emotional Functioning (5 items); Social Functioning (5 items); and School Functioning (5 items). A total Score and two Summary Scores (the Psychosocial Health Summary Score and the Physical Health Summary Score), are calculated. Core scores range from 0 to 100, with higher scores indicative of better quality of life. It was completed by the caregiver as well as children themselves (> 5 years).

Child Opportunity Index (COI): Census-level neighborhood socioeconomic status (SES) index was estimated from the Child Opportunity Index (COI 2.0) after geocoding each participant's current residential address using the US Census Tract database and Geographic Information System software (GIS) ArcGIS Pro 2.8.7 software (Environmental Systems Research Institute, Redlands, CA). The COI is comprised of 29 variables across 3 domains (education, health/environment, and social/economic opportunity) and is calculated using data from several sources (diversitydatakids.org).[12] Nationally adjusted total COI scores range from 0 to 100 and are categorized as: very low (0-20), low (20-40), moderate (40-60), high (60-80), and very high (80-100). Higher COI scores indicate more advantageous neighborhoods.

Randomization

At the end of the baseline visit, participants were randomized to early adenotonsillectomy (eAT, within 4 weeks) or Watchful Waiting with Supportive Care (WWSC). Randomization was stratified by the following factors within site: age (≤ 5 years vs > 5 years); overweight status (body mass index [BMI] >85%ile); and race (African American vs other). These factors were identified to likely influence treatment responses. Given the overall target sample size of 460 and a relative large number of strata (8 strata within each of the 6 sites), a dynamic randomization method, Pocock and Simon's minimization method[13], was implemented in the Data Management System, to ensure treatment arms are balanced with respect to these factors as well as for the number of subjects in each group.

Sleep Education Intervention Materials

Educational material on healthy sleep habits were provided to each child at the baseline visit after research data were collected. Standardized materials recommended by the NIH and pediatric professional sleep societies were used to reinforce optimal sleep, and educational play was also encouraged by providing take-home materials, including: *Sleep in Pre-schoolers (3-5 years); Sleep in School-Aged Children (6-12 years); Garfield Star Sleeper Fun Pad.*

Watchful Waiting with Supportive Care (WWSC)

Children were referred for appropriate usual care for management of comorbidities (e.g., poorly controlled asthma, allergies, etc.) based on initial history and exam. At the end of the trial they were scheduled for re-evaluation by ENT.

Surgical Intervention

In addition to receiving information addressing healthy sleep habits, children were referred for adenotonsillectomy. Total tonsillectomy and removal of obstructing adenoid tissue were performed by or under the supervision of board-certified Otolaryngologists qualified to treat pediatric patients with obstructive sleep disordered breathing. All physicians attested to reviewing study-specific training materials that outlined the procedures for surgical documentation and quality assurance, including documentary photographs (detailed under Quality Control). Surgical procedures included inspection and palpation of the palate to assure there was no evidence of a sub-mucus cleft, inspection of the nasopharynx with grading of the adenoid tissue as mildly (0-33%), moderately (34-66%) or severely (67-100%) obstructing the posterior choanae and removal of the obstructing portions of adenoid tissue. The tonsil size was then assessed bilaterally using a standardized 1-4 scale, all tonsil tissue was removed, and hemostasis was obtained. Removal of the lymphoid tissue was completed using cold dissection, monopolar electro cautery, microdebrider, radiofrequency/coblation or any other recognized surgical technique.

Quality Control

Surgical Subcommittee

A Surgical Subcommittee included ENT leaders from each of the clinical sites. The subcommittee was responsible for standardization of surgery and documentation of techniques, perioperative care, and complications. On an ongoing basis, the subcommittee reviewed data on surgical

efficacy obtained from the intra-operative data sheets completed by the surgeons, as well as samples of intra-operative photographs and all surgical AEs.

Intra-operative photographs were obtained on a 10% sample of participants undergoing AT. Specifically, at each clinical site, the oropharynx and nasopharynx of each tenth consecutive participant undergoing AT was photographed pre- and post-adenotonsillectomy in the operating room under general anesthesia after being orally intubated. Using a mouth-gag to optimize visualization, non-magnified digital photos of the oropharynx before and after tonsillectomy, and of the nasopharynx with the palate retracted using an angled telescope, were taken before and after adenoidectomy. De-identified photos and intra-operative assessments of tonsillar size were transmitted to the Surgery Subcommittee Director to ensure that the grading system for tonsil and adenoid size was consistently applied by all surgeons and that adequate tissue removal was performed. Any noted discrepancies in scoring or inadequate tissue removal were be discussed with the specific institution's lead otolaryngologists within ten working days of photo receipt.

Neurobehavioral Testing Quality Control

The Neurobehavioral Subcommittee was responsible for ensuring uniform standards of testing procedures, as well as oversaw staff training and certification, for neurobehavioral assessments. The Subcommittee, consisting of two board certified psychologists, oversaw the training and certification of blinded test administrators, audited a sample of records and, on an ongoing basis, examined summary reports providing the distributions of neurobehavioral indices by site and technician, monitoring missing values, range checks, and for identifying and examining sources of unusual values such as extreme ratings on every item of a behavior checklist. A checklist of the child and caregiver assessments was used to ensure that assessments were administered and are given in the correct order and to make notes regarding reasons for invalid or missing assessments. For each measure, the test administrator indicated whether the measure was completed and was considered valid, and commented on reasons for considering a measure invalid or not completed. An audit-trail was maintained to log any problems or deviations from suggested procedures, including details of errors and action taken, dates, and the individuals involved. Periodic conference calls were held by the Neurobehavioral Quality Control Subcommittee Directors with test administrators to monitor progress and address procedural questions.

Blinding

Use of a surgical pediatric intervention prevented double blinding. To minimize biases related to unblinding of investigators and study personnel, the study: (1) randomized participants only after

all baseline data were collected; (2) restricted access to randomization information to a scheduler and the surgeon; and (3) participants were asked not to discuss any aspects of treatment with other personnel. The following personnel were blinded: Local PI (unless a surgeon); PSG technologists; research staff responsible for neurobehavioral testing, blood pressure, and anthropometry; additional staff as needed (clinical research nurses, etc.). Unblinded personnel at each site included the pediatric otolaryngologist and one staff member responsible for all tasks related to following intervention-specific protocols (e.g., surgical follow-up) and phone interviews for adverse events and health care utilization.

Adverse Event Adjudication

Adverse events were defined as any unfavorable or unintended sign, symptom, or disease occurring in a participant at any stage following consent. Formal adverse event surveillance was undertaken using a structured interview framework conducted by unblinded research coordinators during each study visit and monthly caregiver phone call. In addition, research coordinators were trained to ensure documentation of spontaneously reported adverse events.

Adverse events were adjudicated as follows:

- Events were considered serious if they resulted in death, were immediately life-threatening, required inpatient hospitalization or prolongation of hospitalization, resulted in persistent or significant disability or incapacity, resulted in congenital anomaly, or otherwise jeopardized the participant's health and required medical or surgical intervention to prevent one of the other outcomes listed.
- Events were considered unexpected if the nature, severity, or frequency was not consistent with the known or foreseeable risk of adverse events described in protocol-related documents; or, the expected natural progression of any underlying disease, disorder, or condition of the participants experiencing the event and the participants predisposing risk factor profile for the event. Prior to study initiation, PATS investigators developed a list of mild, expected events that were not captured as adverse events unless they exceeded the associated description of severity. For example, post-operative throat pain lasting <21 days and not requiring intravenous hydration or unscheduled medical evaluation or treatment was expected to occur in every child undergoing AT, and was not captured as an adverse event.
- Events were considered related to the study if adjudicated as possibly, probably, or definitely related. Unrelated events were those that had no temporal association to study

testing, had an alternative etiology established, and did not follow the known pattern of response to the study test. Related adverse events were further classified as being related to AT vs related to other study procedures.

In addition to the adjudication categories listed above, events were graded on a severity scale of 1 (asymptomatic or mild symptoms requiring no indication other than over the counter medication) to 5 (resulting in death). The severity grade assigned to each event was used for descriptive purposes only, and did not impact whether the event required expedited reporting.

A tiered approach to adjudication was used, whereby unblinded staff at the Data Coordination Center (DCC) completed all aspects of adjudication if the event was clearly grade 1 or grade 2 in severity, and was not a suspected serious adverse event. All potentially serious adverse events and/or all adverse events that were potentially severity grade 3 or higher were adjudicated by an independent medical monitor. In addition, all adverse events that were potentially related to AT were reviewed by the surgical subcommittee.

During the adjudication process, a MedDRA 'lowest level term' was assigned to each adverse event. Lowest-level terms were then grouped into clinically-relevant categories by the trial investigators for descriptive purposes. For example, all ADHD events were combined into one category, while remaining behavioral events such as anxiety were combined into a separate category.

Surgical Complications

In addition to adverse event monitoring described above, complications resulting from surgery beyond the immediate post-operative period were documented during a structured post-surgery phone interview and by reports obtained from the caregiver during interim follow-up, supplemented by records, as appropriate. The Surgical Subcommittee was charged with reviewing any evidence of surgical complications that exceeded expectation of usual care, or any site experiencing excessive problems.

