PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Drug-Induced Sleep Endoscopy Compared to Systematic Adenotonsillectomy in the Management of Obstructive Sleep Apnea in Children: A Systematic Review and Meta-Analysis Protocol.
AUTHORS	Prévost, Anne-Sophie; Hylands, Mathieu; Gervais, Mireille; Praud, Jean-Paul; Battista, Marie-Claude; Déziel-Malouin, Stéphanie; Lachance, Monia; Lamontagne, Francois

VERSION 1 – REVIEW

REVIEWER	Ignacio Tapia, MD, MS Children's Hospital of Philadelphia Philadelphia, PA, USA
REVIEW RETURNED	25-Jan-2019

GENERAL COMMENTS	Dear authors, Thank you for submitting your protocol fro review. It is a very interesting and timely project. I have the following minor concerns: 1-The following sentence: "This is the first review to compare DISE-guided interventions to standard adenotonsillectomy in the treatment of pediatric obstructive sleep apnea" is unnecessary as another study may be published before your review is finalized. Alternatively, you may be the first review but not necessarily the nest review. 2- Outcome measures: I would specify that the AHI is not the most perfect measure as it does not consider daytime functioning or cardiovascular complications but it is the most consistent outcome due to standardization of AASM scoring rules. 3-Analysis:I suggest analyzing apnea severity also as a continuous variable as the categories proposed are accepted but based on experts opinions. If the N is big enough, quantile regression would be a great method of analysis. 4- Socioeconomic status would be another important variable to add if available.

REVIEWER	Yorschua Jalil Hospital Josefina Martinez, Santiago de Chile
REVIEW RETURNED	04-Feb-2019

GENERAL COMMENTS	Its a little beat confuse about the role of DISE, considering that its a diagnostic tool and not a treatment such an adenoidectomy. the second objective it more likely accurate. the intervention is not the DISE. The intervention should be a different treatment strategy (non adenotonsillectomy) guided by DISE. the comparation should be DISE guided versus No DISE

	guided. could be a DISE guided adenotonsillectomy, a cobination of those; in this case could be missunderstanding.

REVIEWER	Debora Maria de Araujo Aguiar Universidade de Pernambuco, Brazil
REVIEW RETURNED	06-Feb-2019

GENERAL COMMENTS	The idea is clearly stated but the difference between the primary and secondary objective could be better described. Thank you for the submission of your study design.

REVIEWER	Ivan D. Florez Univerity of Antioquia, Colombia
REVIEW RETURNED	28-Feb-2019

GENERAL COMMENTS	My comments are focused on the Methods section as per request of the Editor. As I am not expert in the topic of this review protocol I am abstaining to provide comments on other issues. In the eligibility criteria, authors describe some exclusion criteria that are “the negative”of the inclusion criteria. Exclusion criteria are defined as those scenarios that occur in patients/articles that you have already included, but for any particular reason, you need to exclude. (occluding children younger than 1 y, makes no sense, as they haven’t been included. Authors should provide a sub-heading in which clearly describe and explain the comparator (ie.e, comparator: adenotonsillectomy without DISE?) GRADE is used to support the categorization of outcomes in page 5. Please add an appropriate reference for GRADE. I am not confortable with the approach authors are planning for the outcomes. They describe that they have a very long list of outcome that will be categorized as critical, important not critical and not important. I agree that this is a recommended approach. However, I have a couple of comments on that: this approach is mostly recommended for Guideline development groups as they need to prioritize and “filter” outcomes. I would have expected that up to this point the research teamed should have decided (after the discussion with experts), the best outcomes to consider. The list is too long and the more outcome in a review the more likely to obtain significant results because of chance. I suggest on of two approaches here: 1. Continuing like the authors are recommending: not deciding on the outcomes to include until they are analyzed and categorized by discussion. In this case my recommendation is to consider a maximum number of outcomes to choose from and state that they will focus on those that are critical and important not critical, and Not considering those outcomes categorized as “not important”. This is the approach recommended when developing Guidelines. Authors should explicitly describe that after the consultation, those outcomes will be excluded from the review. of course, this categorization need to take into consideration the recommended criteria by GRADE for “grading” these outcomes (mostly that they

