OXFORD

Sleep Research Society®

SLEEPJ, 2021, 1–11

doi: 10.1093/sleep/zsaa229
Advance Access Publication Date: 9 November 2020
Special Article

Special Article

Randomized clinical trials of cardiovascular disease in obstructive sleep apnea: understanding and overcoming bias

Allan I. Pack[1,*], Ulysses J. Magalang[2], Bhajan Singh[3], Samuel T. Kuna[4], Brendan T. Keenan[1,5] and Greg Maislin[1,5]

[1]Division of Sleep Medicine/Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, [2]Division of Pulmonary, Critical Care and Sleep Medicine, The Ohio State University Wexner Medical Center, Columbus, Ohio, [3]West Australian Sleep Disorders Research Institute, Department of Pulmonary Physiology & Sleep Medicine, Sir Charles Gairdner Hospital, Nedlands, WA, Australia, [4]Sleep Medicine Section, Corporal Michael J. Crescenz Veterans Affairs Medical Center, Philadelphia, Pennsylvania, [5]Biostatistics Core, Division of Sleep Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania

*Corresponding author. Allan I. Pack, Division of Sleep Medicine, Department of Medicine, University of Pennsylvania Perelman School of Medicine, Translational Research Laboratories, Suite 2100, 125 South 31st Street, Philadelphia, PA 19104-3403. Email: pack@pennmedicine.upenn.edu.

Abstract

Three recent randomized control trials (RCTs) found that treatment of obstructive sleep apnea (OSA) with continuous positive airway pressure (CPAP) did not reduce rates of future cardiovascular events. This article discusses the biases in these RCTs that may explain their negative results, and how to overcome these biases in future studies.

First, sample selection bias affected each RCT. The subjects recruited were not patients typically presenting for treatment of OSA. In particular, subjects with excessive sleepiness were excluded due to ethical concerns. As recent data indicate that the excessively sleepy OSA subtype has increased cardiovascular risk, subjects most likely to benefit from treatment were excluded. Second, RCTs had low adherence to therapy. Reported adherence is lower than found clinically, suggesting it is in part related to selection bias. Each RCT showed a CPAP benefit consistent with epidemiological studies when restricting to adherent patients, but was underpowered.

Future studies need to include sleepy individuals and maximize adherence. Since it is unethical and impractical to randomize very sleepy subjects to no therapy, alternative designs are required. Observational designs using propensity scores, which are accepted by FDA for studies of medical devices, provide an opportunity. The design needs to ensure covariate balance, including measures assessing healthy user and healthy adherer biases, between regular users of CPAP and non-users. Sensitivity analyses can evaluate the robustness of results to unmeasured confounding, thereby improving confidence in conclusions. Thus, these designs can robustly assess the cardiovascular benefit of CPAP in real-world patients, overcoming biases in RCTs.

Statement of Significance

Randomized control trials of the effect of continuous positive airway pressure on cardiovascular events have been negative, which is inconsistent with epidemiological data. Major biases in these trials could explain the negative results. For ethical reasons, recent trials excluded subjects with excessive sleepiness, the very group at increased risk for cardiovascular events. Moreover, adherence to therapy was inadequate. Future studies need to include and focus on sleepy subjects. Ethical limitations to including these patients can be overcome with observational designs using propensity scores. To obtain a robust treatment effect, these designs need to directly ensure balance of covariates related to cardiovascular events, including measures of healthy user and healthy adherer bias, in patients very compliant to CPAP compared to non-users.

Key words: obstructive sleep apnea; cardiovascular disease; randomized control trials; propensity score matching; bias

Submitted: 12 August, 2020; **Revised:** 28 October, 2020

Introduction

Obstructive sleep apnea (OSA) is an extremely common disorder [1–3]. It is estimated there are likely 1.0 billion people worldwide with undiagnosed OSA [1]. Apneas and hypopneas during sleep result in sleep fragmentation and cyclical intermittent hypoxia in every tissue, although the pattern of deoxygenation/reoxygenation varies between organs [4]. There is an effective and safe therapy—nasal continuous positive airway pressure (CPAP) [5]. Both the daily use of CPAP and its efficacy can be monitored remotely. Given the high prevalence of OSA, there is a need to determine its influence on other chronic disorders. If OSA adversely affects medically important conditions, then treating people with OSA becomes crucial for improving overall health. This understanding can also aid policymakers in determining appropriate reimbursement strategies and screening recommendations. Currently, the evidence for routine screening for OSA is considered inadequate [6].

The first approaches to understanding the impact of OSA in other health conditions included well-conducted epidemiological studies, with control for relevant confounding factors. The Sleep Heart Health Study (SHHS) in community-based samples [7, 8] and the Wisconsin Sleep Cohort Study [9] in state employees in Wisconsin provided important results. OSA was shown to be an independent risk factor for hypertension [10, 11], cardiovascular disease [12], coronary disease [13], stroke [14, 15], carotid atherosclerosis [16], cardiac arrhythmias [17, 18], and mortality [19, 20]. These findings from community samples were supported by studies in clinical cohorts. For example, the Spanish Sleep Research Network demonstrated a substantial increase of cardiovascular events in patients with severe OSA not using CPAP, compared to controls, less severe apneics and severe OSA using CPAP [21, 22]. This was not simply related to a “healthy adherer” bias, since medication refills for multiple classes of medications were the same in CPAP users and non-users [23]. Although impactful, these studies likely over-estimated the effects of OSA. While studies controlled for important covariates, other potential confounders, such as level of exercise/activity [24], dietary differences, [25], and specific distributions of fat [26], were not included.

It has recently been shown that OSA is a heterogeneous disorder with respect to physiology [27] and symptoms [28–32]. Cluster analysis in the Icelandic Sleep Apnea Cohort (ISAC) identified three distinct symptom subtypes of OSA [28]: (1) *disturbed sleep* (e.g. complaints of insomnia); (2) *relatively asymptomatic*; and (3) *excessively sleepy*. These findings have been replicated in sleep centers around the world that participate in the Sleep Apnea Global Interdisciplinary Consortium (SAGIC) [29], and in the community-based Korean Genome Cohort [30] and SHHS [31]. Importantly, only the *excessively sleepy* group shows consistent evidence of increased cardiovascular risk [31]. This is a key observation both for understanding clinical care and interpreting results of recent cardiovascular clinical trials, where adults with OSA and excessive sleepiness have been excluded mainly due to ethical concerns related to increased likelihood of car crashes without treatment [33].

Given the results of epidemiological studies, it was natural to study the role of OSA using randomized control trials (RCTs). There is a view that evidence based on RCTs is essential [34] since randomization is expected to result in good covariate balance between treated and untreated groups when applied to large numbers. Indeed, with the development of sham CPAP [35], successful and important RCTs have been conducted in OSA for sleepiness [36, 37], quality of life [38], and blood pressure [39], including in resistant hypertension [40]. These studies have shown clear benefits of CPAP therapy with respect to relatively short-term endpoints, which has improved clinical care. In contrast, recent RCTs assessing the effects of OSA treatment on cardiovascular events have been negative [41–45]. This raises the question as to whether OSA is indeed of clinical relevance for cardiovascular disease. Alternatively, could there have been methodological issues that impaired the ability of RCTs to detect the benefit of OSA therapy for reduction in cardiovascular events?

While RCTs are considered the optimal approach, they are not without inherent problems. These biases have been described in a recent report by Krauss [46], building on previous literature [47–49]. Consequently, there are proposed and evolving standards for reporting of randomized clinical trials—the Consolidated Standards of Reporting Trials (CONSORT) [50–53]. Although RCTs are the preferred level of evidence in the ideal world, alternative methods for estimating causal treatment effects from real-world data are often required (see below). In OSA, this is particularly true given biases in RCTs that result from more severe patients being under-represented in longer-term studies and the inability to ethically randomize excessively sleepy patients to no therapy given the established benefits of CPAP and the danger sleepy individuals can present to themselves and others.