Treatment Failures

Research coordinators were trained to identify signs and/or symptoms that could potentially indicate treatment failure regardless of the arm assignment. "Treatment failure" was defined as a condition or situation that is observed during either routine interim follow-up phone calls or during

any study visits or clinical visits, or discovered in EMR surveillance, or from other parent or physician contact, and indicated a potential need to change treatment. Once identified as a potential treatment failure, the research coordinator completed written structured forms (obtaining additional information from the medical chart or caregiver), and notified the PI, who made any decisions required for the participant's clinical safety and need for further follow-up/referral (e.g., back to the referring physician). This information was provided to an Independent Medical Monitor who determined if the event met Treatment Failure criteria. This information, including specific reasons for failure including why a physician involved in the child's care determined alternative therapy and which alternative therapy was recommended, was documented on case report forms.

Missing Data Analysis

The primary and secondary analyses utilized a mixed effects model for repeated measurements, which accounts for missing data using a maximum-likelihood-based approach. These methods have been demonstrated to be superior to complete-case analysis or single-imputation methods such as the last observation carried forward, and have been recommended as a preferred method for analyzing clinical trials with incomplete longitudinal data[14,15]. The mixed effects models incorporate all available data from the three study visits, including observations from study participants who had at least one measurement. The mixed effects modeling will yield valid inferences provided that the data are missing at random (MAR)[16]; that is, the missingness is independent of the unobserved measurements conditional on the observed data. When the parameters in the outcome model are functionally independent of the parameters in the missingness process, MAR is ignorable, and likelihood-based inference for the outcome model remains valid when the missing data mechanism is ignored. Our primary and secondary analysis models adjusted for treatment arm, for study site and for the stratification factors of age (> 5 years), overweight status (BMI > 85th percentile), and race (Black/African American), and would yield valid inferences so long as the outcome data are MAR conditional on the covariates included in the model. Analyses are also presented without adjustment for stratification factors used in randomization.

As a sensitivity analysis, we also used multiple imputation by chained equations to re-analyze the two co-primary endpoints under a less restrictive MAR assumption. The imputation model incorporated additional baseline characteristics (demographic information [sex, continuous age, maternal educational attainment, annual household income < $30,000, childhood opportunity index]; medical and physical exam information [continuous BMI z-score, BMI weight category,

maximum tonsil grade, urinary cotinine levels, diagnosed asthma, current ADHD medication]; SDB symptoms and PSG information [PSQ-SRBDS score, ESS total score, OSA-18, apnea hypopnea index, oxygen desaturation index, arousal index]; behavioral information [CBCL total problems T-score]; and enrollment during the COVID-19 pandemic) as well as the study site, treatment arm, and available longitudinal outcome measures. We generated 30 imputed datasets and fit the primary analysis mixed effects models to each of these datasets. Point estimates and variances from the fitted models were then combined using the standard Rubin method[17].

Stratified Analysis by Subgroups

The following mixed effects model was used to examine subgroup differences with respect to each of the two primary endpoints and stratification factors:

$$Y_{ij} = \beta_0 + \beta_1 T6_{ij} + \beta_2 T12_{ij} + \beta_3 T6_{ij} \times eAT_{ij} + \beta_4 T12_{ij} \times eAT_{ij} + \beta_5 Z_{ij} + \beta_6 T6_{ij} \times Z_{ij} + \beta_7 T12_{ij} \times Z_{ij} + \beta_8 T6_{ij} \times eAT_{ij} \times Z_{ij} + \beta_9 T12_{ij} \times eAT_{ij} \times Z_{ij} + b_i + E_{ij},$$

where subgroups are defined by levels of the variable Z_{ij} and where additional adjustment was made for the main effects of the stratification factors and study site (not shown in the model). β_9 captures treatment effect heterogeneity across levels of Z_{ij}: in the stratum with $Z_{ij} = 0$, the 12-month intervention effect is given by β_4, while in the stratum with $Z_{ij} = 1$, the 12-month intervention effect is given by $\beta_4 + \beta_9$.

Stratification factors were: age (> 5 years old), sex (male vs. female), racial/ethnic minority status (minority vs. non-minority), obesity status (obese vs. overweight/healthy weight/failure to thrive), second hand smoke exposure (urinary cotinine levels ≥ 5 ng/mL vs. < 5 ng/m), diagnosed asthma (yes vs. no), high maternal education (4-year college education or greater vs. some college or less), low annual household income (<$30,000 vs. ≥$30,000), childhood opportunity index below the national median level (<50 vs ≥ 50), elevated CBCL total problems t-score at enrolment (≥ 60 vs. < 60), elevated PSQ-SRBDS score (≥ 0.33 vs. < 0.33), elevated ESS total score (≥ 10 vs. < 10), elevated OSA-18 (≥ 60 vs. < 60), enlarged tonsil size (tonsillar hypertrophy grade III/IV vs. II), and AHI (<1 vs ≥1).

Assessment of the Impact of the COVID-19 Pandemic

On March 11, 2020, the World Health Organization declared coronavirus 19 (COVID-19) to be a global pandemic, while the recruitment and follow-up of the PATS study were still ongoing. Taking this date as a proxy for the start of the pandemic, we classified each child's study visit as either

before or during the pandemic. We then performed a sensitivity analysis to assess the impact of the COVID-19 pandemic on the primary study findings.

To do so, we included a time-dependent indicator of COVID-19 onset into the mixed effects analysis of the BRIEF and GNG endpoints. We also considered the interaction of this indicator with the visit month (baseline, 6 months, or 12 months), the trial arm, and the visit month by trial arm interaction.

Supplementary Results

Cross-Over Rate

Of the 227 individuals randomized to watchful waiting with supportive care, 13 (5.7%) were documented to have crossed over and undergone surgery; conversely, 19 (8.2%; 13 active cross-overs, 6 unknown surgical status) of the 231 individuals randomized to early adenotonsillectomy had no recorded surgical intervention.

Treatment Failures

A total of six children in the WWSC were adjudicated to be Treatment Failures, reflecting persistent and/or worsening snoring and SDB symptoms (including worsening of snoring, observed apneas, tonsillitis, and behavioral problems); these children subsequently pursued surgery following ENT evaluation.

Completeness of Data

ETable 1 summarizes the rates at which complete data were available for the two primary endpoints: the caregiver-reported BRIEF GEC T-score and the GNG sustained attention d-prime. At 12-months post-randomization, 392 participants across both the WWSC and eAT arms (85.6% of the analysis population) had complete information with respect to the BRIEF endpoint and 368 participants (80.3%) had complete information with respect to the GNG endpoint. These data completion proportions did not differ between trial arms (difference in proportion for the BRIEF endpoint: -1.5% [95% CI, -7.9% to 4.9%]; difference in proportion for the GNG endpoint: -1.4% [95% CI, -8.7% to 5.9%]).

ETables 3 and 4a,b show the results of the primary analysis using multiple imputation with equations.

Prevalence of clinically meaningful behavioral or sleep-related symptom scores at baseline, 6-months, and 12 months, by randomization arm

Upon enrollment into PATS, 109 children (23.9% of the primary analysis population) had caregiver reported BRIEF GEC T-scores ≥ 65, 127 children (28.2%) had CBCL total problem T-scores ≥60, 343 children (75.2%) had PSQ-SRBDS scores ≥ 0.33, 107 children (23.6%) had ESS scores ≥10, and 134 children (29.4%) had OSA-18 scores ≥ 60. These results, as well as all subsequent analyses/proportions, represent complete case analyses, i.e., are restricted to children with complete baseline (or, when relevant, 12 month) data with respect to the clinical measure of interest (or all components of the composite outcome of interest). In this case, the number of children with complete baseline information with respect to caregiver-reported BRIEF GEC T-scores, CBCL total problems T-scores, PSQ-SRBDS scores, mESS scores, and OSA-18 scores were 457, 451, 456, 454, and 456, respectively. All told, 356 study participants (79.8% of the 446 participants with complete baseline SRBD-related symptom information across all five measures) had at least one marker of clinically significant SRBD-related symptoms/morbidity at baseline: 179 (79.9%) in the WWSC arm and 177 (79.7%) in the eAT arm (RR,1.00; 95% CI, 0.91 to 1.10). (ETable 7).

Surgical Complications

Of 224 surgical procedures, no intra-operative complications were reported. Intra-operative blood loss was <25 cc in 99% of cases, with a maximum intra-operative blood loss of 30 cc.

A total of 30 Adverse Events were reported as possibly, probably or definitely related to Surgery. Five children required surgical intervention for post-operative bleeding. No deaths or cases of naso-regurgitation or velo-pharyngeal insufficiency were reported. (ETable 11)

Impact of the COVID-19 Pandemic

Two hundred seventy-eight patients completed their study participation prior to the onset of the pandemic, meaning either that they completed all three study visits, or that they were lost to follow up, withdrew consent, or expressed that they were not interested in participating in the study further before March 11, 2020. One hundred twenty-four patients had scheduled study visits both prior to and during the pandemic, and 54 patients enrolled in and completed the trial entirely during the pandemic.