	should be patient-important outcomes and avoid surrogate outcomes) 2. A more advanced approach would be to do this consultation ASAP, and take decision on the outcomes to be included (based on GRADE classification, excluding the “not important” ones. A recommended number of outcomes to consider in review is “no more than 8”. Sometimes 8-10 may be needed (providing a rationale of this decision). But more than 10 is not recommended. Searching Medline and Pubmed will very likely retrieve 98-99% similar results. I suggest to use one of them to be more efficient. Authors describe the Screening process in duplicate and the criteria to go to full text review stage. However, there is no mention of how the Full-text review stage will be undertaken. Similar to how they describe the Screening process, they should provide a short paragraph describing how the Full text review will be conducted. Authors will use Cochrane RoB tool for RCT. However, they are panning to use the CLARITY group tool for observational studies. I strongly recommend authors to use the ROBINS-I tool to assess the RoB of Non-randomized studies as recommended by Cochrane. Since THE GRADE approach is explained in page 8, the following paragraph at the bottom of page 6 is not needed: If any domain presents a potential source of bias, then the report will be graded as high risk of bias. We will evaluate the overall quality of data across studies for each outcome using GRADE methodology [16]. I suggest authors to clearly state the GRADE criteria that will be considered for the GRADE assessment. And, for instance, what I2 threshold will be used to rate down the Inconsistency criterion Why not using a Standardized mean difference to pool results from continuous different scales ? Authors plan to present RCT and nonrandomized results separately. I agree with this approach. First, one question: Does this mean authors are not planning to pool all the results from RCT and Non-RCT in one single estimate? I am not recommending the latter, I am just highlighting that authors should explicitly describe if they are or they are not going to pool all the results, and if they are going to do it, do they have any pre-specified criteria to do so? Second, I suggest authors explicitly state that as they will present results separately, they should present 2 GRADE tables (one for each effect estimate: RCTs and Non-RCTSs I like the way authors repented the pre-established variables for subgroup analyses and their rationale (describing a hypothetical direction of the effect)

	the sensitivity analysis is not completely describe. Since studies will have at least six ROB criteria evaluated, how will authors decide if a study is of low RoB or Unclear? I mean, how many criteria should be Unclear to consider the study as High RoB? Only one will be enough? GRADE description in page 8, requires the appropriate citations.

REVIEWER	Bokai Wang University of Rochester, U.S.A.
REVIEW RETURNED	08-Mar-2019

GENERAL COMMENTS	1. Change as "with a variable success rate between 12% and 83% [10, 11]." 2. The primary objective and secondary objective seem confusing. 3. In section "Investigations of heterogeneity", are hypothesis like "hypothesizing that more severe obstructive sleep apnea will be more likely to benefit from DISE-directed therapy;" related to the statistical analysis. 4. Since the data for this protocol are searched online. How to guarantee the search criteria won't impact the results of the analysis.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

1. The following sentence: "This is the first review to compare DISE-guided interventions to standard adenotonsillectomy in the treatment of pediatric obstructive sleep apnea" is unnecessary as another study may be published before your review is finalized. Alternatively, you may be the first review but not necessarily the nest review.

Response: Your point is well taken. We have replaced the sentence with the following:

(Strengths and Limitations, p.2)
• This systematic review will offer a comprehensive, rigorous assessment of the literature pertaining to the use of Drug-Induced Sleep Endoscopy for a common pediatric condition.

2. Outcome measures: I would specify that the AHI is not the most perfect measure as it does not consider daytime functioning or cardiovascular complications but it is the most consistent outcome due to standardization of AASM scoring rules.

Response: Thank you for pointing this out. We have integrated this suggestion to the manuscript as follows:

(Introduction, p.3) AHI is an imperfect metric, as it does not consider relevant variables such as cardiovascular complications or daytime functioning. It is nonetheless the outcome most likely to be consistently reported, given the standardization of scoring rules by the American Academy of Sleep Medicine [1].

3. Analysis: I suggest analyzing apnea severity also as a continuous variable as the categories proposed are accepted but based on experts opinions. If the N is big enough, quantile regression would be a great method of analysis.

Response: Thank you for this suggestion. We have added quantile regression to the protocol:

(Prespecified subgroup analyses, p.7) If the number of eligible studies is sufficient, we will also explore heterogeneity using quantile regression to analyze apnea severity as a continuous variable. We will require at least ten studies in the meta-analysis in order to perform a quantile regression, as recommended by Cochrane guidelines [2].Moreover, studies included in this analysis will need to report the estimated treatment effect, associated variance, and covariate values [3]. In order to account for the residual heterogeneity between studies, we will perform a random-effects meta-regression.

4. Socioeconomic status would be another important variable to add if available.

Response: We agree that socioeconomic status is a potentially important variable that we overlooked in our subgroup analysis plan. Since this can be represented by many different metrics (income, postal code, educational attainment, etc.), we formulated the subgroup and hypothesis as follows:

(Prespecified subgroup analyses, p.7) (7) Socioeconomic status (higher vs lower, as defined in each individual study), hypothesizing that patients of lower socioeconomic status will be more likely to benefit from DISE-directed interventions.

Reviewer: 2

1. Its a little beat confuse about the role of DISE, considering that its a diagnostic tool and not a treatment such an adenoidectomy. the second objective it more likely accurate.
the intervention is not the DISE. The intervention should be a different treatment strategy (non adenotonsillectomy) guided by DISE. the comparation should be DISE guided versus No DISE guided. could be a DISE guided adenotonsillectomy, a cobination of those; in this case could be missunderstanding.