In this article, we focus on the potential challenges and biases in recently conducted RCTs of cardiovascular endpoints of OSA [41–45] and describe how these may have led to their negative conclusions. We then detail the rationale for, and the implementation of, alternative approaches to obtaining evidence of causality from carefully designed observational studies, some of which have been utilized within the context of these same negative RCTs. Given the likely biases in the recent RCTs, we argue that we cannot simply conclude that CPAP has no benefit in reducing cardiovascular events based on these recent negative results. As Black indicated many years ago, “the false conflict between those who advocate randomized trials in all situations and those who believe observational data provide evidence needs to be replaced with mutual recognition of the complimentary roles of these two approaches.” [54].

RCTs on Role of CPAP in Prevention of Cardiovascular Events

We first briefly describe the study populations and primary outcomes of recent randomized trials of cardiovascular endpoints in OSA.

The sleep apnea cardiovascular endpoints (SAVE) study

The SAVE trial was a secondary prevention study, including participants with coronary or cerebrovascular disease and an oxygen desaturation index (4%) of ≥12 events/h, based on an ApneaLink (ResMed) device, recruited from 89 clinical centers in seven countries [41, 42]. Notably, the ApneaLink does not include an assessment of respiratory effort or oral airflow and hence can misclassify events. Subjects with marked excessive sleepiness (Epworth Sleepiness Scale [ESS] score > 15) and/or severe hypoxemia (oxygen saturation <80% for >10% of recording time) were excluded. Participants (n = 2,717) were randomized into a CPAP treatment group or usual care with no specific therapy for OSA. There was no difference between randomized arms with respect to incidence of the primary composite endpoint of death from any cardiovascular cause, myocardial infarction (including silent), stroke, hospitalization for heart failure, acute coronary syndrome (including unstable angina), or TIA (hazard ratio and 95% confidence interval [CI] with CPAP = 1.10 [0.91, 1.32]; p = 0.34) [41]. Thus, the RCT was negative.

The impact of sleep apnea syndrome in the evolution of acute coronary syndrome—effect of intervention with CPAP (ISAACC) study

The ISAACC study [43, 44] recruited individuals who had just been hospitalized for acute coronary syndrome with an apnea–hypopnea index (AHI) ≥15 events/h, based on a respiratory polygraphy performed 24 to 72 h after admission, across 15 hospitals in the Spanish Sleep Network. Subjects were excluded if they had an ESS >10, indicating elevated sleepiness. Participants with OSA (n = 1,264) were randomized to CPAP or usual care. There was no difference between arms in the incidence of the primary composite outcome of first cardiovascular event—cardiovascular death or non-fatal events (acute myocardial infarction, non-fatal stroke, hospital admission for heart failure, and new hospitalizations for unstable angina or TIA)—with a hazard ratio (95% CI) with CPAP of 0.89 (0.68, 1.17) (p = 0.40) [43]. Thus, this study was also negative.

The randomized intervention with continuous positive airway pressure in CAD and OSA (RICCADSA) study

The RICCADSA study [45] was also a secondary prevention trial, and included participants with angiography-demonstrated coronary disease who had undergone a revascularization procedure (surgical or percutaneous) and had an AHI ≥15 events/h identified by an overnight sleep study at home (cardiorespiratory polygraphy) or in-laboratory PSG. Subjects with an ESS ≥10 were excluded. Participants (n = 244) were randomized to either auto-titrating CPAP or no positive airway pressure. There was no difference in incidence of the primary outcome—a composite of repeat revascularization, new myocardial infarction, stroke or death attributed to cardiovascular causes—between those with and without CPAP (hazard ratio [95% CI] with CPAP = 0.80 [0.46, 1.41]; p = 0.45). Thus, the study was also negative with respect to the primary endpoint.

Understanding Challenges and Bias in RCTs and the Impact on OSA Trials for Assessment of Cardiovascular Benefit of CPAP

We now consider the key challenges and sources of bias in recent RCTs of cardiovascular endpoints in OSA, and how they may have contributed to the observed negative results.

Sample selection bias

Bias that results from the specific characteristics of the study sample that was randomized, when compared to the target population as a whole, is referred to as sample selection bias. For recent cardiovascular trials of OSA, it is important to ask if the recruited participants are representative of real-world patients. The answer is no. Thus, sample selection bias is a fundamental problem with all of these studies. There is a developing interest in ensuring that we are studying real-world patients [55].

To illustrate this point, we emphasize that due to ethical concerns with randomization, all recent trials excluded subjects with excessive sleepiness, based on different thresholds [41–45]. This resulted in study samples with levels of sleepiness considerably lower than typically seen in clinical practice. Given recent data indicating that adults with OSA exhibiting excessive sleepiness are at greatest cardiovascular risk [31, 56, 57], we believe this represents a major source of bias that contributed to the observed negative results in each trial.

Beyond excluding excessively sleepy individuals, selection bias may result from where and how included participants were recruited. While individuals with OSA from the general population can be a convenient source of recruitment, they tend to be relatively asymptomatic, raising questions as to the clinical significance of their disease [9, 58]. Recent randomized trials have focused on diagnosing OSA among individuals with established cardiovascular disease, as opposed to identifying adults with clinically diagnosed OSA. Recruitment in the SAVE study was initially based on the George's Institute Stroke and Cardiology network and while it expanded to include sleep centers, many of the subjects recruited were not individuals presenting with symptoms of OSA. Thus, the sample was not representative of typical adults seeking treatment. A key cause of this bias is the reality that symptomatic patients are less willing to be randomized to a study arm that receives no treatment for an extended period of follow-up and/or their providers are less likely to recommend participation. This point is perhaps most clearly illustrated in the Apnea Positive Pressure Long-term Efficacy Study (APPLES) [59], where authors noted that due to the required willingness to defer effective treatment for 6 months in the sham arm, "a majority of these participants [78%; C. Kushida, personal communication] were recruited from advertisements rather than clinically referred for OSA." Given the length of follow-up required to observe cardiovascular endpoints, this issue is likely even more of a barrier in the recent RCTs.

In addition to bias in recruitment approaches, all trials were designed as secondary prevention studies. In fact, the ISAACC study identified individuals immediately following an acute coronary event. Although we do not know the impact of acute cardiovascular disease on sleep-disordered breathing, other studies have waited until the patient stabilized before assessing OSA severity [60]. While recruiting individuals with pre-existing disease increases the rate of new cardiovascular events, statistical power is not necessarily enhanced unless the treatment effect size is as clinically meaningful for secondary prevention as it is for primary prevention.

Bias due to low adherence to therapy

For a treatment to be effective, one presumes that individuals assigned to therapy will adequately adhere. For CPAP, a clinically accepted criterion for adequate usage is an average of ≥4 h/night. Even at this threshold, additional sleep likely occurs without wearing CPAP, leading to unprotected sleep with attendant apneas, hypopneas, and intermittent oxygen desaturation. A lack of adherence will diminish the benefits of therapy, leading to negative results. As detailed below and illustrated in Figure 1, low adherence to CPAP therapy is another major source of bias in each of the recent RCTs of cardiovascular endpoints in OSA. This bias is likely related, in part, to the sample selection bias described above, as people not presenting clinically for diagnosis of OSA may be less likely to accept and adhere to treatment. As a clear illustration of the impact of lack of adherence, in each of the RCTs more favorable results that are consistent with observational studies were found in secondary analyses restricted to adherent individuals.