Prior to the onset of the COVID-19 pandemic, there was a decrease in mean caregiver-reported BRIEF GEC T-scores in both the eAT arm and the WWSC arm, but this decrease did not differ by treatment arm. The same pattern persisted after the pandemic; the estimated difference in the effect of eAT on the BRIEF endpoint during vs. prior to the pandemic was 0.2 [95% CI: -4.7 to 5.1]. For the GNG endpoint, the estimated difference in the effect of eAT during vs. prior to the pandemic was -0.1 [95% CI, -0.8 to 0.5]. Our analysis did not reveal any evidence to suggest a differential effect of eAT on the BRIEF endpoint or the GNG endpoint between the pre-pandemic and during-pandemic periods.

Supplementary Tables

ETable 1. Data availability proportions for each of the co-primary endpoints, shown by randomization group

	Early Adenotonsillectomy (n = 231)	Watchful Waiting (n = 227)
Caregiver BRIEF GEC T-score		
Baseline	230 (99.6%)	227 (100%)
6-month visit	187 (81.0%)	189 (83.3%)
12-month visit	196 (84.8%)	196 (86.3%)
Go/No-Go sustained attention d'		
Baseline	229 (99.1%)	222 (97.8%)
6-month visit	161 (69.7%)	161 (70.9%)
12-month Tisit	184 (79.7%)	184 (81.1%)

Abbreviations: BRIEF, Behavior Rating Inventory of Executive Function; GEC, global executive composite.

ETable 2a. Mixed effects modeling results for the BRIEF GEC t-score endpoint

	Point Estimate (95% CI) (n = 458)
Early Adenotonsillectomy	
Mean at baseline	53.0 (49.6 to 56.6)
Change from baseline to 6 months	-2.1 (-3.3 to -0.9)
Change from baseline to 12 months	-3.0 (-4.2 to -1.8)
Watchful Waiting	
Mean at baseline	53.7 (50.3 to 57.1)
Change from baseline to 6 months	-1.3 (-2.5 to -0.1)
Change from baseline to 12 months	-2.1 (-3.3 to -0.9)

Abbreviations: BRIEF, Behavior Rating Inventory of Executive Function; CI, confidence interval; GEC, global executive composite.

ETable 2b. Mixed effects modeling results for the sustained attention Go/No-Go d' endpoint

	Point Estimate (95% CI) (n = 455)
Early Adenotonsillectomy	
Mean at baseline	2.1 (1.8 to 2.4)
Change from baseline to 6 months	0.2 (0.1 to 0.4)
Change from baseline to 12 months	0.2 (0.1 to 0.4)
Watchful Waiting	
Mean at baseline	2.2 (1.9 to 2.5)
Change from baseline to 6 months	0.1 (0.0 to 0.3)
Change from baseline to 12 months	0.2 (0.0 to 0.3)

Abbreviations: CI, confidence interval.

ETable 3. Primary outcome measures under multiple imputation of the missing outcome data[a]

	Mean (SD)							
	Early Adenotonsillectomy			Watchful Waiting				
	Baseline	12 mo.	Change from Baseline to 12 mo.	Baseline	12 mo.	Change from Baseline to 12 mo.	Effect Size – Difference in 12 mo. Changes (95% CI)[b]	P Value[b]
Caregiver BRIEF GEC T-score[c]	55.3 (12.2)	52.3 (11.4)	-3.0 (9.5)	56.0 (12.5)	53.8 (11.3)	-2.2 (9.0)	-0.85 (-2.57 to 0.86)	0.33
Go/No-Go sustained attention d'[d]	2.0 (1.1)	2.2 (1.1)	0.2 (1.2)	2.1 (1.0)	2.3 (1.0)	0.2 (1.2)	0.05 (-0.16 to 0.26)	0.64

Abbreviations: BRIEF, Behavior Rating Inventory of Executive Function; CI, confidence interval; GEC, global executive composite; SD, standard deviation.

[a]The imputation model incorporated additional baseline characteristics (demographic information, medical and physical exam information, sleep-related breathing disorder and polysomnography information, behavioral information, and enrollment during the COVID-19 pandemic) as well as the study site, treatment group, and available longitudinal outcome measures. See Missing Data Analysis subsection of the Supplementary Appendix for more details.

[b]The estimated between-group difference in mean change from baseline to 12 months and the corresponding *P* value are from a linear mixed-effects model with prespecified adjustment for stratification factors (age ≥ 5, overweight/ status, Black/African American) and site effect.

[c]The BRIEF GEC section comprises summary measures of behavioral regulation, emotion regulation, and cognitive regulation (BRIEF-2, for children ages 5 to 18) or inhibitory self-control, flexibility, and emergent metacognition (BRIEF-P, for preschool-aged children). Caregiver scores ranged from 33 to 102, with higher scores indicating worse functioning.

[d]The Go/No-Go sustained attention d' is a signal detection measure that combines a child's true positive rate on an attention task (correct response to the target stimuli) with their false alarm rate (incorrect response to the non-target stimuli). Scores ranged from -0.9 to 4.5, with higher scores indicating greater sustained attention.

ETable 4a**.** Mixed effects modeling results for the BRIEF GEC T-score endpoint under multiple imputation of the missing outcome data

	Point Estimate (95% CI)
Early Adenotonsillectomy	
Mean at Baseline	52.8 (49.4 to 56.2)
Change from baseline to 6 months	-1.9 (-3.2 to -0.7)
Change from baseline to 12 months	-3.0 (-4.3 to -1.8)
Watchful Waiting	
Mean at Baseline	53.5 (50.1 to 56.9)
Change from baseline to 6 months	-1.3 (-2.5 to -0.1)
Change from baseline to 12 months	-2.2 (-3.4 to -1.0)
Abbreviations: BRIEF, Behavior Rating Inventory of Executive Function; CI, confidence interval; GEC, global executive composite.	

ETable 4b**.** Mixed effects modeling results for the sustained attention Go/No-Go d' endpoint under multiple imputation of the missing outcome data

	Point Estimate (95% CI)
Early Adenotonsillectomy	
Mean at Baseline	2.1 (1.8 to 2.4)
Change from baseline to 6 months	0.2 (0.0 to 0.4)
Change from baseline to 12 months	0.2 (0.1 to 0.4)
Watchful Waiting	
Mean at Baseline	2.2 (1.9 to 2.5)
Change from baseline to 6 months	0.1 (0.0 to 0.3)
Change from baseline to 12 months	0.2 (0.0 to 0.3)
Abbreviations: CI, confidence interval.	

ETable 5. Polysomnography measures

	Early Adenotonsillectomy[a] (n=231)			Watchful Waiting[a] (n=227)				
	Baseline (n=231)	12 mo. (n=154)	Change in Baseline to 12 mo. (n=154)	Baseline (n=227)	12 mo. (n=152)	Change in Baseline to 12 mo. (n=152)	Unadjusted Effect Size – Difference in 12 mo. Changes[b] (95% CI) (n=458)	Adjusted Effect Size – Difference in 12 mo. Changes[b] (95% CI) (n=458)
Apnea-hypopnea index[c]	0.5 (0.1-1.1)	0.3 (0.1-0.6)	-0.2 (-0.6-0.2)	0.6 (0.3-1.1)	0.7 (0.2-1.8)	0.1 (-0.3-0.9)	-0.61[d] (-0.93 to -0.28)	-0.60[d] (-0.93 to -0.27)
Arousal index[e]	5.2 (4.3-6.7)	5.0 (4.1-6.5)	0.0 (-1.3-0.9)	5.4 (4.3-6.8)	5.7 (4.5-7.2)	0.4 (-0.8-1.5)	-0.09[d] (-0.17 to -0.01)	-0.09[d] (-0.17 to 0.00)
% N1 Sleep[f]	7.1 (3.2)	6.6 (3.4)	-0.5 (4.1)	7.6 (3.7)	7.6 (3.7)	0.2 (4.1)	-0.61 (-1.48 to 0.26)	-0.59 (-1.47 to 0.28)
% N2 Sleep[f]	44.6 (8.0)	47.0 (6.9)	2.4 (7.7)	44.6 (7.5)	45.3 (6.3)	0.7 (7.6)	1.71 (0.09 to 3.33)	1.73 (0.10 to 3.36)
% N3 Sleep[f]	31.0 (7.9)	29.1 (7.4)	-1.8 (8.7)	30.5 (7.7)	29.6 (6.6)	-1.3 (8.2)	-0.84 (-2.64 to 0.96)	-0.84 (-2.64 to 0.96)
% REM Sleep[g]	17.2 (4.7)	17.4 (4.5)	-0.1 (5.0)	17.2 (4.7)	17.5 (4.5)	0.4 (5.2)	-0.28 (-1.35 to 0.80)	-0.30 (-1.38 to 0.78)

Abbreviations: CI, confidence interval; REM, rapid eye movement.