Response: We apologize if the wording of our objectives was not sufficiently clear. Thank you for drawing attention to this fact.

We recognize that DISE is a diagnostic technology rather than a therapeutic intervention. However, there is a growing impetus for conducting randomized trials and meta-analyses of diagnostic interventions [4]. Indeed, criteria have been proposed to determine whether a specific diagnostic tool is of clinical benefit. For example, a diagnostic intervention should (1) incur fewer cost and/or side-effects while being as accurate as current tools, (2) decrease the need for other interventions, or (3) lead to superior management decisions [5] . We believe that DISE has the potential to optimize the therapies offered to patients with OSA, hence fulfilling the latter criterion.

Evaluating a diagnostic intervention's overall benefit requires studies that assess both its intrinsic characteristics and its downstream effects on management decisions and, ultimately, clinical outcomes. Randomized controlled trials are essential in this respect [6] . By extension, meta-analyses of these trials will provide the best available estimates of overall effect.

The dilemma our study seeks to address is that of a generic patient presenting with OSA. Should the patient be prescribed a DISE before deciding on a surgical plan, or should he undergo adenotonsillectomy right away? At the current time, we do not know whether incorporating DISE this way improves outcomes.

Our primary objective is to compare “DISE + targeted management” vs. “adenotonsillectomy for all patients”, with the latter group representing the current standard of care. “DISE + targeted management” will presumably consist of non-surgical management for some patients.

Our secondary objective is more in line with your suggestion. We agree that it is important to determine, within the subgroup of patient undergoing surgery, whether those selected using a pre-operative DISE have better outcomes than those that undergo adenotonsillectomy without further workup. However, this second question does not address the fundamental clinical dilemma of a patient presenting with OSA, since it ignores patients selected for DISE who are oriented towards non-surgical management strategies. It is entirely possible that DISE is of no benefit whatsoever in improving outcomes of surgical procedure yet nonetheless reduces overall surgical morbidity by avoiding unnecessary operations. As such, we do not believe that comparing surgical techniques (DISE guided vs. non-DISE-guided) is sufficient for the purpose of this review’s primary objective.

We have amended the manuscript as follows to make our objectives and rationale as clear as possible:

(Introduction, p.3) It is conceivable that DISE may improve rates of OSA cure, if the ensuing surgical intervention is better tailored to a patient’s specific anatomic abnormality. Alternatively, it is possible that DISE will only avoid unnecessary adenotonsillectomy in patients who would not have benefited from a surgical intervention either way. In this latter scenario, we would expect similar rates of OSA improvement and decreased surgical morbidity in patients undergoing DISE.

(Objectives, p.4) Our primary objective is to determine whether children with OSA should undergo DISE followed by targeted therapy, or routine adenotonsillectomy without additional preoperative workup. The latter case reflects the current standard of care. Our primary research question is therefore as follows: In children with OSA, does DISE-guided management (surgical and/or non-surgical) lead to improved cure rates (normal polysomnography), compared to first-line adenotonsillectomy without additional preoperative workup?

Our secondary objective is to determine, within the more limited subgroup of patients that ultimately undergo a surgical procedure, whether those selected with preoperative DISE have improved outcomes. Our secondary research question is therefore as follows: In children with OSA, do surgical interventions guided by pre-operative DISE lead to improved cure rates (normal polysomnography), compared to first-line adenotonsillectomy without further preoperative workup?

Reviewer: 3

1. The idea is clearly stated but the difference between the primary and secondary objective could be better described. Thank you for the submission of your study design.

Response: Thank you for your comment. We have amended the primary and secondary study objectives as follows:

(Objectives, p.4) Our primary objective is to determine whether children with OSA should undergo DISE followed by targeted therapy, or routine adenotonsillectomy without additional preoperative workup. The latter case reflects the current standard of care. Our primary research question is therefore as follows: In children with OSA, does DISE-guided management (surgical and/or non-surgical) lead to improved cure rates (normal polysomnography), compared to first-line adenotonsillectomy without additional preoperative workup?

Our secondary objective is to determine, within the more limited subgroup of patients that ultimately undergo a surgical procedure, whether those selected with preoperative DISE have improved outcomes. Our secondary research question is therefore as follows: In children with OSA, do surgical interventions guided by pre-operative DISE lead to improved cure rates (normal polysomnography), compared to first-line adenotonsillectomy without further preoperative workup?

Reviewer: 4

My comments are focused on the Methods section as per request of the Editor. As I am not expert in the topic of this review protocol I am abstaining to provide comments on other issues.

1. In the eligibility criteria, authors describe some exclusion criteria that are “the negative” of the inclusion criteria. Exclusion criteria are defined as those scenarios that occur in patients/articles that you have already included, but for any particular reason, you need to exclude. (occluding children younger than 1 y, makes no sense, as they haven’t been included.