In the SAVE study [41, 42], despite an initial run-in period that achieved an average usage of 5.2 h/night, CPAP usage declined over the first year to 3.5 ± 2.4 h/night and was only 3.3 ± 2.3 h/night at final follow-up. Only 42% of subjects in the CPAP treated group achieved acceptable adherence (≥4 h/night). Strong evidence that lack of adherence may have biased results is shown through a secondary propensity score matching analysis conducted by the SAVE investigators [41] (this technique is described in more detail below). When creating a matched sample of 561 adherent participants from the CPAP arm and 561 participants receiving usual care, there were fewer CV events among the CPAP arm (hazard ratio [95% CI] = 0.80 [0.60, 1.07]; $p = 0.13$). This estimate is consistent with both observational data and evidence from the meta-analysis of Yu et al. [61]. However, the sample size for this secondary analysis was underpowered for significance [62].

Similarly, the adherence to CPAP was extremely low in the ISAACC study [43, 44]. At one year after starting CPAP, average compliance was only 2.8 ± 2.6 h/night, with only 227 of 629 patients in the CPAP arm (36%) achieving "good adherence" (≥4 h/night on average). A propensity score analysis comparing those

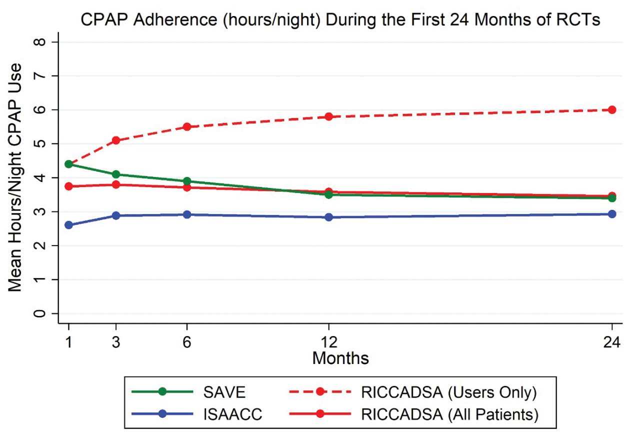

Figure 1. Average CPAP compliance (hours/night) over the first 24 months reported in recent RCTs of cardiovascular endpoints in OSA. Each study shows sub-optimal adherence throughout the study. The SAVE study (green line) [41] shows a progressive decline in average hours of use per night, while the ISAACC study (blue line) [43] shows low adherence even during the early phase of the study. In the RICCADSA study, [45] increased CPAP adherence throughout the study is driven by the fact that estimates were derived only among those that continued using CPAP (dashed red line). When incorporating no usage in those reported to have stopped using CPAP (solid red line), estimated adherence levels are similar to both the SAVE and ISAACC trials.

achieving "good adherence" to those with usual care showed a hazard ratio of 0.80 (95% CI: 0.52, 1.23; $p = 0.32$) favoring the CPAP arm. Once again, this estimate is similar in magnitude to that found in SAVE [41] and in statin trials [63], but was achieved in a very limited sample size with a lack of statistical power to detect significance.

Finally, in the RICCADSA study [45], only 76 of 122 participants who started CPAP (62.3%) were still using therapy at 1 year. In those using CPAP, the average adherence was 5.8 ± 1.7 h/night. Thus, the bias in this study was due to the relatively large number of participants who stopped using CPAP (37.7%). Once again emphasizing the bias caused by lack of adherence, the incidence of the composite endpoint was 2.31 (95% CI: 0.96, 5.54) per 100 person-years in those using CPAP ≥4 h/night on average, compared to 5.32 (95% CI: 3.96, 7.15) per 100 person-year in those using CPAP <4 h/night. The study observed a significant decreased risk of cardiovascular events among those using CPAP ≥4 h/night compared to <4 h/night or no CPAP (hazard ratio [95% CI] of 0.29 [0.10, 0.86]; $p = 0.026$).

Therefore, the lack of adequate adherence to CPAP was a major source of bias in each of the recent RCTs of cardiovascular endpoints. Moreover, it can be argued that the adherence found in these RCTs is not representative of that found clinically in adults with OSA [64]. When adherent participants are compared to non-adherent patients, often employing causal analysis approaches to account for confounding factors, each study showed some evidence suggesting a benefit of CPAP. While unmeasured confounding, such as healthy user or healthy adherer bias (see below), could explain some of these positive benefits, results from these secondary analyses were also consistent with epidemiological data. Issues with adherence are likely to be further impacted in longer-term studies, emphasizing the need for developing and utilizing programs to enhance CPAP adherence [65].

Small sample size and lack of power

Another challenge with recent RCTs is an inadequate sample size. In multiple instances, reasonable and consistent effect sizes were observed (particularly when accounting for lack of adherence), but samples were underpowered to declare statistical significance, resulting in negative conclusions.

Towards this point, Javaheri et al. [66] have recently argued that inadequate sample size is a major problem, estimating that between 8,000 and 12,000 subjects *per arm* are required. Thus, while the sample size for the SAVE study ($n = 2{,}717$) may seem large [41], it remains underpowered [66]. The sample size for ISAACC was considerably smaller than for the SAVE trial. Based on an expected 25% risk reduction in the CPAP arm and an assumed 12% to 20% rate of new cardiovascular events in the first year among patients with acute coronary syndrome [21], 1,264 patients were randomized to CPAP ($n = 633$) or usual care ($n = 631$). Given the very low CPAP compliance rate, and an observed risk reduction of only 20% among compliant patients, the study was not powered to detect a difference. Finally, the sample size in the RICCADSA study [45] was extremely small, with a total of 244 patients. This sample size estimate was based on available evidence more than 10 years prior to the final publication, at which point there had been no studies of CPAP in revascularized patients with CAD and OSA [45].

Therefore, while each study provided information on how the initial sample sizes were determined, all ultimately suffered from a lack of statistical power. Larger trials are needed to robustly determine the benefit of CPAP on cardiovascular endpoints. These trials need to be in real-world patients with the disorder.

Challenges with composite endpoints

Each of the recent RCTs also utilized a composite endpoint, combining fatal and non-fatal events, including myocardial infarction, stroke, heart failure, acute coronary syndrome, unstable angina, and/or TIA. The use of composite endpoints can improve statistical power by increasing the incidence of events when all components are impacted by treatment in the same direction, or reduce statistical power if treatment affects components differently [66]. Thus, although composite endpoints have advantages, there are also challenges to this approach [67]. Typically, each component of the composite endpoint is weighted equally. However, this may not be optimal since some endpoint are more clinically important than others (e.g. sudden cardiovascular death compared to TIA). Relatedly, studies using equally-weighted composite endpoints may not adequately consider the concept of competing risks, such as fatal events prohibiting future occurrence of non-fatal events. While only evaluating the first incident event is one commonly used approach to address competing risks, this may not adequately capture the broader impact of therapy on cardiovascular endpoints/risk.

CPAP may also differentially affect certain components of the composite score. There is some evidence that OSA may have more marked effects on the cerebrovascular system than the coronary system. Vibration of the carotid arteries secondary to snoring may lead to plaques in the carotid arteries [68]. A loss of autoregulation of the cerebrovascular system has also been described in patients with OSA and is corrected by CPAP treatment [69]. Consistent with this idea, secondary propensity score analyses in the SAVE trial suggested a CPAP benefit specific to stroke and a composite of only cerebral events [41]. In contrast, the ISAACC trial found no particular benefit of CPAP on stroke events within the propensity score-matched sample [43]. Two recent meta-analyses including subjects in these trials adherent to therapy also suggested a stronger benefit of CPAP for cerebrovascular events compared to coronary events [70, 71]. However, some caution may be warranted when interpreting these results since results from the SAVE trial may dominate the findings due to the much larger number of participants compared to other recent studies. Notably, the observation of a stronger effect of CPAP on stroke and cerebral events in the SAVE study could reflect selection bias, as recruitment in China was initially based on the George's Institute Stroke and Cardiology network population, enriching the sample for stroke and cerebral events. In fact, 63.1% of SAVE participants [41] were recruited in China, where stroke is the leading cause of death [72]. Supporting this point, approximately 5% of the SAVE sample experienced a stroke [41], compared to only 1% of the ISAACC sample [43].