[a]Means (SD) or medians (interquartile range) for heavily skewed data.
[b]The estimated between-group differences in mean change from baseline to 12 months are from linear mixed-effects models. The adjusted analysis controls for stratification factors (age > 5, overweight status, Black/African American) and site effects, while the unadjusted analysis does not.
[c]The apnea-hypopnea index (AHI) is the average number of apnea or hypopnea (hypopneas with ≥ 3% oxygen desaturation or arousal) events per hour of sleep by polysomnography, with higher scores indicating more severe obstructive sleep apnea.

[d]Calculated under log-transformation of the endpoint.
[e]The arousal index is the average number of arousals per hour of sleep by polysomnography. Arousal indices ranged from 1.4 to 18.9, with higher scores indicating greater sleep disruption.
[f]Sleep stages N1, N2, and N3 are all considered non-rapid eye movement sleep, and each sleep stage is progressively deeper. The percentage of time spent in N1 sleep ranged from 1.03% to 25.8%, the percentage of time spent in N2 sleep ranged from 20.4% to 82.30%, and the percentage of time spent in N3 sleep ranged from 0.0% to 61.2%.
[g]The percentage of time spent in rapid eye movement (REM) sleep ranged from 0.0% to 31.8%.

ETable 6. Analysis of primary and key secondary outcome measures, with and without adjustment for stratification factors

	Mean (SD)								
	Early Adenotonsillectomy (n=231)			Watchful Waiting (n=227)					
	Baseline	12 mo.	Change from Baseline to 12 mo.	Baseline	12 mo.	Change from Baseline to 12 mo.	Unadjusted Effect Size – Difference in 12 mo. Changes (95% CI)[a] (n=458)	Adjusted Effect Size – Difference in 12 mo. Changes (95% CI)[a] (n=458)	P Value[a]
Co-primary outcomes									
Caregiver BRIEF GEC T-score[b]	55.3 (12.2) [n=230]	52.1 (11.3) [n=196]	-3.1 (9.4) [n=195]	56.0 (12.5)	53.7 (11.2) [n=196]	-1.9 (8.6) [n=196]	-0.98 (-2.68 to 0.72)	-0.96 (-2.66 to 0.74)	0.27
Go/No-Go sustained attention d'[c]	2.0 (1.1) [n=229]	2.2 (1.1) [n=184]	0.2 (1.2) [n=182]	2.1 (1.0) [n=222]	2.3 (1.0) [n=184]	0.1 (1.2) [n=182]	0.05 (-0.17 to 0.27) [n=455]	0.05 (-0.18 to 0.27) [n=455]	0.68
Secondary outcomes							FDR-Adjusted CI[d]	FDR-Adjusted CI[d]	FDR-Adjusted *P* Value[d]
Pegboard dexterity (average)[e]	32.5 (14.6) [n=227]	27.4 (9.1) [n=187]	-5.3 (7.4) [n=183]	32.8 (11.9)	26.7 (6.4) [n=187]	-5.9 (8.0) [n=187]	0.82 (-0.85 to 2.49)	0.76 (-0.92 to 2.43)	0.37
Caregiver BRIEF subscales[b]									
Emotional control T-score	54.2 (11.8) [n=230]	52.0 (11.5) [n=196]	-2.0 (9.8) [n=195]	54.8 (13.0)	53.5 (11.5) [n=196]	-0.9 (9.4) [n=196]	-0.94 (-2.98 to 1.10)	-0.92 (-2.96 to 1.12)	0.37

Inhibit T-score	55.7 (12.2) [n=230]	52.3 (11.2) [n=196]	-3.1 (9.5) [n=195]	56.5 (12.5)	53.7 (11.8) [n=196]	-2.7 (8.6) [n=196]	-0.37 (-2.30 to 1.56)	-0.36 (-2.30 to 1.57)	0.74
Plan/organize T-score	52.5 (11.6) [n=230]	49.9 (10.4) [n=196]	-2.6 (10.4) [n=195]	53.4 (11.6)	50.8 (10.5) [n=196]	-2.1 (10.0) [n=196]	-0.24 (-2.35 to 1.86)	-0.22 (-2.34 to 1.89)	0.82
Shift T-score	53.2 (11.6) [n=230]	50.9 (10.8) [n=196]	-2.4 (9.6) [n=195]	53.4 (11.4)	52.4 (10.9) [n=196]	-0.6 (9.4) [n=196]	-1.47 (-3.46 to 0.52)	-1.42 (-3.41 to 0.57)	0.17
Working memory T-score	55.6 (11.5) [n=230]	52.7 (11.0) [n=196]	-2.7 (8.7) [n=195]	56.0 (11.6)	54.2 (10.8) [n=196]	-1.4 (9.0) [n=196]	-1.16 (-3.06 to 0.73)	-1.15 (-3.05 to 0.74)	0.23
Caregiver-reported CBCL[f]									
Total problems T-score	53.0 (11.0) [n=227]	48.4 (10.8) [n=186]	-4.5 (9.0) [n=183]	53.3 (11.3) [n=224]	51.6 (10.9) [n=182]	-1.4 (7.5) [n=182]	-3.10 (-4.92 to -1.29) [n=454]	-3.09 (-4.90 to -1.28) [n=454]	< 0.001
Externalizing problems T-score	51.1 (10.8) [n=227]	48.1 (10.3) [n=186]	-3.2 (8.5) [n=183]	51.2 (11.7) [n=224]	49.6 (11.2) [n=182]	-1.6 (8.2) [n=182]	-1.56 (-3.36 to 0.24) [n=454]	-1.54 (-3.34 to 0.26) [n=454]	0.09
Internalizing problems T-score	51.8 (11.2) [n=227]	47.8 (10.9) [n=186]	-3.8 (10.0) [n=183]	52.1 (11.3) [n=224]	51.1 (11.0) [n=182]	-0.7 (8.3) [n=182]	-3.07 (-5.09 to -1.05) [n=454]	-3.05 (-5.07 to -1.04) [n=454]	0.003
Attentional problems T-score	57.5 (8.2) [n=227]	55.2 (6.6) [n=186]	-2.3 (7.1) [n=183]	57.3 (7.8) [n=183]	56.0 (6.7) [n=182]	-1.1 (5.8) [n=182]	-1.19 (-2.55 to 0.17) [n=454]	-1.19 (-2.55 to 0.17) [n=454]	0.09

PSQ-SRBD scale[g]	0.5 (0.2) [n=229]	0.2 (0.2) [n=189]	-0.2 (0.2) [n=187]	0.5 (0.2)	0.4 (0.2) [n=193]	-0.1 (0.2) [n=193]	-0.16 (-0.20 to -0.12)	-0.16 (-0.20 to -0.12)	< 0.001
mESS score[h]	6.9 (4.7) [n=227]	5.0 (5.3) [n=188]	-1.8 (4.9) [n=184]	6.9 (4.6)	6.2 (5.1) [n=193]	-0.7 (4.4) [n=193]	-1.16 (-2.12 to -0.19)	-1.18 (-2.15 to -0.21)	0.01
OSA-18[i]	51.2 (15.7) [n=229]	35.6 (13.9) [n=188]	-15.8 (14.4) [n=186]	52.7 (17.4)	46.5 (17.3) [n=193]	-6.0 (14.6) [n=193]	-9.69 (-12.79 to -6.59)	-9.75 (-12.84 to -6.65)	< 0.001
Caregiver-reported PedsQL[j]									
Total score	75.9 (13.2) [n=229]	78.4 (16.0) [n=189]	2.1 (14.9) [n=187]	77.7 (12.8) [n=226]	75.0 (15.9) [n=193]	-2.6 (15.0) [n=193]	4.75 (1.43 to 8.07) [n=457]	4.76 (1.44 to 8.09) [n=457]	0.005
Physical score	79.5 (19.1) [n=229]	81.1 (21.9) [n=189]	0.7 (23.5) [n=187]	82.1 (16.3) [n=226]	76.4 (23.2) [n=193]	-5.3 (24.3) [n=193]	6.44 (1.20 to 11.67) [n=457]	6.53 (1.29 to 11.78) [n=457]	0.01
Psychosocial score	73.9 (13.6) [n=229]	77.0 (14.8) [n=189]	2.8 (14.2) [n=187]	75.3 (14.1) [n=226]	74.3 (14.5) [n=193]	-1.1 (13.3) [n=193]	3.89 (0.90 to 6.89) [n=457]	3.88 (0.89 to 6.88) [n=457]	0.01
Body mass index (%ile)	65.0 (30.0)	70.4 (27.4) [n=187]	5.1 (14.3) [n=187]	62.0 (32.1)	66.2 (31.4) [n=188]	3.2 (13.3) [n=188]	1.84 (-0.91 to 4.60)	1.86 (-0.88 to 4.60)	0.18
Systolic blood pressure (%ile)	63.6 (23.0) [n=218]	58.6 (27.9) [n=180]	-4.5 (28.9) [n=173]	57.6 (27.0) [n=216]	61.6 (26.0) [n=184]	4.8 (31.4) [n=177]	-9.11 (-15.60 to -2.62) [n=451]	-9.02 (-15.49 to -2.54) [n=451]	0.006