Response: Thank you for pointing that out. It is indeed redundant information. We have amended the eligibility criteria to read as follows:

(Participants, p.4) Our population is limited to surgically naïve children (≥ 1 and < 18 years of age) with confirmed obstructive or mixed sleep apnea, defined by an obstructive apnea index (OAI) > 1/h or an obstructive apnea and hypopnea index (oAHI) >1.5/h ascertained by PSG. We will exclude studies whose populations include congenital craniofacial malformations, neurologic or muscular disease impacting respiratory function (e.g. cerebral palsy, muscular dystrophy) and patients with previous airway surgery unless these patients account for less than 10% of the total sample size or there is data available for the subgroup of patients without these characteristics. We will not exclude patients with laryngomalacia.

2. Authors should provide a sub-heading in which clearly describe and explain the comparator (ie.e, comparator: adenotonsillectomy without DISE?)

Response: We agree that this is a mandatory inclusion. We have amended the interventions section to read as follows:

(Intervention/comparator, p.4) The intervention of interest is DISE performed before a first-line surgical therapy for OSA is attempted. The comparator is adenotonsillectomy for all patients presenting with OSA without preoperative DISE. This procedure removes tissue in the nasopharynx and oropharynx, thereby potentially relieving OSA when these are the sites of obstruction [7]. Multiple techniques (e.g. cold steel, monopolar or bipolar diathermy, coblation]) are reported in the literature [8-11]. Complications of adenotonsillectomy include post-operative bleeding, pain, dehydration, post-obstructive pulmonary edema, velopharyngeal insufficiency and death.

3. GRADE is used to support the categorization of outcomes in page 5. Please add an appropriate reference for GRADE.

Response: Thank you for pointing out this omission. The reference had been added (Outcome measures, p. 5).

4. I am not confortable with the approach authors are planning for the outcomes. They describe that they have a very long list of outcome that will be categorized as critical, important not critical and not

important. I agree that this is a recommended approach. However, I have a couple of comments on that:

this approach is mostly recommended for Guideline development groups as they need to prioritize and “filter” outcomes. I would have expected that up to this point the research teamed should have decided (after the discussion with experts), the best outcomes to consider. The list is too long and the more outcome in a review the more likely to obtain significant results because of chance. I suggest on of two approaches here:

1. Continuing like the authors are recommending: not deciding on the outcomes to include until they are analyzed and categorized by discussion. In this case my recommendation is to consider a maximum number of outcomes to choose from and state that they will focus on those that are critical and important not critical, and Not considering those outcomes categorized as “not important”. This is the approach recommended when developing Guidelines. Authors should explicitly describe that after the consultation, those outcomes will be excluded from the review. of course, this categorization need to take into consideration the recommended criteria by GRADE for “grading” these outcomes (mostly that they should be patient-important outcomes and avoid surrogate outcomes)

2. A more advanced approach would be to do this consultation ASAP, and take decision on the outcomes to be included (based on GRADE classification, excluding the “not important” ones.

A recommended number of outcomes to consider in review is “no more than 8”. Sometimes 8-10 may be needed (providing a rationale of this decision). But more than 10 is not recommended.

Response: Thank you for pointing this out. We agree that our method for selecting outcomes of interest could be improved. Following consultation with our multidisciplinary group, we have selected 8 outcomes that are either “critical” or “important but not critical” for consideration in this review’s analysis. The paper has been amended as follows:

(Outcome measures, p.5) The primary outcome will be the normalization of either the obstructive apnea index (OAI ≤ 1/h) or apnea hypopnea index (AHI≤ 1.5/h). Given that we anticipate that there will be few comparative studies addressing our specific research question, we will not exclude studies based on outcomes assessed. However, we will prespecify which secondary outcomes to include in our formal analysis and GRADE summary tables. Outcomes were selected and prioritized following a consultation with otorhinolaryngologists and patient advocates. We followed GRADE recommendations and favored patient-important outcomes and those that are not surrogate outcomes [12]. Outcomes of “low importance” will not be included in our analysis. Outcomes graded as “critical” were death and acute postoperative respiratory failure. Outcomes deemed to be “important but not critical” included the proportion of patients cured of OSA, the proportion of patients undergoing an adenotonsillectomy, post-operative bleeding, the number of interventions requiring general anesthesia, overall cost and quality of life.

5. Searching Medline and Pubmed will very likely retrieve 98-99% similar results. I suggest to use one of them to be more efficient.

Response: We appreciate that there is a significant overlap between these two search methods, given that PubMed essentially searches the MEDLINE database. The protocol has been amended to refer solely to MEDLINE as follows:

(Search strategy, p.5) We will perform a search in MEDLINE, MEDLINE In-Process, CINAHL, EMBASE and The Cochrane Library (CENTRAL database).