Ultimately, the evidence that CPAP has a differential benefit on cerebrovascular events compared to coronary events is not particularly strong. Thus, the present state of knowledge seems to support the continued use of a primary composite endpoint to evaluate the cardiovascular benefit of CPAP. However, studies need to carefully consider the potential challenges with composite endpoints and develop plans to address these in the design stage. In all cases, treatment group differences in the components of a composite outcome should be summarized to provide a complete clinical interpretation. Ideally, studies should be designed to maintain power to perform secondary analyses evaluating the possible differential benefit of CPAP on cerebrovascular compared to coronary events. Given the lower incidence of these component-specific events, this will require larger sample sizes than if only focused on the overall composite. In addition, more careful consideration and characterization of the underlying physiological response to CPAP, beyond the AHI, may help to understand the specific mechanisms through which treatment affects these endpoints. In this regard, studies should quantify novel variables that can be extracted from the overnight sleep study [73]. The overall hypoxic burden may be particularly important since it has been associated with cardiovascular mortality [74].

How Do We Include Excessively Sleepy Patients in Clinical Trials of Cardiovascular disease?

Selection bias introduced by exclusion of excessively sleepy individuals in each of these RCTs [41, 43, 45] of cardiovascular endpoints deserves special consideration. Recent literature suggests this is the very group likely to benefit the most from CPAP in terms of reduction of future cardiovascular events [31]. If true, excluding people who are excessively sleepy clearly limits a trial's ability to observe a positive effect. Earlier randomized trials including sleepy patients demonstrated benefits of CPAP for short-term outcomes such as sleepiness [75, 76], quality of life [38, 77], and blood pressure [39]. This has led some to argue that there is no value in studying excessively sleepy patients – they will be treated regardless of the cardiovascular benefits. This, in our view, misses the point. Establishing a benefit of CPAP for reducing risk of cardiovascular events, which is best accomplished by including sleepy individuals, is likely to help reverse recent claims of insufficient evidence to support routine screening for OSA [6].

There are a number of possible ways to include excessively sleepy individuals in trials of cardiovascular endpoints, outlined below.

1. **Standard RCT including excessively sleepy individuals.** Since the benefit of CPAP with respect to cardiovascular outcomes is uncertain, one may argue that it is ethical to randomize even excessively sleepy patients into CPAP treatment or usual care [78]. As evidenced by the exclusion of these individuals in recent trials, it seems clear this is not a viable strategy [41–45]. This approach will be challenging for IRBs to approve since excessively sleepy patients are presumed to have an increased risk of vehicular crashes if untreated [33], posing a risk not only to themselves, but also to others on the road. Moreover, given the known benefits of CPAP for improving sleepiness, potential participants and/or their providers will have concerns about being randomized, creating unavoidable selection bias [59]. Given the high prevalence of OSA and cardiovascular disease, some may also consider

a traditional RCT with an extremely large sample (e.g. 50,000 patients) and very short follow-up period (e.g. <3 months) as a feasible way to maintain statistical power for detecting cardiovascular events without ethnical limitations. However, this approach is unlikely to avoid the selection bias caused by symptomatic participants being less willing to participate, as noted in the APPLES study [59], and only addresses the very short-term cardiovascular benefits of CPAP.

2. **Randomizing to multiple treatment arms.**
   A second approach might be to conduct a randomized trial comparing usual CPAP care with an arm that is designed to enhance and sustain CPAP adherence. Since both arms include active treatment, there are no longer ethical concerns about randomizing people who are excessively sleepy. As opposed to comparing CPAP users to non-users, the primary analysis would evaluate whether the arm with enhanced adherence shows improved outcomes compared to those with usual adherence. While this approach overcomes the ethical concerns, there are practical concerns for conducting such a study. Since the primary analysis compares two active treatment groups, risk differences between the arms are expected to be considerably smaller than when comparing treated and untreated subjects. Thus, these studies will likely require very large sample sizes to maintain statistical power.
3. **Pharmacologically treating sleepiness in an RCT.**
   A third approach, recently advocated by Javaheri et al. [66] is to randomize all patients to active CPAP or control, but simultaneously prescribe FDA-approved therapeutics that promote wakefulness (e.g. modafinil, solriamfetol, or pitolisant) to excessively sleepy patients in both arms. Ultimately, this is not reflective of typical clinical practice, resulting in concerns about generalizability. Moreover, this type of blanket pharmacological treatment may lead to additional concerns, and it is unknown whether drugs will result in other effects that confound interpretation of the benefits of CPAP in their absence. Thus, the validity of this approach remains unclear.
4. **Apply techniques for estimating causal treatment effects in observational data.**
   A fourth strategy is to utilize analysis techniques designed to reduce bias in estimated treatment effects from observational real-world data, such as methods relying on propensity scores [79, 80]. This approach is particularly applicable when it is unethical to randomize individuals into no treatment [81], as is the case with excessively sleepy people with OSA. While this approach has had limited application in studies of OSA [82], it has become more widely used in a number of other diseases, including in cardiovascular disease [81], as well as in determining the efficacy of devices [83]. Supporting the validity of this approach, the FDA Center for Devices and Radiological Health (CDRH) has indicated that propensity score methods are appropriate to support approval of medical devices, which would include CPAP, for studies using non-randomized controls as long as subject-level data are available for both covariates and outcomes [84]. Given that data on CPAP use can be obtained remotely, we are in an ideal position to apply these propensity score designs to compare outcomes in individuals who are adherent to CPAP to those who do not use the therapy.

Overcoming Bias in RCTs Using Real-World Observational Data

Propensity score matching of real-world observational data to estimate causal treatment effects represents a promising method for overcoming the described biases in recent RCTs of cardiovascular endpoints in OSA. In fact, several of these RCTs used these same techniques in secondary analyses comparing adherent and non-adherent (or control) participants and demonstrated suggestive benefits of CPAP. We propose that studies should take this approach for their primary analyses. We now discuss this approach and address concerns about its use.

What is a propensity score?

A propensity score (PS) is the probability that an individual who received the study intervention compared to control as a function of relevant baseline covariates. "Relevant baseline covariates" include all potentially confounding factors. Given two individuals with the same propensity score value, one receiving the study intervention and one the control intervention, "then we could imagine that these two subjects were 'randomly' assigned to each group in the sense of being equally likely to be treated or control" [85]. Under the assumption of no unobserved confounding (more on this below), treatment group comparisons may proceed as if participants have been randomized. Thus, "The propensity score is the observational study analog of complete randomization in randomized experiments in the sense that its use is not intended to increase precision but only to eliminate systematic biases in treatment-control comparisons" [86]. Notably, Braitman and Rosenbaum [87] suggest that propensity score methods are particularly well-suited when the outcome event is rare, but the treatment is more common; there are a large number of individuals in each treatment group, and there are many observed covariates. This is because the PS modeling focuses on the prediction of treatment rather than on the prediction of the rare outcome. This is likely the case for CPAP trials evaluating cardiovascular endpoints.

Overarching concepts in PS design application

The goal of a PS design is to create a group of treated and untreated participants, balanced with respect to baseline covariates, in which to assess differences in outcomes due to an intervention. The theoretical underpinnings of PS designs are derived from the potential outcome perspective, sometimes referred to as Rubin's Causal Model (RCM) [88]. For valid causal interpretations, it is necessary for the non-randomized groups to share the same range of propensity scores; subjects in one group who have propensity scores appreciably larger or smaller than any subject in the other group may be excluded from the chosen PS design (i.e. "trimmed"). While subjects from both arms may be trimmed, it is important to note that excluding subjects receiving the study treatment has negative implications for some causal estimates, and regulators (e.g. FDA) typically prefer no trimming in the intervention group to avoid issues with regulatory labeling.