Diastolic blood pressure (%ile)	55.5 (20.5) [n=218]	51.5 (22.2) [n=179]	-4.9 (22.5) [n=172]	52.2 (21.5) [n=216]	54.6 (22.1) [n=184]	2.2 (20.8) [n=177]	-6.61 (-11.70 to -1.53) [n=451]	-6.52 (-11.59 to -1.45) [n=451]	0.01
Heart rate (b.p.m.)	80.8 (9.1)	78.8 (8.9) [n=154]	-2.0 (7.7) [n=154]	81.2 (9.5)	79.7 (9.5) [n=152]	-1.8 (7.9) [n=152]	-0.30 (-2.21 to 1.61)	-0.23 (-2.13 to 1.67)	0.82
Apnea-hypopnea index[k]	Baseline, No. (%)	12 mo., No. (%)		Baseline, No. (%)	12 mo. No. (%)		Unadjusted Effect Size – Risk Difference at 12 mo., % (FDR-Adjusted CI)[d,l]	Adjusted Effect Size – Risk Difference at 12 mo., % (FDR-Adjusted CI)[d,l]	
AHI ≥ 3	2 (0.9%)	2/154 (1.3%)		1 (0.4%)	20/152 (13.2%)		-11.48 (-17.82 to -5.14) [n=306]	-11.17 (-17.46 to -4.89) [n=306]	< 0.001
AHI ≥ 5	0 (0.0%)	0/154 (0.0%)		0 (0.0%)	11/152 (7.2%)		-7.11 (-11.74 to -2.48) [n=306]	-7.14 (-11.76 to -2.52) [n=306]	0.002

Abbreviations: BRIEF, Behavior Rating Inventory of Executive Function; CBCL, Child Behavior Checklist; CI, confidence interval; FDR, false discovery rate; GEC, global executive composite; mESS, modified Epworth Sleepiness Scale; NA, not applicable; OSA, obstructive sleep apnea; PedsQL, Pediatric Quality of Life Inventory; PSQ-SRBD, Sleep-Related Breathing Disorder scale of the Pediatric Sleep Questionnaire; SD, standard deviation.

[a]The estimated between-group differences in mean change from baseline to 12 months and the corresponding *P* value are from linear mixed-effects models. The adjusted analysis controls for stratification factors (age >5, overweight status, Black/African American) and site effects, while the unadjusted analysis does not. The reported p value is from the adjusted analysis. Both analyses incorporate information from all study participants with at least one measurement across the three study visits.

[b]The BRIEF GEC section comprises summary measures of behavioral regulation, emotion regulation, and cognitive regulation (BRIEF-2, for children ages 5 to 18) or inhibitory self-control, flexibility, and emergent metacognition (BRIEF-P, for preschool-aged children). Caregiver GEC scores ranged from 33 to 102, with higher scores indicating worse functioning. Five subscales are shared between both the BRIEF-2 and BRIEF-P version. Scores ranged from 35 to 96 on the BRIEF emotional control subscale, from 36 to 93 on the BRIEF inhibit subscale, from 32 to 94 on the BRIEF plan/organize subscale, from 37 to 87 on the BRIEF shift subscale, and from 36 to 98 on the

BRIEF working memory subscale. On each scale, higher scores again indicate worse functioning. Thirty-five patients were lost to follow up at 12 months in the early adenotonsillectomy (eAT) group and 31 in the watchful waiting with supportive care (WWSC) group. All 458 patients contributed information to the difference in differences analysis.

[c]The Go/No-Go sustained attention d' is a signal detection measure that combines a child's true positive rate on an attention task (correct response to the target stimuli) with their false alarm rate (incorrect response to the non-target stimuli). Scores ranged from -0.9 to 4.5, with higher scores indicating greater sustained attention. Forty-seven patients were lost to follow up at 12 months in the eAT group and 43 in the WWSC group. Three patients were excluded from the difference in differences analysis.

[d]*P* values for all secondary outcomes are adjusted so that the set of hypotheses with p values below 0.05 exactly corresponds to the set of hypotheses that would be rejected under the Benjamini and Hochberg (1995) procedure for controlling the FDR at 0.05[18]. This means that, among all effects considered statistically significant at the 0.05 level (using the adjusted *P* values), we would expect 5% to be truly null. 22 analyses were included in the multiplicity adjustment set. The 95% confidence intervals for all 22 of these secondary endpoints are adjusted for multiple comparisons following the procedure of Benjamini and Yekutieli (2005)[19] with a nominal coverage level of 97%.

[e]The average of NIH Toolbox 9-hole pegboard dexterity test times (in seconds) from both the dominant and non-dominant hand. Average times ranged from 15.5 seconds to 118 seconds, with higher average times indicating lower manual dexterity. Forty-four patients were lost to follow up at 12 months in the eAT group and 40 in the WWSC group. All 458 patients contributed information to the difference in differences analysis.

[f]Scores ranged from 28 to 86 on the CBCL externalizing problems scale, from 29 to 88 on the CBCL internalizing problems scale, from 50 to 97 on the CBCL attentional problems scale, and from 24 to 84 on the CBCL total problems scale (comprising all three of the previous scales). On each scale, higher scores indicating greater problems. Forty-five patients were lost to follow up at 12 months in both the eAT group and the WWSC group. Four patients were excluded from the difference in differences analysis.

[g]Scores on the PSQ-SRBD range from 0 to 1, with higher scores indicating greater severity. Forty-two patients were lost to follow up at 12 months in the eAT group and 34 in the WWSC group. All 458 patients contributed information to the difference in differences analysis.

[h]Scores on the mESS range from 0 to 24, with higher scores indicating greater sleepiness. Forty-three patients were lost to follow up at 12 months in the eAT group and 34 in the WWSC group. All 458 patients contributed information to the difference in differences analysis.

[i]Scores on the OSA-18 quality of life survey range from 18 to 126, with higher scores indicating a greater negative effect of sleep-disordered breathing on quality of life. Forty-three patients were lost to follow up at 12 months in the eAT group and 34 in the WWSC group. All 458 patients contributed information to the difference in differences analysis.

[j]The PedsQL total score comprises performance on four subscales: emotional functioning, social functioning, and school functioning (summarized by the psychosocial functioning score) and physical functioning (summarized by the physical functioning score). Scores on all scales range from 0 to 100, with higher scores indicating better quality of life. Forty-two patients were lost to follow up at 12 months in the eAT group and 34 in the WWSC group. One patient was excluded from the difference in differences analysis.

[k]The apnea-hypopnea index (AHI) is the average number of apnea or hypopnea (≥ 3% oxygen desaturation) events per hour of sleep, with higher scores indicating more severe obstructive sleep apnea. AHIs were rounded to the nearest tenth. An AHI of 3 or greater at 12 months indicates a progression of disease over the course of the study; an AHI of 5 or greater at 12 months indicates progression to moderate obstructive sleep apnea. Seventy-seven patients were lost to follow up at 12 months in the eAT group and 75 in the WWSC group. One hundred fifty-two patients were excluded from the risk difference analysis.

[I]The between-group difference in prevalence at 12 months was estimated using a linear regression model with robust standard errors, fit to all patients with complete baseline and 12-month information. Point estimates and confidence intervals adjust for baseline apnea-hypopnea index, stratification factors (age >5, overweight status, Black/African American) and site effect. The confidence intervals are also adjusted for multiple comparisons, with a nominal coverage level of 97%.

ETable 7. Prevalence of clinically meaningful behavioral and sleep-related symptom scores at baseline, 6-months, and 12 months, by randomization group

	Early Adenotonsillectomy, No. (%) (n=231)			Watchful Waiting, No. (%) (n=227)		
	Baseline	6 mo.	12 mo.	Baseline	6 mo.	12 mo.
BRIEF GEC T-score ≥ 65[a]	52/230 (22.6%)	33/187 (17.6%)	25/196 (12.8%)	57/227 (25.1%)	39/189 (20.6%)	35/196 (17.9%)
CBCL total problems T-score ≥ 60[b]	65/227 (28.6%)	29/168 (17.3%)	30/186 (16.1%)	62/224 (27.7%)	45/175 (25.7%)	44/182 (24.2%)
PSQ-SRBD scale ≥ 0.33[c]	176/229 (76.9%)	52/184 (28.3%)	48/189 (25.4%)	167/227 (73.0%)	123/187 (65.8%)	106/193 (54.9%)
mESS total score ≥ 10[d]	50/227 (22.0%)	25/184 (13.6%)	30/188 (16.0%)	57/227 (25.1%)	42/187 (22.5%)	44/193 (22.8%)
OSA-18 ≥ 60[e]	66/229 (28.8%)	13/184 (7.1%)	13/188 (6.9%)	68/227 (30.0%)	43/187 (23.0%)	41/193 (21.2%)
Frequent loud snoring[f]	125/229 (54.6%)	11/184 (6.0%)	15/188 (8.0%)	125/227 (55.1%)	81/187 (43.3%)	69/193 (35.8%)

Abbreviations: BRIEF, Behavior Rating Inventory of Executive Function; CBCL, Child Behavior Checklist; CI, confidence interval; GEC, global executive composite; mESS, modified Epworth Sleepiness Scale; NA, not applicable; OSA, obstructive sleep apnea; PSQ-SRBD, Sleep-Related Breathing Disorder scale of the Pediatric Sleep Questionnaire.