6. Authors describe the Screening process in duplicate and the criteria to go to full text review stage. However, there is no mention of how the Full-text review stage will be undertaken. Similar to how they describe the Screening process, they should provide a short paragraph describing how the Full text review will be conducted.

Response: Thank you for highlighting this omission. The manuscript has been amended to read as follows:

(Study records, p.5) Two reviewers will independently screen titles and abstracts in duplicate using the Rayyan electronic platform (Qatari Computing Research Institute): https://rayyan.qcri.org/.

We will proceed to full-text review unless both reviewers agree to exclude a report. Both reviewers will assess full-text reports independently and in duplicate using the same electronic platform. Disagreements will be resolved by consensus or third-reviewer adjudication.

7. Authors will use Cochrane RoB tool for RCT. However, they are panning to use the CLARITY group tool for observational studies. I strongly recommend authors to use the ROBINS-I tool to assess the RoB of Non-randomized studies as recommended by Cochrane.

Response: The ROBINS-I tool indeed seems to be the most appropriate instrument available for assessing the risk of bias in non-randomized studies. The Risk of bias assessment section has been amended to read as follows:

(Risk of bias assessment, p.6) For non-randomized trials, we will use the Cochrane Collaboration's ROBINS-I tool [13, 14]. This tool is based on the principle that each non-randomized study seeks to reproduce the results of an "ideal" randomized controlled trial. Sources of bias are defined as the differences between the two studies that significantly alter the results of the non-randomized study. ROBINS-I addresses the following domains as potential sources of bias: confounding, selection bias, intervention classification, deviation from anticipated interventions, missing outcome data, method of measuring outcomes, and selective outcome reporting.

8. Since THE GRADE approach is explained in page 8, the following paragraph at the bottom of page 6 is not needed:

If any domain presents a potential source of bias, then the report will be graded as high risk of bias. We will evaluate the overall quality of data across studies for each outcome using GRADE methodology [16].

Response: Thank you for pointing that out. The paragraph in question has been removed.

9. I suggest authors to clearly state the GRADE criteria that will be considered for the GRADE assessment. And, for instance, what I2 threshold will be used to rate down the Inconsistency criterion

Response: We have modified the section of the protocol which describes the GRADE assessment to briefly describe the different criteria as well as the specific thresholds that will be used. The new section has been modified as follows:

(Interpretation of results, p.8-9) The Grading of Recommendations, Assessment, Development and Evaluation framework will be used to report the overall quality of evidence and our confidence in estimates of effect. This framework considers the overall risk of bias, imprecision, inconsistency across studies, indirectness and the likelihood of publication bias [15]. We will classify the quality of evidence for each outcome across studies as being "very low", "low", "moderate", or "high".

Confidence in effect estimates will be rated down for overall risk of bias if any study included in the analysis is graded as “high risk of bias” [16].

Imprecision refers to the width of the 95% confidence interval surrounding the overall estimate of effect for an outcome. If clinical decision-making would differ based on whether the upper or lower bound of the confidence interval represented the truth, then the outcome will be rated down for imprecision [17].

Inconsistency refers to the variation in results across different studies. We will explore inconsistency by assessing the similarity of estimates, overlap of 95% CIs, as well as the Chi-squared test (with significance established at $p<0.05$) and I2 statistic (with “substantial heterogeneity” defined as an I2 greater than 50%) [18, 19] . We will present a transparent rationale justifying the decision to rate down for inconsistency based on these factors and on whether it is explained by our a priori subgroup effects.

Indirectness refers to the degree to which clinical outcomes are surrogate rather than patient-important outcomes [20]. We will rate down for indirectness if studies fail to address patient-important outcomes directly.

Publication bias refers to the bias that is introduced to a body of evidence if positive studies are more likely to have been published than negative studies. We will rate down for publication bias if the arcsine test, Egger’s test or a visual funnel plot are suggestive of significant publication bias [21].

10. Why not using a Standardized mean difference to pool results from continuous different scales?

Response: Although we recognize that the standardized mean difference (SMD) is a very common way of pooling results from different scales, we are reluctant to rely on it for our main analysis. The SMD has several disadvantages. Since it is calculated and reported as SD units, it is difficult for clinicians to interpret. The SMD also varies according to the variability and heterogeneity of scores reported. Hence, studies of heterogeneous populations will yield smaller SMDs than trials of less heterogeneous populations, even if the true magnitude of effect is identical.

We also recognize that each method of presenting and pooling such data has its associated pitfalls. We have therefore chosen the methods described, which seek to maximize both the interpretability and transparency of our results and analyses. We will also be happy to include the SMD as a sensitivity analysis. The manuscript has been amended to include the SMD as a sensitivity analysis as well as to clarify the analysis plan as follows:

(Summarizing data and treatment effect, p.6) Outcomes reported on different scales, such as quality of life, will be presented according to the previously published recommendations of Thorlund et al. [22]. These recommendations include the use of two or more complimentary methods to present results in units that are easily interpreted by clinicians, for example as natural units of a familiar instrument or as a Number Needed to Treat. We will also present these data as standardized mean differences, as a sensitivity analysis.