In addition to the requirement of propensity score overlap, another fundamental principle is the separation of the PS design phase from the analysis of outcomes [86, 89]. This separation avoids the type of analysis bias that occurs with typical covariate adjustment. Best practices indicate that outcome data should be sequestered from the analyst who is determining the PS design. Unblinding of outcomes occurs only after all stakeholders reach a consensus regarding the suitability of the PS design.

In practice, to implement these approaches a first design stage similar to that of an RCT is conducted, including careful specification of hypotheses and power analyses, with an additional focus on identifying the rich set of baseline covariates associated with outcomes (e.g. composite cardiovascular endpoint) that are potentially distributed differently between the two treatment groups (e.g. those with and without adequate CPAP adherence). For propensity score studies that include prospective recruitment, increasing the planned sample size at least among controls (e.g. by 20%) should be considered to maintain power in the context of trimming. When enough time has elapsed to evaluate CPAP adherence in all enrolled subjects, a second design stage implementing the chosen PS methodology is completed. As noted above, the analyst implementing the PS design should remain blinded to any outcomes until the design is determined to be acceptable by stakeholders. When outcome data are finally unblinded, valid treatment group comparisons that account for the PS design may then be performed.

Covariate selection process

An often-cited limitation of propensity score approaches is that, unlike randomization, bias reduction can only be based on observed variables. However, bias from unobserved covariates can be removed to the extent that the unobserved covariates are correlated with the set of included PS covariates [90]. This underscores the need to include a rich set of clinically relevant covariates that are likely to be associated with any unobserved covariates. Importantly, when choosing covariates it has been shown that one should focus on those variables that are likely to be associated with the outcome. In fact, the inclusion of covariates only associated with likelihood of treatment (and not outcome) can reduce statistical precision [91]. A distinct advantage of the propensity score approach is that it allows simultaneous statistical control for a potentially much larger number of variables that could be supported in stratified analyses or a multiple variable regression model that contains individual factors. Ultimately, obtaining input from clinical experts and stakeholders regarding the most essential covariates to measure and include in a PS model is a crucial endeavor for assuring a robust estimate of the treatment effect.

Specific types of propensity score designs

There are four types of PS designs: (1) subclassification, (2) matching, (3) weighting, and (4) covariate adjustment [92]. Covariate adjustment using the PS involves the outcome variable and so violates the fundamental principle of separating the PS design from the outcomes analysis; thus, it is not further mentioned here. Subclassification, matching, and weighting are PS design approaches often capable of achieving at least as much covariate balance as an RCT, leading to effective bias reduction that improves estimates of causal effects.

In PS *subclassification*, subjects in both groups are partitioned into subclasses (e.g. quintiles) such that the PS values are relatively homogenous within the subclass. While fewer or more subclasses could be considered, five has been found to remove over 90% of the confounding due to a continuous variable [93]. Treatment effects are estimated within each subclass and then statistically pooled to determine the overall treatment effect. There are many ways to construct a PS model for use in PS subclassification [94]. Often, logistic regression is utilized. Regardless of the modeling approach, it is imperative to evaluate imbalance not only in main effects, but also in squared terms and interactions among covariates. Important higher-ordered terms should be included in the PS model for the design to result in good balance across the set of covariates. An alternative to logistic regression capable of incorporating covariate interactions is Classification and Regression Trees (CART) [95] or other machine learning approaches [96]. One method for implementing PS subclassification is to utilize a sequential heuristic with multiple steps that can be repeated as needed to estimate propensity scores, identify important higher-ordered terms, trim subjects with insufficient propensity score overlap, and, ultimately, achieve a valid PS design [97].

There are also many *matching* designs capable of reducing bias and preserving the ability to estimate meaningful causal estimands [98]. Matching on the propensity score or other metrics (e.g. Mahalanobis distance) is an effective approach—particularly when there is a surplus of controls. Matching within "caliper" guarantees that all matched pairs are "close enough," but may not allow all study treatment subjects to be matched. If 1:1 matching is employed, analysis methods appropriate for matched pairs may be used. Other methods allow multiple matches, which may be beneficial to avoid limitations in external validity when excluding treated patients.

The third approach is *weighting* individual observations according to a specific function of the propensity score, which produces specific causal effect estimates consistent with the potential outcomes approach [98]. To estimate the average treatment effect on the treated (ATT), weights are equal to 1 for treated subjects and equal to the odds of study treatment exposure (PS divided by 1 minus the PS) among control subjects [99, 100]. When applied to the difference in the proportion with a cardiovascular event, this is the causal estimate of the change in the event rate had an adherent CPAP user been a non-user instead. An alternative weighting scheme, known as the inverse probability of treatment weighting (IPTW) produces the average treatment effect (ATE). IPTW uses weights of 1/PS for treated and 1/(1 – PS) for controls. Of note, results from PS subclassification can also be used to determine the ATT estimate by weighting the subclass-specific treatment differences according to the number of study treatment subjects in each subclass. The ATE estimate can also be determined by weighting the subclass-specific treatment differences according to the total sample size per subclass. As in other designs, estimates derived using weighting should be restricted to a sample with similar PS overlap to avoid extrapolation.

Determining the success of the PS design

Ultimately, the success of the PS modeling in achieving covariate balance must be confirmed before unblinding to have confidence in adequate bias reduction. This is most readily accomplished through comparisons of standardized mean differences between treatment groups (on both the original and absolute scales) before and after accounting for the PS design [101]. Guidelines provided by Cohen [102] can be used to interpret the magnitude of these differences as small (0.2), moderate (0.5), or large (0.8). However, values as small as 0.10 may reflect a meaningful imbalance. For normally distributed covariates, a standardized mean difference of 0.10 implies a 7.7% non-overlap in distributions [101]. The allowable difference likely depends on the importance of the covariate. A simple and compelling visual comparison of successful bias reduction is summarized in a "Love Plot," as originally described by Ahmed et al. [103]. This plot summarizes standardized mean differences with and without incorporating the PS design. Graphical illustrations of the propensity score overlap are also helpful in demonstrating that the treatment groups have sufficient covariate overlap for sensible causal estimates. Often, the logit, or log odds of the PS, is used for these purposes. Notably, there may be cases in which the chosen PS approach is unable to achieve an adequate design. As such, the "propensity score technique allows the straightforward assessment of whether the treatment groups overlap enough regarding baseline covariates to allow for a sensible treatment comparison" [83].

Some common critiques of PS designs in observational data

Superiority of randomization

While randomization is thought of as superior to observational designs, the examples of recent RCTs of cardiovascular endpoints in OSA have shown there is often substantial bias in RCTs. Although a perfectly conducted RCT may provide an unbiased estimate of the treatment effect, this estimate is not all that useful when derived in the wrong target population. Studies need to be done in real-world patients [55]. Moreover, while RCTs create balance in expectation, when applied in small numbers there are often residual differences in baseline covariates. By studying real-world patients, including the excessively sleepy, and purposefully designing a sample with balanced covariates, propensity score designs are meant to directly mitigate selection bias inherent in these randomized trials. As such, in many cases, PS designs can provide more useful and less biased estimates than RCTs.