[a]The BRIEF GEC section comprises summary measures of behavioral regulation, emotion regulation, and cognitive regulation (BRIEF-2, for children ages 5 to 18) or inhibitory self-control, flexibility, and emergent metacognition (BRIEF-P, for preschool-aged children). Caregiver scores ranged from 33 to 102, with higher scores indicating worse functioning. A T-score of 65 or greater is considered potentially clinically elevated.

[b]The CBCL total problems summary scale comprises internalizing, externalizing, social, thought, and attention problems. Scores ranged from 24 to 84, with higher scores indicating greater emotional, social, and behavioral problems. A T-score of 60 or greater indicates that the child is at risk for clinical problem behaviors.

[c]Scores on the PSQ-SRBD range from 0 to 1, with higher scores indicating greater severity. A score of 0.33 or greater suggests a high risk for a pediatric sleep-related breathing disorder.
[d]Scores on the mESS range from 0 to 24, with higher scores indicating greater sleepiness. A score of 10 or greater represents excessive daytime sleepiness.
[e]Scores on the OSA-18 quality of life survey range from 18 to 126, with higher scores indicating a greater negative effect of sleep-disordered breathing on quality of life. A score of 60 or greater represents a moderate to severe negative effect.
[f]Snoring was assessed using item 1a of the OSA-18 quality of life survey, which uses a Likert scale to ask about the frequency of loud snoring over the last four weeks. Possible responses ranged from “none of the time” to “all of the time”, and loud snoring was considered frequent if it occurred “a good bit of the time”, “most of the time”, or “all of the time”.

ETable 8. Exploratory analysis of the prevalence of clinically meaningful behavioral and sleep symptom outcomes at 12 months, with and without adjustment for stratification factors

	Symptom Prevalence at 12 Months, No. (%)					
	Early Adenotonsillectomy (n=231)	Watchful Waiting (n=227)	Unadjusted Risk Difference, % (95% CI)[a]	Adjusted Risk Difference, % (95% CI)[a]	Unadjusted Relative Risk (95% CI)[b]	Adjusted Relative Risk (95% CI)[b]
BRIEF GEC T-score ≥ 65[c]	25/196 (12.8%)	35/196 (17.9%)	-4.4 (-10.8 to 1.9) [n=391]	-4.0 (-10.2 to 2.3) [n=391]	0.80 (0.52 to 1.24) [n=391]	0.83 (0.53 to 1.29) [n=391]
CBCL total problems T-score ≥ 60[d]	30/186 (16.1%)	44/182 (24.2%)	-7.9 (-14.9 to -0.9) [n=365]	-7.6 (-14.6 to -0.7) [n=365]	0.67 (0.47 to 0.95) [n=365]	0.66 (0.46 to 0.94) [n=365]
PSQ-SRBD scale ≥ 0.33[e]	48/189 (25.4%)	106/193 (54.9%)	-31.4 (-39.6 to -23.2) [n=380]	-30.9 (-39.1 to -22.8) [n=380]	0.43 (0.34 to 0.56) [n=380]	0.43 (0.33 to 0.55) [n=380]
mESS total score ≥ 10[f]	30/188 (16.0%)	44/193 (22.8%)	-6.2 (-13.2 to 0.9) [n=377]	-5.9 (-12.7 to 1.0) [n=377]	0.69 (0.47 to 1.01) [n=377]	0.65 (0.45 to 0.94) [n=377]
OSA-18 ≥ 60[g]	13/188 (6.9%)	41/193 (21.2%)	-13.2 (-19.3 to -7.2) [n=379]	-13.0 (-19.1 to -7.0) [n=379]	0.39 (0.22 to 0.67) [n=379]	0.36 (0.21 to 0.62) [n=379]
Frequent loud snoring[h]	15/188 (8.0%)	69/193 (35.8%)	-27.4 (-34.7 to -20.2) [n=379]	-27.3 (-34.5 to -20.1) [n=379]	0.22 (0.13 to 0.37) [n=379]	0.22 (0.13 to 0.38) [n=379]

Abbreviations: BRIEF, Behavior Rating Inventory of Executive Function; CBCL, Child Behavior Checklist; CI, confidence interval; GEC, global executive composite; mESS, modified Epworth Sleepiness Scale; NA, not applicable; OSA, obstructive sleep apnea; PSQ-SRBD, Sleep-Related Breathing Disorder scale of the Pediatric Sleep Questionnaire.

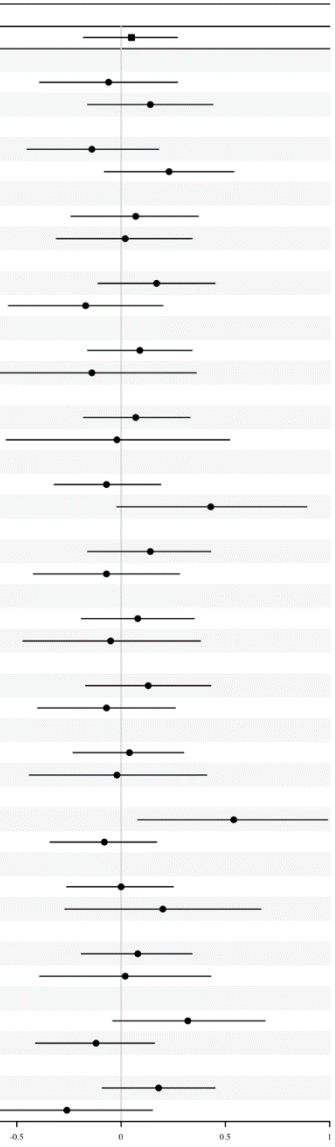
[a]Difference in prevalence at 12 months was estimated using a linear regression model with robust standard errors, fit to all patients with complete baseline and 12-month information. Point estimates and confidence intervals from the unadjusted model control for the value of the symptom scale at baseline only, while those from the adjusted model additionally control for stratification factors (age >5, overweight status, Black/African American) and site effect.
[b]Relative risk was estimated using a Poisson regression model with log link and robust standard errors, fit to all patients with complete baseline and 12-month information. Point estimates and confidence intervals from the unadjusted model control for the value of the symptom scale at baseline only, while those from the adjusted model additionally control for stratification factors (age >5, overweight status, Black/African American) and site effect.
[c]The BRIEF GEC section comprises summary measures of behavioral regulation, emotion regulation, and cognitive regulation (BRIEF-2, for children ages 5 to 18) or inhibitory self-control, flexibility, and emergent metacognition (BRIEF-P, for preschool-aged children). Caregiver scores ranged from 33 to 102, with higher scores indicating worse functioning. A T-score of 65 or greater is considered potentially clinically elevated.
[d]The CBCL total problems summary scale comprises internalizing, externalizing, social, thought, and attention problems. Scores ranged from 24 to 84, with higher scores indicating greater emotional, social, and behavioral problems. A T-score of 60 or greater indicates that the child is at risk for clinical problem behaviors.
[e]Scores on the PSQ-SRBD range from 0 to 1, with higher scores indicating greater severity. A score of 0.33 or greater suggests a high risk for a pediatric sleep-related breathing disorder.
[f]Scores on the mESS range from 0 to 24, with higher scores indicating greater sleepiness. A score of 10 or greater represents excessive daytime sleepiness.
[g]Scores on the OSA-18 quality of life survey range from 18 to 126, with higher scores indicating a greater negative effect of sleep-disordered breathing on quality of life. A score of 60 or greater represents a moderate to severe negative effect.
[h]Snoring was assessed using item 1a of the OSA-18 quality of life survey, which uses a Likert scale to ask about the frequency of loud snoring over the last four weeks. Possible responses ranged from “none of the time” to “all of the time”, and loud snoring was considered frequent if it occurred “a good bit of the time”, “most of the time”, or “all of the time”.