11. Authors plan to present RCT and nonrandomized results separately. I agree with this approach. First, one question: Does this mean authors are not planning to pool all the results from RCT and Non-RCT in one single estimate? I am not recommending the latter, I am just highlighting that authors should explicitly describe if they are or they are not going to pool all the results, and if they are going to do it, do they have any pre-specified criteria to do so?

Second, I suggest authors explicitly state that as they will present results separately, they should present 2 GRADE tables (one for each effect estimate: RCTs and Non-RCTSs

Response: We agree that this could have been stated more clearly. RCT's and non-randomized trials will be analyzed separately. There are no conditions under which we will pool randomized and non-randomized data. Although we will present only one GRADE table, each outcome will be presented on two different rows, one for trials and the other for observational studies. Our final interpretation of results will rely on the estimate of effect providing the highest degree of certainty (e.g., data from high-quality clinical trials if available). The manuscript has been clarified as follows:

(Summarizing data and treatment effect, p.6) We will analyze and present randomized trials and observational studies separately. There are no conditions under which we will pool results from randomized and non-randomized studies.

(Interpretation of results, p.9) We will present our results in a Summary of Findings table to represent individual outcomes across studies as well as the quality of evidence for each outcome [23]. Results from observational studies and randomized trials will be presented separately as different rows within the same table. Our final interpretation of results will rely on the estimate of effect providing the highest degree of certainty (e.g., data from high-quality clinical trials if available).

12. I like the way authors repented the pre-established variables for subgroup analyses and their rationale (describing a hypothetical direction of the effect)

Response: Thank you.

13. the sensitivity analysis is not completely describe. Since studies will have at least six ROB criteria evaluated, how will authors decide if a study is of low RoB or Unclear? I mean, how many criteria should be Unclear to consider the study as High RoB? Only one will be enough?

Response: Studies for which any single domain is "high risk of bias" or "unclear risk of bias" will be considered "high risk of bias." As such, only studies that are unequivocally at low risk of bias will be included in the "low risk of bias" group. We have added the following clarification to the manuscript:

(Risk of bias assessment, p.6) For both randomized and non-randomized studies, if any domain presents a potential source of bias (unclear or high risk of bias), then the report will be graded as high risk of bias. [24].

14. GRADE description in page 8, requires the appropriate citations.

Response: Thank you for pointing out this oversight. The section has been amended – as per another reviewer's comments – and the relevant citations have been added. Th section now reads as follows:

(Interpretation of results, p.8-9) The Grading of Recommendations, Assessment, Development and Evaluation framework will be used to report the overall quality of evidence and our confidence in estimates of effect. This framework considers the overall risk of bias, imprecision, inconsistency across studies, indirectness and the likelihood of publication bias [15]. We will classify the quality of evidence for each outcome across studies as being "very low", "low", "moderate", or "high".

Confidence in effect estimates will be rated down for overall risk of bias if any study included in the analysis is graded as "high risk of bias" [16].

Imprecision refers to the width of the 95% confidence interval surrounding the overall estimate of effect for an outcome. If clinical decision-making would differ based on whether the upper or lower bound of the confidence interval represented the truth, then the outcome will be rated down for imprecision [17].

Inconsistency refers to the variation in results across different studies. We will explore inconsistency by assessing the similarity of estimates, overlap of 95% CIs, as well as the Chi-squared test (with significance established at $p<0.05$) and I2 statistic (with “substantial heterogeneity” defined as an I2 greater than 50%) [18, 19] . We will present a transparent rationale justifying the decision to rate down for inconsistency based on these factors and on whether it is explained by our a priori subgroup effects.

Indirectness refers to the degree to which clinical outcomes are surrogate rather than patient-important outcomes [20]. We will rate down for indirectness if studies fail to address patient-important outcomes directly.

Publication bias refers to the bias that is introduced to a body of evidence if positive studies are more likely to have been published than negative studies. We will rate down for publication bias if the arcsine test, Egger’s test or a visual funnel plot are suggestive of significant publication bias [21].

We will present our results in a Summary of Findings table to represent individual outcomes across studies as well as the quality of evidence for each outcome [23] . Results from observational studies and randomized trials will be presented separately as different rows within the same table. Our final interpretation of results will rely on the estimate of effect providing the highest degree of certainty (e.g., data from high-quality clinical trials if available).

Reviewer: 5

1. Change as "with a variable success rate between 12% and 83% [10, 11]."

Response: Thank you for pointing that out, that modification has been made (p.3).