Healthy user and healthy adherer bias

Another common concern of propensity score designs in which patients select their treatment status is the potential for "healthy user" and "healthy adherer" bias [104]. Both of these biases reflect the idea that individuals who choose to use or adhere to a target therapy of interest (e.g. CPAP) are likely to use other preventative services (e.g. influenza vaccine) or be more adherent to other interventions (e.g. medications or exercise) [105]. Thus, benefits attributed to the target therapy may be caused by these underlying healthy behaviors. For example, previous cardiovascular trials have shown that individuals in the placebo arm who used the placebo had better outcomes than those who did not [106]. While this is a valid concern, established covariates capturing these effects can be directly measured and included in the set of covariates used in the PS design. There are validated instruments to assess diet [107] and exercise levels [108–110]. There are known lifestyle and socioeconomic factors that contribute to healthy user bias [111]. Information on preventative services (screening and vaccinations) and prescribed medications are increasingly available through electronic health records; these data have recently been combined to develop a "healthy user index" [112]. Healthy adherer bias can be assessed by examining prescription refills, as has been done in studies of OSA [23], as well as compliance with follow-up for clinical visits. In general, credible PS designs should include all identifiable factors that are associated with outcomes of interest and that may differ between treated and not treated subjects. Enumerating causes of potential selection bias, such as healthy user and healthy adherer bias, should be viewed as an essential step in the initial design phase of any PS design. As described below, sensitivity analyses quantifying the potential impact of unmeasured confounders can also help to mitigate these concerns if relevant covariates are unavailable.

Unmeasured or hidden confounding

The validity of causal estimates from a PS design relies on the assumption of no unobserved confounders. The credibility of this assumption is enhanced if the covariate list is comprehensive and multidimensional since this makes it more likely that any unobserved covariates are at least partially associated with and indirectly adjusted for by the set of observed covariates. Sensitivity analyses can also be performed to determine the magnitude of associations between a theoretical unobserved covariate and both the exposures and outcomes of interest that would be required to nullify the observed treatment effects [113–116]. If these magnitudes are very large, then the results from the observational study are robust with regard to unobserved covariates. In contrast, if only small associations could reverse the findings, then results from PS designs are not robust. In particular, VanderWeele and Ding [116] have recently introduced the concept of the "E-value," which can be used to perform sensitivity analyses of hidden bias with minimal assumptions. "The E-value is defined as the minimum strength of association, on the risk ratio scale, that an unmeasured confounder would need to have with both the treatment and the outcome to fully explain away a specific treatment–outcome association, conditional on the measured covariates" [116]. Routine reporting of such sensitivity analyses promises to increase the scientific credibility of the results from this approach to observational studies.

Conclusion

Recently published RCTs on the effect of CPAP on cardiovascular outcomes in patients with OSA have major challenges and biases, which likely explains the negative results. Thus, we assert that it is premature to conclude that CPAP treatment does not reduce cardiovascular events. A particular challenge is the need to include real-world patients [55] with OSA who are excessively sleepy. If one accepts that it is unethical to randomize these individuals into no treatment for a long period of time, as we do here, alternative approaches to randomization are required. Even if considered ethical [78], there is the practical barrier that symptomatic patients and/or their providers will decline participation in trials where patients will be untreated for a long period of time [59]. This is the very situation where strategies to reduce bias and derive causal estimates from observational data are of value. Thus, we propose that propensity score designs are the optimal approach to address the impact of CPAP on cardiovascular events. We appreciate that this assertion will be controversial and will require a willingness of investigators to consider new strategies. Other fields have realized the importance of this type of approach. Hopefully, this commentary will stimulate constructive discussion.

Funding

Funding provided the National Institutes of Health grant P01 HL094307.

Disclosure Statement

Financial Disclosure: A.I.P. is the John Miclot Professor of Medicine. Funds for this endowment are provided by the Philips Respironics Foundation. S.T.K. has received grant support from Philips Respironics. U.J.M. has received grant support from Hill-Rom and Philips Respironics.

Non-financial Disclosure: The authors have no potential conflict of interest.

References

1. Benjafield AV, *et al*. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *Lancet Respir Med*. 2019;**7**(8):687–698.
2. Lyons MM, *et al*. Global burden of sleep-disordered breathing and its implications. *Respirology*. 2020;**25**(7):690–702.
3. Lim DC, *et al*. Obstructive sleep apnea: update and future. *Annu Rev Med*. 2017;**68**:99–112.
4. Reinke C, *et al*. Effects of different acute hypoxic regimens on tissue oxygen profiles and metabolic outcomes. *J Appl Physiol*. 2011;**111**(3):881–890.
5. Sullivan CE, *et al*. Reversal of obstructive sleep apnoea by continuous positive airway pressure applied through the nares. *Lancet*. 1981;**1**(8225):862–865.
6. Jonas DE, *et al*. Screening for obstructive sleep apnea in adults: evidence report and systematic review for the US Preventive Services Task Force. *JAMA*. 2017;**317**(4):415–433.
7. Quan SF, *et al*. The Sleep Heart Health Study: design, rationale, and methods. *Sleep*. 1997;**20**(12):1077–1085.
8. Redline S, *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.
9. Young T, *et al*. The occurrence of sleep-disordered breathing among middle-aged adults. *N Engl J Med*. 1993;**328**(17):1230–1235.
10. Nieto FJ, *et al*. Association of sleep-disordered breathing, sleep apnea, and hypertension in a large community-based study. Sleep Heart Health Study. *JAMA*. 2000;**283**(14):1829–1836.
11. Peppard PE, *et al*. Prospective study of the association between sleep-disordered breathing and hypertension. *N Engl J Med*. 2000;**342**(19):1378–1384.
12. Shahar E, *et al*. Sleep-disordered breathing and cardiovascular disease: cross-sectional results of the Sleep Heart Health Study. *Am J Respir Crit Care Med*. 2001;**163**(1):19–25.
13. Gottlieb DJ, *et al*. Prospective study of obstructive sleep apnea and incident coronary heart disease and heart failure: the sleep heart health study. *Circulation*. 2010;**122**(4):352–360.
14. Redline S, *et al*. Obstructive sleep apnea-hypopnea and incident stroke: the sleep heart health study. *Am J Respir Crit Care Med*. 2010;**182**(2):269–277.
15. Arzt M, *et al*. Association of sleep-disordered breathing and the occurrence of stroke. *Am J Respir Crit Care Med*. 2005;**172**(11):1447–1451.
16. Gunnarsson SI, *et al*. Obstructive sleep apnea is associated with future subclinical carotid artery disease: thirteen-year follow-up from the Wisconsin sleep cohort. *Arterioscler Thromb Vasc Biol*. 2014;**34**(10):2338–2342.
17. Mehra R, *et al*.; Sleep Heart Health Study. Association of nocturnal arrhythmias with sleep-disordered breathing: The Sleep Heart Health Study. *Am J Respir Crit Care Med*. 2006;**173**(8):910–916.
18. Tung P, *et al*. Obstructive and central sleep apnea and the risk of incident atrial fibrillation in a community cohort of men and women. *J Am Heart Assoc*. 2017;**6**(7):e004500.
19. Young T, *et al*. Sleep disordered breathing and mortality: eighteen-year follow-up of the Wisconsin sleep cohort. *Sleep*. 2008;**31**(8):1071–1078.
20. Punjabi NM, *et al*. Sleep-disordered breathing and mortality: a prospective cohort study. *PLoS Med*. 2009;**6**(8):e1000132.
21. Marin JM, *et al*. Long-term cardiovascular outcomes in men with obstructive sleep apnoea-hypopnoea with or without treatment with continuous positive airway pressure: an observational study. *Lancet*. 2005;**365**(9464):1046–1053.
22. Campos-Rodriguez F, *et al*. Cardiovascular mortality in women with obstructive sleep apnea with or without continuous positive airway pressure treatment: a cohort study. *Ann Intern Med*. 2012;**156**(2):115–122.
23. Villar I, *et al*. Medication adherence and persistence in severe obstructive sleep apnea. *Sleep*. 2009;**32**(5):623–628.
24. Lavie CJ, *et al*. Exercise and the cardiovascular system: clinical science and cardiovascular outcomes. *Circ Res*. 2015;**117**(2):207–219.
25. Bowen KJ, *et al*. Nutrition and cardiovascular disease—an update. *Curr Atheroscler Rep*. 2018;**20**(2):8.
26. Després JP. Body fat distribution and risk of cardiovascular disease: an update. *Circulation*. 2012;**126**(10):1301–1313.
27. Zinchuk AV, *et al*. Polysomnographic phenotypes and their cardiovascular implications in obstructive sleep apnoea. *Thorax*. 2018;**73**(5):472–480.
28. Ye L, *et al*. The different clinical faces of obstructive sleep apnoea: a cluster analysis. *Eur Respir J*. 2014;**44**(6):1600–1607.
29. Keenan BT, *et al*. Recognizable clinical subtypes of obstructive sleep apnea across international sleep centers: a cluster analysis. *Sleep*. 2018;**41**(3):zsx214.
30. Kim J, *et al*. Symptom-based subgroups of Koreans with obstructive sleep apnea. *J Clin Sleep Med*. 2018;**14**(3):437–443.
31. Mazzotti DR, *et al*. Symptom subtypes of obstructive sleep apnea predict incidence of cardiovascular outcomes. *Am J Respir Crit Care Med*. 2019;**200**(4):493–506.
32. Pien GW, *et al*. Changing faces of obstructive sleep apnea: treatment effects by cluster designation in the Icelandic Sleep Apnea Cohort. *Sleep*. 2018;**41**(3). doi: 10.1093/sleep/zsx201
33. Sassani A, *et al*. Reducing motor-vehicle collisions, costs, and fatalities by treating obstructive sleep apnea syndrome. *Sleep*. 2004;**27**(3):453–458.
34. Javaheri S, *et al*. Response. *Chest*. 2020;**157**(4):1047–1048.
35. Jenkinson C, *et al*. Comparison of therapeutic and subtherapeutic nasal continuous positive airway pressure for obstructive sleep apnoea: a randomised prospective parallel trial. *Lancet*. 1999;**353**(9170):2100–2105.
36. Crook S, *et al*. Minimum important difference of the Epworth Sleepiness Scale in obstructive sleep apnoea: estimation from three randomised controlled trials. *Thorax*. 2019;**74**(4):390–396.
37. McDaid C, *et al*. A systematic review of continuous positive airway pressure for obstructive sleep apnoea-hypopnoea syndrome. *Sleep Med Rev*. 2009;**13**(6):427–436.