ETable 9a. Treatment effects for the caregiver-reported BRIEF GEC T-score in various subgroups

	Effect Estimate
All Patients	*-0.96 (-2.66, 0.74)*
Age (in years)	
5 or less years old	-0.76 (-3.30, 1.78)
6 or more years old	-1.11 (-3.41, 1.18)
Sex	
Female	-1.54 (-3.96, 0.88)
Male	-0.42 (-2.82, 1.99)
Racial/Ethnic Minority	
No	-0.53 (-2.87, 1.80)
Yes	-1.46 (-3.97, 1.05)
BMI Percentile	
≤ 85th BMI percentile	-1.82 (-3.97, 0.32)
> 85th BMI percentile	0.55 (-2.25, 3.35)
Obesity Status	
< 95th BMI percentile	-0.34 (-2.24, 1.56)
≥ 95th BMI percentile	-3.62 (-7.45, 0.21)
Secondhand Smoke Exposure	
< 5 ng/mL	-0.90 (-2.88, 1.08)
≥ 5 ng/mL	-0.27 (-4.47, 3.94)
Diagnosed Asthma	
No	-0.94 (-2.90, 1.02)
Yes	-0.88 (-4.35, 2.59)
Maternal Education (4-Year College or Greater)	
No	-0.58 (-2.84, 1.68)
Yes	-1.42 (-4.03, 1.19)
Annual Household Income < $30,000	
No	-1.01 (-3.09, 1.08)
Yes	-0.99 (-4.31, 2.34)
Childhood Opportunity Index	
≥ 50	0.54 (-1.75, 2.83)
< 50	-2.78 (-5.33, -0.24)
CBCL Total Problems T-Score	
< 60	0.29 (-1.68, 2.25)
≥ 60	-3.59 (-6.75, -0.44)
PSQ-SRBDS Score	
< 0.33	1.02 (-2.45, 4.50)
≥ 0.33	-1.50 (-3.46, 0.46)
ESS Total Score	
< 10	-1.90 (-3.83, 0.04)
≥ 10	2.75 (-0.86, 6.37)
OSA-18	
< 60	-0.96 (-2.97, 1.04)
≥ 60	-1.03 (-4.11, 2.05)
Maximum Tonsil Grade	
II	-0.32 (-3.06, 2.42)
III & IV	-1.36 (-3.54, 0.82)
AHI level	
AHI < 1	-1.32 (-3.37, 0.73)
AHI ≥ 1	0.00 (-3.09, 3.09)

ETable 9b. Treatment effects for the Go/No-go CPT in various subgroups

	Effect Estimate
All Patients	*0.05 (-0.18, 0.27)*
Age (in years)	
5 or less years old	-0.06 (-0.39, 0.27)
6 or more years old	0.14 (-0.16, 0.44)
Sex	
Female	-0.14 (-0.45, 0.18)
Male	0.23 (-0.08, 0.54)
Racial/Ethnic Minority	
No	0.07 (-0.24, 0.37)
Yes	0.02 (-0.31, 0.34)
BMI Percentile	
≤ 85th BMI percentile	0.17 (-0.11, 0.45)
> 85th BMI percentile	-0.17 (-0.54, 0.20)
Obesity Status	
< 95th BMI percentile	0.09 (-0.16, 0.34)
≥ 95th BMI percentile	-0.14 (-0.65, 0.36)
Secondhand Smoke Exposure	
< 5 ng/mL	0.07 (-0.18, 0.33)
≥ 5 ng/mL	-0.02 (-0.55, 0.52)
Diagnosed Asthma	
No	-0.07 (-0.32, 0.19)
Yes	0.43 (-0.02, 0.89)
Maternal Education (4-Year College or Greater)	
No	0.14 (-0.16, 0.43)
Yes	-0.07 (-0.42, 0.28)
Annual Household Income < $30,000	
No	0.08 (-0.19, 0.35)
Yes	-0.05 (-0.47, 0.38)
Childhood Opportunity Index	
≥ 50	0.13 (-0.17, 0.43)
< 50	-0.07 (-0.40, 0.26)
CBCL Total Problems T-Score	
< 60	0.04 (-0.23, 0.30)
≥ 60	-0.02 (-0.44, 0.41)
PSQ-SRBDS Score	
< 0.33	0.54 (0.08, 0.99)
≥ 0.33	-0.08 (-0.34, 0.17)
ESS Total Score	
< 10	0.00 (-0.26, 0.25)
≥ 10	0.20 (-0.27, 0.67)
OSA-18	
< 60	0.08 (-0.19, 0.34)
≥ 60	0.02 (-0.39, 0.43)
Maximum Tonsil Grade	
II	0.32 (-0.04, 0.69)
III & IV	-0.12 (-0.41, 0.16)
AHI level	
AHI < 1	0.18 (-0.09, 0.45)
AHI ≥ 1	-0.26 (-0.66, 0.15)



ETable 9c. Treatment effects for the Apnea-Hypopnea Index (on a log-transformed scale) outcome in various subgroups

	Effect Estimate
All Patients	*-0.60 (-0.93, -0.27)*
Age (in years)	
5 or less years old	-0.53 (-1.02, -0.05)
6 or more years old	-0.66 (-1.10, -0.23)
Sex	
Female	-0.73 (-1.19, -0.27)
Male	-0.43 (-0.88, 0.02)
Racial/Ethnic Minority	
No	-0.40 (-0.86, 0.07)
Yes	-0.79 (-1.26, -0.33)
BMI Percentile	
≤ 85th BMI percentile	-0.48 (-0.89, -0.07)
> 85th BMI percentile	-0.81 (-1.35, -0.26)
Obesity Status	
< 95th BMI percentile	-0.63 (-1.00, -0.27)
≥ 95th BMI percentile	-0.52 (-1.27, 0.23)
Secondhand Smoke Exposure	
< 5 ng/mL	-0.64 (-1.02, -0.27)
≥ 5 ng/mL	-0.46 (-1.24, 0.32)
Diagnosed Asthma	
No	-0.61 (-0.99, -0.24)
Yes	-0.62 (-1.29, 0.05)
Maternal Education (4-Year College or Greater)	
No	-0.50 (-0.93, -0.07)
Yes	-0.78 (-1.29, -0.28)
Annual Household Income < $30,000	
No	-0.54 (-0.95, -0.13)
Yes	-0.78 (-1.41, -0.16)
Childhood Opportunity Index	
≥ 50	-0.66 (-1.10, -0.22)
< 50	-0.56 (-1.04, -0.08)
CBCL Total Problems T-Score	
< 60	-0.61 (-1.00, -0.23)
≥ 60	-0.53 (-1.17, 0.11)
PSQ-SRBDS Score	
< 0.33	-0.04 (-0.70, 0.63)
≥ 0.33	-0.80 (-1.17, -0.42)
ESS Total Score	
< 10	-0.40 (-0.77, -0.03)
≥ 10	-1.43 (-2.10, -0.77)
OSA-18	
< 60	-0.66 (-1.05, -0.27)
≥ 60	-0.50 (-1.11, 0.11)
Maximum Tonsil Grade	
II	-0.39 (-0.93, 0.14)
III & IV	-0.69 (-1.11, -0.28)
AHI level	
AHI < 1	-0.51 (-0.86, -0.16)
AHI ≥ 1	-1.03 (-1.56, -0.50)

ETable 10. Primary outcomes, stratified by the state of the COVID-19 pandemic

	Early Adenotonsillectomy			Watchful Waiting			
	Baseline	12 mo.	Change from Baseline to 12 mo.	Baseline	12 mo.	Change from Baseline to 12 mo.	Effect Size – Difference in 12 mo. Changes (95% CI)
Prior to the COVID-19 pandemic[b]							
Caregiver BRIEF GEC T-score[c]	53.1 (49.6 to 56.6)	50.2 (46.6 to 53.8)	-2.9 (-4.4 to -1.4)	53.6 (50.2 to 57.0)	51.3 (47.8 to 54.8)	-2.3 (-3.8 to -0.8)	-0.6 (-2.7 to 1.5)
Go/No-Go sustained attention d'[d]	2.1 (1.8 to 2.3)	2.4 (2.1 to 2.6)	0.3 (0.1 to 0.5)	2.2 (1.9 to 2.5)	2.4 (2.1 to 2.7)	0.2 (0.0 to 0.4)	0.1 (-0.2 to 0.3)
During the COVID-19 pandemic[b]							
Caregiver BRIEF GEC T-score[c]	54.4 (49.9 to 58.9)	50.3 (46.5 to 54.1)	-4.1 (-7.0 to -1.2)	56.4 (51.6 to 61.2)	52.8 (48.9 to 56.6)	-3.7 (-6.9 to -0.4)	-0.4 (-4.8 to 4.0)
Go/No-Go sustained attention d'[d]	2.1 (1.7 to 2.5)	2.1 (1.8 to 2.5)	0.0 (-0.3 to 0.4)	2.1 (1.6 to 2.6)	2.2 (1.9 to 2.6)	0.1 (-0.3 to 0.5)	-0.1 (-0.6 to 0.5)

Abbreviations: BRIEF, Behavior Rating Inventory of Executive Function; CI, confidence interval; COVID-19, coronavirus disease 2019; GEC, global executive composite.

[a]The within-group (changes in) means and the between-group differences in mean change from baseline to 12 months are all estimated from a linear mixed-effects model. The model includes interaction terms involving a visit-specific pandemic onset indicator and adjustment for stratification factors (age >5, overweight status, Black/African American) and site effects. All estimates are reported as mean (95% CI).

[b]The pandemic onset date is taken to be March 11, 2020, the date on which the World Health Organization declared the COVID-19 outbreak to be a global pandemic.

[c]The BRIEF GEC section comprises summary measures of behavioral regulation, emotion regulation, and cognitive regulation (BRIEF-2, for children ages 5 to 18) or inhibitory self-control, flexibility, and emergent metacognition (BRIEF-P, for preschool-aged

children). Caregiver GEC scores ranged from 33 to 102, with higher scores indicating worse functioning. All 458 patients contributed information to the sensitivity analysis.
[d]The Go/No-Go sustained attention d' is a signal detection measure that combines a child's true positive rate on an attention task (correct response to the target stimuli) with their false alarm rate (incorrect response to the non-target stimuli). Scores ranged from -0.9 to 4.5, with higher scores indicating greater sustained attention. Four hundred fifty-five patients contributed information to the sensitivity analysis.