2. The primary objective and secondary objective seem confusing.

Response: We agree that the wording of our objectives could be improved. We have modified the manuscript accordingly as follows:

(Objectives, p.4) Our primary objective is to determine whether children with OSA should undergo DISE followed by targeted therapy, or routine adenotonsillectomy without additional preoperative workup. The latter case reflects the current standard of care. Our primary research question is therefore as follows: In children with OSA, does DISE-guided management (surgical and/or non-surgical) lead to improved cure rates (normal polysomnography), compared to first-line adenotonsillectomy without additional preoperative workup?

Our secondary objective is to determine, within the more limited subgroup of patients that ultimately undergo a surgical procedure, whether those selected with preoperative DISE have improved outcomes. Our secondary research question is therefore as follows: In children with OSA, do surgical interventions guided by pre-operative DISE lead to improved cure rates (normal polysomnography), compared to first-line adenotonsillectomy without further preoperative workup?

3. In section "Investigations of heterogeneity", are hypothesis like "hypothesizing that more severe obstructive sleep apnea will be more likely to benefit from DISE-directed therapy;" related to the statistical analysis.

Response: The hypotheses that we have prespecified will not have an impact on the statistical analysis. We state them in the protocol only to underline our a priori hypotheses and therefore to maximize the transparency of our eventual interpretation of the results.

4. Since the data for this protocol are searched online. How to guarantee the search criteria won't impact the results of the analysis.

Response: We agree that flawed search criteria could have a profound effect on our ultimate analysis. To avoid such a pitfall, we developed our search criteria using standard systematic review methodology, with the assistance of an experience medical librarian. We also made every effort to ensure that our selection criteria are not skewing our sample, by making the search itself as inclusive as possible. However, we recognize that this risk cannot be 100% eliminated. This is one of the reasons why we believe in the importance of publishing a protocol a priori, in the interest of transparency.

REFERENCES

1. Berry, R.B., The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. 2018: American Academy of Sleep Medicine.
2. Higgins, J.P. and S. Green, Cochrane handbook for systematic reviews of interventions. 2008.
3. Thompson, S.G. and J.P. Higgins, How should meta-regression analyses be undertaken and interpreted? Statistics in medicine, 2002. 21(11): p. 1559-1573.
4. Rodger, M., T. Ramsay, and D. Fergusson, Diagnostic randomized controlled trials: the final frontier. Trials, 2012. 13(1): p. 137.
5. Guyatt, G.H., et al., A framework for clinical evaluation of diagnostic technologies. CMAJ: Canadian Medical Association Journal, 1986. 134(6): p. 587.
6. Lord, S.J., L. Irwig, and R.J. Simes, When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Annals of internal medicine, 2006. 144(11): p. 850-855.
7. Bhattacharjee, R., et al., Adenotonsillectomy outcomes in treatment of obstructive sleep apnea in children: a multicenter retrospective study. American journal of respiratory and critical care medicine, 2010. 182(5): p. 676-683.
8. Martin, G., Complications associées aux instruments chirurgicaux utilisés pour les amygdalectomies. Une, 2012.
9. Shapiro, N.L. and N. Bhattacharyya, Cold dissection versus coblation-assisted adenotonsillectomy in children. The Laryngoscope, 2007. 117(3): p. 406-410.
10. Park, A., et al., Subtotal bipolar tonsillectomy does not decrease postoperative pain compared to total monopolar tonsillectomy. International journal of pediatric otorhinolaryngology, 2007. 71(8): p. 1205-1210.
11. Glade, R.S., et al., Coblation adenotonsillectomy: an improvement over electrocautery technique? Otolaryngology—Head and Neck Surgery, 2006. 134(5): p. 852-855.
12. Guyatt, G.H., et al., GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol, 2011. 64(4): p. 395-400.
13. Sterne, J.A., et al., ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. Bmj, 2016. 355: p. i4919.

14. Schunemann, H.J., et al., GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. J Clin Epidemiol, 2018.
15. Balshem, H., et al., GRADE guidelines: 3. Rating the quality of evidence. Journal of clinical epidemiology, 2011. 64(4): p. 401-406.
16. Guyatt, G.H., et al., GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). Journal of clinical epidemiology, 2011. 64(4): p. 407-415.
17. Guyatt, G.H., et al., GRADE guidelines 6. Rating the quality of evidence—imprecision. Journal of clinical epidemiology, 2011. 64(12): p. 1283-1293.
18. Guyatt, G.H., et al., GRADE guidelines: 7. Rating the quality of evidence—inconsistency. Journal of clinical epidemiology, 2011. 64(12): p. 1294-1302.
19. Higgins, J.P., et al., Measuring inconsistency in meta-analyses. Bmj, 2003. 327(7414): p. 557-560.
20. Guyatt, G.H., et al., GRADE guidelines: 8. Rating the quality of evidence—indirectness. Journal of clinical epidemiology, 2011. 64(12): p. 1303-1310.
21. Guyatt, G.H., et al., GRADE guidelines: 5. Rating the quality of evidence—publication bias. Journal of clinical epidemiology, 2011. 64(12): p. 1277-1282.
22. Thorlund, K., et al., Pooling health-related quality of life outcomes in meta-analysis-a tutorial and review of methods for enhancing interpretability. Res Synth Methods, 2011. 2(3): p. 188-203.
23. Guyatt, G., et al., GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. Journal of clinical epidemiology, 2013. 66(2): p. 151-157.
24. Guyatt, G.H., et al., GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ, 2008. 336: p. 924-926.