38. Siccoli MM, *et al*. Effects of continuous positive airway pressure on quality of life in patients with moderate to severe obstructive sleep apnea: data from a randomized controlled trial. *Sleep*. 2008;**31**(11):1551–1558.
39. Pepperell JC, *et al*. Ambulatory blood pressure after therapeutic and subtherapeutic nasal continuous positive airway pressure for obstructive sleep apnoea: a randomised parallel trial. *Lancet*. 2002;**359**(9302):204–210.
40. Martínez-García MA, *et al*.; Spanish Sleep Network. Effect of CPAP on blood pressure in patients with obstructive sleep apnea and resistant hypertension: the HIPARCO randomized clinical trial. *JAMA*. 2013;**310**(22):2407–2415.
41. McEvoy RD, *et al*.; SAVE Investigators and Coordinators. CPAP for prevention of cardiovascular events in obstructive sleep apnea. *N Engl J Med*. 2016;**375**(10):919–931.
42. Antic NA, *et al*. The Sleep Apnea cardioVascular Endpoints (SAVE) Trial: rationale, ethics, design, and progress. *Sleep*. 2015;**38**(8):1247–1257.
43. Sanchez-de-la-Torre M, *et al*. Effect of obstructive sleep apnoea and its treatment with continuous positive airway pressure on the prevalence of cardiovascular events in patients with acute coronary syndrome (ISAACC study): a randomised controlled trial. *Lancet Respir Med*. 2020;**8**(4):359–367.
44. Esquinas C, *et al*.; Spanish Sleep Network. Rationale and methodology of the impact of continuous positive airway pressure on patients with ACS and nonsleepy OSA: the ISAACC Trial. *Clin Cardiol*. 2013;**36**(9):495–501.
45. Peker Y, *et al*. Effect of positive airway pressure on cardiovascular outcomes in coronary artery disease patients with nonsleepy obstructive sleep apnea. The RICCADSA Randomized Controlled Trial. *Am J Respir Crit Care Med*. 2016;**194**(5):613–620.
46. Krauss A. Why all randomised controlled trials produce biased results. *Ann Med*. 2018;**50**(4):312–322.
47. Chan AW, *et al*. Epidemiology and reporting of randomised trials published in PubMed journals. *Lancet*. 2005;**365**(9465):1159–1162.
48. Cartwright N. Are RCTs the gold standard? *BioSocieties*. 2007;**2**(1):11–20.
49. Altman DG. Comparability of randomised groups. *Statistician*. 1985;**34**(1): 125–136.
50. Moher D. CONSORT: an evolving tool to help improve the quality of reports of randomized controlled trials. Consolidated Standards of Reporting Trials. *JAMA*. 1998;**279**(18):1489–1491.
51. Rennie D. CONSORT revised–improving the reporting of randomized trials. *JAMA*. 2001;**285**(15):2006–2007.
52. Moher D, *et al*. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;**340**:c869.
53. Juszczak E, *et al*. Reporting of multi-arm parallel-group randomized trials: extension of the CONSORT 2010 statement. *JAMA*. 2019;**321**(16):1610–1620.
54. Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ*. 1996;**312**(7040):1215–1218.
55. Sherman RE, *et al*. Real-world evidence—what is it and what can it tell us? *N Engl J Med*. 2016;**375**(23):2293–2297.
56. Xie J, *et al*. Excessive daytime sleepiness independently predicts increased cardiovascular risk after myocardial infarction. *J Am Heart Assoc*. 2018;**7**(2):e007221.
57. Kapur VK, *et al*.; Sleep Heart Health Study Group. Sleep disordered breathing and hypertension: does self-reported sleepiness modify the association? *Sleep*. 2008;**31**(8):1127–1132.
58. Arnardottir ES, *et al*. Obstructive sleep apnoea in the general population: highly prevalent but minimal symptoms. *Eur Respir J*. 2016;**47**(1):194–202.
59. Kushida CA, *et al*. 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.
60. Fan J, *et al*. Association of obstructive sleep apnea with cardiovascular outcomes in patients with acute coronary syndrome. *J Am Heart Assoc*. 2019;**8**(2):e010826.
61. Yu J, *et al*. Association of positive airway pressure with cardiovascular events and death in adults with sleep apnea: a systematic review and meta-analysis. *JAMA*. 2017;**318**(2):156–166.
62. Gottlieb DJ. Does obstructive sleep apnea treatment reduce cardiovascular risk?: it is far too soon to say. *JAMA*. 2017;**318**(2):128–130.
63. Baigent C, *et al*.; Cholesterol Treatment Trialists' (CTT) Collaboration. Efficacy and safety of more intensive lowering of LDL cholesterol: a meta-analysis of data from 170,000 participants in 26 randomised trials. *Lancet*. 2010;**376**(9753):1670–1681.
64. Cistulli PA, *et al*. Short-term CPAP adherence in obstructive sleep apnea: a big data analysis using real world data. *Sleep Med*. 2019;**59**:114–116.
65. Sawyer AM, *et al*. Where to next for optimizing adherence in large-scale trials of CPAP? *Sleep Med Clin*. 2020 (in press).
66. Javaheri S, *et al*. CPAP treatment and cardiovascular prevention: we need to change the design and implementation of our trials. *Chest*. 2019;**156**(3):431–437.
67. Irony TZ. The “Utility” in composite outcome measures: measuring what is important to patients. *JAMA*. 2017;**318**(18):1820–1821.
68. Lee SA, *et al*. Heavy snoring as a cause of carotid artery atherosclerosis. *Sleep*. 2008;**31**(9):1207–1213.
69. Foster GE, *et al*. Effects of continuous positive airway pressure on cerebral vascular response to hypoxia in patients with obstructive sleep apnea. *Am J Respir Crit Care Med*. 2007;**175**(7):720–725.
70. Khan SU, *et al*. A meta-analysis of continuous positive airway pressure therapy in prevention of cardiovascular events in patients with obstructive sleep apnoea. *Eur Heart J*. 2018;**39**(24):2291–2297.
71. Javaheri S, *et al*. Continuous positive airway pressure adherence for prevention of major adverse cerebrovascular and cardiovascular events in obstructive sleep apnea. *Am J Respir Crit Care Med*. 2020;**201**(5):607–610.
72. Wang J, *et al*. Risk factors for stroke in the Chinese population: a systematic review and meta-analysis. *J Stroke Cerebrovasc Dis*. 2017;**26**(3):509–517.
73. Lim DC, *et al*.; SAGIC Investigators. Reinventing polysomnography in the age of precision medicine. *Sleep Med Rev*. 2020;**52**:101313.
74. Azarbarzin A, *et al*. The hypoxic burden of sleep apnoea predicts cardiovascular disease-related mortality: the Osteoporotic Fractures in Men Study and the Sleep Heart Health Study. *Eur Heart J*. 2019;**40**(14):1149–1157.
75. Craig SE, *et al*. Continuous positive airway pressure improves sleepiness but not calculated vascular risk in patients with minimally symptomatic obstructive sleep apnoea: the MOSAIC randomised controlled trial. *Thorax*. 2012;**67**(12):1090–1096.