ETable 11a. Adverse events by randomized arm, according to seriousness and relatedness to study procedures

	Early Adenotonsillectomy		**Watchful Waiting**	
	n events		*n* events	
	Serious	*Non-serious*	*Serious*	*Non-serious*
Adverse events related to adenotonsillectomy [a]				
Post-operative pain		12		
Post-operative bleeding	5	8	1	
Dehydration		3		
Aspiration pneumonia		1		
Adverse events related to other study procedures [b]				
Hives (allergy to tape used in PSG)		1		
Vomiting (immediately following blood draw)		1		
Adverse events unrelated to study procedures				
ADHD		4		6
Allergy		5		4
Asthma		19	2	15
Behavioral issues other than ADHD		9		8

Exacerbation of SDB		1		5
Gastrointestinal tract illness		38		54
Headache or migraine		3		7
Infections other than respiratory and ear		53		82
Lower respiratory tract illness	1	37	1	30
Sleep issues other than SDB		4		6
Trauma/injury	5	18	6	32
Upper respiratory tract or ear illness		138		188
All other events	2	30	1	29

[a] The following events were pre-specified as not requiring reporting: intra-operative blood less ≤7 mL/kg; post-operative pain, hoarseness, or difficulty swallowing lasting <21 days and not requiring intravenous hydration or unscheduled evaluation or treatment; post-operative blood-tinged oral or nasal secretions lasting <72 hours; velopharyngeal insufficiency lasting <2 months and not requiring evaluation or treatment.

[b] The following events were pre-specified as not requiring reporting: skin irritation associated with adhesives used in PSGs lasting <2 days; temporary depigmentation under areas of PSG sensor attachment lasting <1 month; poor sleep during PSG; temporary pain at the site of the phlebotomy lasting <48 hours; bleeding or bruising at the site of the phlebotomy not requiring evaluation or treatment; anxiety surrounding behavioral testing not requiring psychiatric attention.

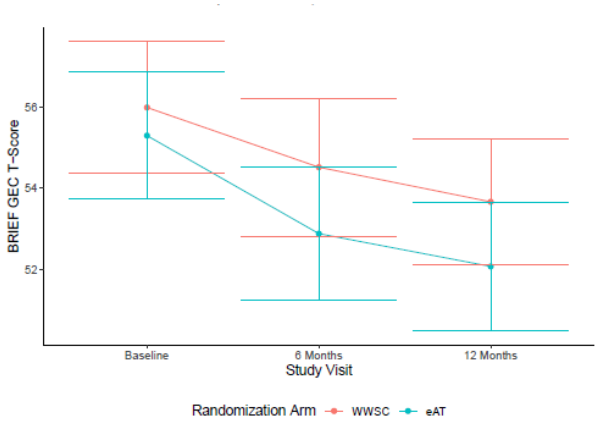
ETable 11b. Participants with unrelated adverse events by randomized arm

	Early Adenotonsillectomy *n* participants		**Watchful Waiting** *n* participants	
	Serious	*Non-serious*	*Serious*	*Non-serious*
ADHD		4		6
Allergy		5		4
Asthma		15	1	12
Behavioral issues other than ADHD		9		8
Exacerbation of SDB		1		5
Gastrointestinal tract illness		34		43
Headache or migraine		3		7
Infections other than respiratory and ear		42		59
Lower respiratory tract illness	1	33	1	23
Sleep issues other than SDB		4		6
Trauma/injury	5	18	6	29
Upper respiratory tract or ear illness		90		100
All other events	2	27	1	24

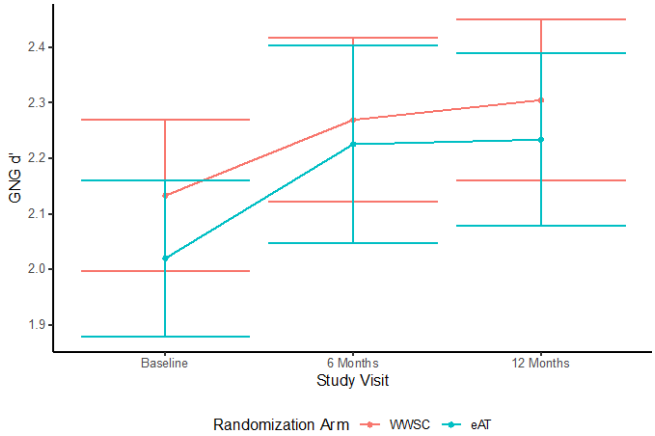
Unrelated adverse events were defined as any unfavorable or unintended sign, symptom, or disease occurring in a participant at any stage following consent not identified to be related to the intervention or study procedure

Supplementary Figure

EFigure 1. Longitudinal trajectory of mean BRIEF GEC T-score and 95% confidence intervals within each study arm and at each study visit, obtained from a mixed effects model



EFigure 1b. Longitudinal trajectory of mean unadjusted Go/No-Go Sustained Attention d' scores and 95% confidence intervals within each study arm and at each study visit, obtained from a mixed effects model



eReferences

1. Redline S, Sanders MH, Lind BK, et al. Methods for obtaining and analyzing unattended polysomnography data for a multicenter study. Sleep Heart Health Research Group. *Sleep*. 1998;21(7):759-767.

2. Kumar HVM, Schroeder JW, Gang Z, Sheldon SH. Mallampati score and pediatric obstructive sleep apnea. *J Clin Sleep Med*. 2014;10(9):985-990. doi:10.5664/jcsm.4032

3. Rosner B, Cook N, Portman R, Daniels S, Falkner B. Determination of blood pressure percentiles in normal-weight children: some methodological issues. *Am J Epidemiol*. 2008;167(6):653-666. doi:10.1093/aje/kwm348

4. Wiebe SA, Sheffield TD, Andrews Espy K. Separating the fish from the sharks: a longitudinal study of preschool response inhibition. *Child Dev*. 2012;83(4):1245-1261. doi:10.1111/j.1467-8624.2012.01765.x

5. Clark CAC, Cook K, Wang R, et al. Psychometric properties of a combined go/no-go and continuous performance task across childhood. *Psychol Assess*. January 12, 2023. doi:10.1037/pas0001202

6. NIH Toolbox 9-Hole Pegboard. Accessed August 9, 2022. http://www.nihtoolbox.org/WhatAndWhy/Motor/Dexterity/Pages/NIH-Toolbox-9-Hole-Pegboard-Dexterity-Test.aspx

7. Smith YA, Hong E, Presson C. Normative and validation studies of the Nine-hole Peg Test with children. *Percept Mot Skills*. 2000;90(3 Pt 1):823-843. doi:10.2466/pms.2000.90.3.823

8. Gioia GA, Isquith PK, Guy SC, Kenworthy L, Denckla MB. *Behavior Rating Inventory of Executive Function (BRIEF)*. Mt. Washington Pediatric Hospital; 1999.

9. Gioia GA, Isquith PK, Retzlaff PD, Espy KA. Confirmatory factor analysis of the Behavior Rating Inventory of Executive Function (BRIEF) in a clinical sample. *Neuropsychol Dev Cogn C Child Neuropsychol*. 2002;8(4):249-257.

10. Isquith PK, Crawford JS, Espy KA, Gioia GA. Assessment of executive function in preschool-aged children. *Ment Retard Dev Disabil Res Rev*. 2005;11(3):209-215. doi:10.1002/mrdd.20075

11. Chervin RD, Ellenberg SS, Hou X, et al. Prognosis for spontaneous resolution of OSA in children. *Chest*. 2015;148(5):1204-1213. doi:10.1378/chest.14-2873

12. Acevedo-Garcia D, Noelke C, McArdle N, et al. Racial and ethnic inequities in children's neighborhoods: evidence from the new child opportunity index 2.0. *Health Aff (Millwood)*. 2020;39(10):1693-1701. doi:10.1377/hlthaff.2020.00735

13. Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*. 1975;31(1):103-115. doi:10.2307/2529712

14. Ware JH, Harrington D, Hunter DJ, D'Agostino, Sr. RB. Missing data. . *N Engl J Med*. 2012;367(14):1353-1354.

15. Molenberghs G, Thijs H, Jansen I, et al. Analyzing incomplete longitudinal clinical trial data. *Biostatistics*. 2004;5(3):445-464. doi:10.1093/biostatistics/5.3.445

16. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581-592. doi:10.1093/biomet/63.3.581

17. Rubin DB. *Multiple Imputation for Nonresponse in Surveys* . Vol 18. John Wiley & Sons; 2004.

18. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *J Royal Stat Soc*. 1995;57:289-300.

19. Benjamini Y, Yekutieli D. False discovery rate–adjusted multiple confidence intervals for selected parameters. *J Am Stat Assoc*. 2005;100(469):71-81. doi:10.1198/016214504000001907