VERSION 2 – REVIEW

REVIEWER	Ignacio E. Tapia, MD, MS Children's Hospital of Philadelphia University of Pennsylvania Philadelphia, PA
REVIEW RETURNED	07-Jul-2019

GENERAL COMMENTS	Dear authors, Thank you for submitting your protocol. This is an interesting question that needs to be answered. The protocol is well describes, the statistics are sound. I agree with quantile regression for this purpose. My only clarification is whether children with Down syndrome or other genetic syndromes such as 22q deletion will be included. Can you please clarify this? If so, how ill results be stratified?

REVIEWER	Ivan D Florez University of Antioquia, Colombia McMaster University, Canada
REVIEW RETURNED	24-Jun-2019

GENERAL COMMENTS	Authors have addressed all my comments and they have made changes accordingly. I am satisfied with the responses and the current version of the manuscript

REVIEWER	Bokai Wang Department of Biostatistics and Computational Biology University of Rochester

	Rochester, NY, USA
REVIEW RETURNED	17-Jul-2019

GENERAL COMMENTS	1. On line 49-50, it mentions "(5) Sex, hypothesizing that male patients will be more likely to benefit from DISE-directed interventions". Any previous research or arguments for this? 2. On line 52-53, it mentions "(6) Ethnicity (white, black or other), hypothesizing that African American patients will be more likely to benefit from DISE-directed interventions". Any previous research or arguments for this?

VERSION 2 – AUTHOR RESPONSE

Reviewer: 4

Authors have addressed all my comments and they have made changes accordingly. I am satisfied with the responses and the current version of the manuscript

Thank you!

Reviewer: 1

Dear authors,

Thank you for submitting your protocol. This is an interesting question that needs to be answered. The protocol is well describes, the statistics are sound. I agree with quantile regression for this purpose. My only clarification is whether children with Down syndrome or other genetic syndromes such as 22q deletion will be included. Can you please clarify this? If so, how ill results be stratified?

Thank you for your comments. We had a long discussion about whether include or exclude genetic syndromes like Down and 22q11 deletion. Children with those type of syndromes will be excluded. As you can see in our participant section:

“we will exclude studies whose populations include congenital craniofacial malformations, neurologic or muscular disease impacting respiratory function (e.g. cerebral palsy, muscular dystrophy) […] unless these patients account for less than 10% of the total sample size or there is data available for the subgroup of patients without these characteristics.”

We fully recognize that patients with syndromes affecting airway anatomy are an important subset of the population of interest to practitioners. However, we are concerned that any effect of the intervention may be diluted if the population included is excessively heterogeneous. Moreover, we are concerned that the sample size of included studies will be too small to allow for meaningful stratification. Given that the American Academy of Pediatrics’ guidelines do not apply to children with respiratory anomalies attributable to genetic disorders, we have decided not to include these patients.

Reviewer: 5

1. On line 49-50, it mentions "(5) Sex, hypothesizing that male patients will be more likely to benefit from DISE-directed interventions". Any previous research or arguments for this?

Thank you for pointing this out. In fact, male sex is a risk factor for OSA persisting from childhood to adolescence, in the absence of treatment. It is also a risk factor for the presence of sleep-disordered breathing syndrome in adolescents. We are not aware of evidence establishing male sex as a risk factor for OSA persistence or recurrence after adenotonsillectomy. Hence, we prefer to delete the mention of “male sex” as being an argument for benefiting more from the performance of DISE.

2. On line 52-53, it mentions "(6) Ethnicity (white, black or other), hypothesizing that African American patients will be more likely to benefit from DISE-directed interventions". Any previous research or arguments for this?

We appreciate you comment. African-American patients have been found to be at increased risk of recurrence after tonsillectomy and adenoidectomy. As stated by Amin et al.: "Gain velocity in BMI, BMI and being African American (odds ratios, 4–6/unit change/yr; 1.4/unit and 15, respectively) provided equal amounts of predictive power to the risk of recurrence of SDB."

Hence, we believe this is a relevant subgroup consideration.

Amin R, Anthony L, Somers V, et al. Growth velocity predicts recurrence of sleep-disordered breathing 1 year after adenotonsillectomy. Am J Respir Crit Care Med. 2008;177(6):654–659. doi:10.1164/rccm.200710-1610OC