76. Engleman HM, *et al*. Effect of continuous positive airway pressure treatment on daytime function in sleep apnoea/hypopnoea syndrome. *Lancet*. 1994;**343**(8897):572–575.
77. Weaver TE, *et al*. Continuous positive airway pressure treatment of sleepy patients with milder obstructive sleep apnea: results of the CPAP Apnea Trial North American Program (CATNAP) randomized clinical trial. *Am J Respir Crit Care Med*. 2012;**186**(7):677–683.
78. Brown DL, *et al*. Ethical issues in the conduct of clinical trials in obstructive sleep apnea. *J Clin Sleep Med*. 2011;**7**(1):103–108.
79. Rosenbaum PR, *et al*. The central role of the prospensity score in observational studies for causal effects. *Biometrika*. 1983;**70**:41–55.
80. Rubin DB. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Serv Outcomes Res Methodol*. 2001;**2**:169–188.
81. Deb S, *et al*. A review of propensity-score methods and their use in cardiovascular research. *Can J Cardiol*. 2016;**32**(2):259–265.
82. Keenan BT, *et al*. Obstructive sleep apnoea treatment and fasting lipids: a comparative effectiveness study. *Eur Respir J*. 2014;**44**(2):405–414.
83. Yue LQ. Statistical and regulatory issues with the application of propensity score analysis to nonrandomized medical device clinical studies. *J Biopharm Stat*. 2007;**17**(1):1–13; discussion 15.
84. FDA Center for Devices and Radiological Health. *Design Considerations for Pivotal Clinical Investigations for Medical Devices: Guidance for Industry, Clinical Investigators, Institutional Review Boards and Food and Drug Administration Staff*. Silver Spring, MD: Center for Devices and Radiological Health, FDA; 2013.
85. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*. 1998;**17**(19):2265–2281.
86. Rubin DB. For objective causal inference, design trumps analysis. *Ann Appl Stat*. 2008;**2**(3):808–840.
87. Braitman LE, *et al*. Rare outcomes, common treatments: analytic strategies using propensity scores. *Ann Intern Med*. 2002;**137**(8):693–695.
88. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. *J Am Stat Assoc*. 2005;**100**(469):322–331.
89. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med*. 2007;**26**(1):20–36.
90. Zhou Z, *et al*. Discussion of: Statistical and regulatory issues with the application of propensity score analysis of nonrandomized medical device clinical studies. *J Biopharm Stat*. 2007;**17**:25–27.
91. Brookhart MA, *et al*. Variable selection for propensity score models. *Am J Epidemiol*. 2006;**163**(12):1149–1156.
92. Austin PC, *et al*. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Stat Med*. 2006;**25**(12):2084–2106.
93. Cochran WG, *et al*. Controlling bias in observational studies: a review. *Indian J Stat, Ser A*. 1973;**35** (4):417–446.
94. Imbens GW, *et al*. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge University Press; 2015.
95. Breiman L, *et al*. *Classification and Regression Trees*. Wadsworth & Brooks/Cole Advanced Books & Software. Boca Raton, FL: Chapman & Hall/CRC, Taylor & Francis Group; 1984.
96. Westreich D, *et al*. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J Clin Epidemiol*. 2010;**63**(8):826–833.
97. Maislin G, *et al*. Design of non-randomized medical device trials based on sub-classification using propensity score quintiles. *Proc Joint Stat Meet, Biopharm Stat*. 2010:2182–2196.
98. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;**25**(1):1–21.
99. Sato T, *et al*. Marginal structural models as a tool for standardization. *Epidemiology*. 2003;**14**(6):680–686.
100. Stürmer T, *et al*. Insights into different results from different causal contrasts in the presence of effect-measure modification. *Pharmacoepidemiol Drug Saf*. 2006;**15**(10):698–709.
101. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*. 2009;**28**(25):3083–3107.
102. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale: Lawrence Erlbaum Associates; 1988.
103. Ahmed A, *et al*. Heart failure, chronic diuretic use, and increase in mortality and hospitalization: an observational study using propensity score methods. *Eur Heart J*. 2006;**27**(12):1431–1439.
104. Shrank WH, *et al*. Healthy user and related biases in observational studies of preventive interventions: a primer for physicians. *J Gen Intern Med*. 2011;**26**(5):546–550.
105. Silverman SL, *et al*. Healthy users, healthy adherers, and healthy behaviors? *J Bone Miner Res*. 2011;**26**(4):681–682.
106. Granger BB, *et al*.; CHARM investigators. Adherence to candesartan and placebo and outcomes in chronic heart failure in the CHARM programme: double-blind, randomised, controlled clinical trial. *Lancet*. 2005;**366**(9502): 2005–2011.
107. Thompson FE, *et al*. Development and evaluation of the national cancer institute's dietary screener questionnaire scoring algorithms. *J Nutr*. 2017;**147**(6):1226–1233.
108. van Poppel MN, *et al*. Physical activity questionnaires for adults: a systematic review of measurement properties. *Sports Med*. 2010;**40**(7):565–600.
109. Lee PH, *et al*. Validity of the International Physical Activity Questionnaire Short Form (IPAQ-SF): a systematic review. *Int J Behav Nutr Phys Act*. 2011;**8**:115.
110. Craig CL, *et al*. International physical activity questionnaire: 12-country reliability and validity. *Med Sci Sports Exerc*. 2003;**35**(8):1381–1395.
111. Kinjo M, *et al*. Potential contribution of lifestyle and socioeconomic factors to healthy user bias in antihypertensives and lipid-lowering drugs. *Open Heart*. 2017;**4**(1):e000417.
112. Eurich DT, *et al*. Development and validation of an index score to adjust for Healthy User Bias in Observational Studies. *J Popul Ther Clin Pharmacol*. 2017;**24**(3):e79–e89.
113. Rosenbaum PR. Sensitivity to hidden bias. In: Rosenbaum PR, ed. *Observational Studies*. New York: Springer; 2002.
114. Liu W, *et al*. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prev Sci*. 2013;**14**(6):570–580.
115. Vanderweele TJ, *et al*. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology*. 2011;**22**(1):42–52.
116. VanderWeele TJ, *et al*. Sensitivity analysis in observational research: introducing the E-value. *Ann Intern Med*. 2017;**167**(4):268–274.