Cochrane Australia

Systematic review of evidence on the clinical effectiveness of aromatherapy

Technical report prepared by Cochrane Australia

5 March 2024

Trusted evidence.
Informed decisions.
Better health.

Contents

Scope of the technical report

This Technical Report includes a complete description of the methods for the review (Appendices A, B and G), results of the search and prioritisation process (Appendix A), and abbreviations used in the report (Appendix I).

It also includes an overview of Appendices C-F which are listed below but presented in separate files.

Appendices contained in this file are in light grey rows. Those in separate files are in blue rows.

Appendix A. Study eligibility criteria, identification and selection
Appendix B. Data collection, analysis and interpretation of findings
Appendix C. Lists of excluded studies, public submissions, studies awaiting classification, studies in the evidence inventory (1 file)
Appendix D. Extended results and citations for studies included in the evidence synthesis (1 file)
Appendix E. Characteristics of studies included in the evidence synthesis and evidence inventory (4 files)
Appendix F. Risk of bias assessments for studies contributing to meta-analyses (1 file)
Appendix G. Differences between the protocol and the review and methods not used
Appendix H. Response to methodological review
Appendix I. Abbreviations and list of measures

Appendix A. Study eligibility criteria, identification and selection

Overview of Appendix A

Appendix A is comprised of Appendices A1-A7 (below). These Appendices report the methods (grey rows) and results (blue rows) from the first four stages of the review (Figure A, 1-4). These stages encompass the initial specification of questions to be addressed in the synthesis and criteria for including studies in the review, the specification and implementation of search methods, and the selection of studies. From this set of studies, we compiled information about the populations and outcomes addressed in randomised trials eligible for the review. This information was reviewed by the NHMRC, NTWC and NTREAP in order to confirm populations and outcomes for inclusion in the evidence synthesis.

Appendix A1. Review questions and criteria for considering studies for this review
Appendix A2. Search methods for identification of studies
Appendix A3. Methods for selecting studies
Appendix A4. Results of the search
Appendix A5. Prioritisation process: methods used to refine the questions addressed in the synthesis
Appendix A6. Final framework: synthesis questions and criteria for including studies in each synthesis
Appendix A7. Summary of inclusion decisions based on the final framework

Appendices A1-A3 and A5 report the pre-specified methods from the protocol endorsed by NTWC, prospectively registered on the International prospective register of systematic reviews (PROSPERO ID CRD42021268244), and published in the open access journal Systematic Reviews [1]. Appendix A6 reports the framework that resulted from the prioritisation process shown in Figure A and described in Appendix A5. The framework was finalised prior to commencing data extraction (Figure A, panel 5). It defines the scope of the evidence synthesis and specifies the synthesis questions and associated PICO (population, intervention, comparator, outcome) criteria for including studies in each synthesis. Studies that met the eligibility criteria for the review but not the evidence synthesis are reported on the evidence inventory (Appendix E3).

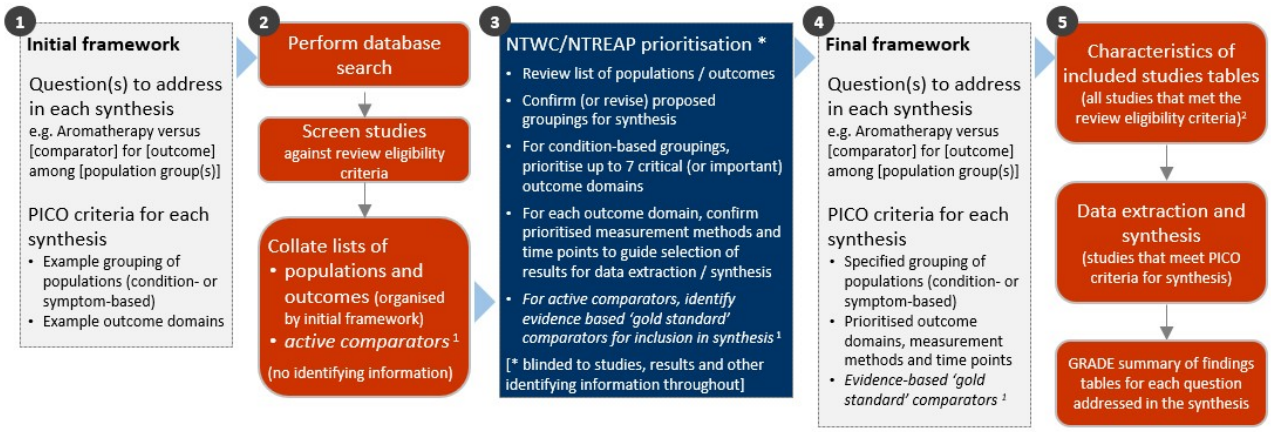

Fig A | Staged approach for developing the questions and analytic framework for this review. [1]Active comparators were not considered in the prioritisation process because there were few studies with active comparators. These studies (or study arms) are included on the evidence inventory. [2] Separate tables are presented for studies included for the evidence synthesis (Appendix E1, E2) and those in the evidence inventory (Appendix E3)

Appendix A1. Review questions and criteria for considering studies

The overall objective of this systematic review is to examine the evidence for the clinical effectiveness of aromatherapy in preventing and/or treating injury, disease, medical conditions or preclinical conditions [2]. The questions for the review follow (framed as primary and secondary objectives). An initial analytic framework for the review was presented in the protocol to illustrate the breadth of questions and a possible structure for the synthesis, with indicative populations and outcome domains (Figure A1.1). The framework was refined through the prioritisation process (described in Appendix A5) leading to the final framework and criteria for including studies in the synthesis (Appendix A6). Outcomes listed in the objectives were agreed through the prioritisation process.

Primary objectives were to address the following questions

1. What is the effect of *aromatherapy (delivered by any mode)* compared to an inactive control (placebo, no intervention or usual care) among people with any condition, pre-condition, injury or risk factor on outcomes for which aromatherapy is used (pain, nausea and vomiting, sleep quality, fatigue, emotional functioning and mental health, health-related quality of life, physical function)?
2. What is the effect of *aromatherapy (delivered by massage)* compared to *massage alone* among people with any condition, pre-condition, injury or risk factor on outcomes for which aromatherapy is used (outcomes as per objective 1)?

Secondary objectives related to the following questions

3. What is the effect of *aromatherapy (delivered by any mode)* compared to an inactive control (placebo, no intervention or usual care) on outcomes for each underlying condition, pre-condition, injury or risk factor (for example, effects on sleep disruption among people living with cancer, people with chronic insomnia, people with chronic pain, or people with dementia)?
4. What is the effect of *aromatherapy (delivered by massage)* compared to *massage alone* on outcomes for each underlying condition, pre-condition, injury or risk factor (for example, effects on sleep disruption among people living with cancer, people with chronic insomnia, people with chronic pain or people with dementia)?
5. What evidence exists examining the effects of aromatherapy compared to active comparators?
6. What evidence exists on the effects of aromatherapy compared to inactive controls or other treatments, for conditions that were not prioritised for the review?

We planned to examine the effects of aromatherapy compared to "gold standard" treatments in exceptional circumstances (studies at low risk of bias that could be combined in a synthesis). This criterion was not met.

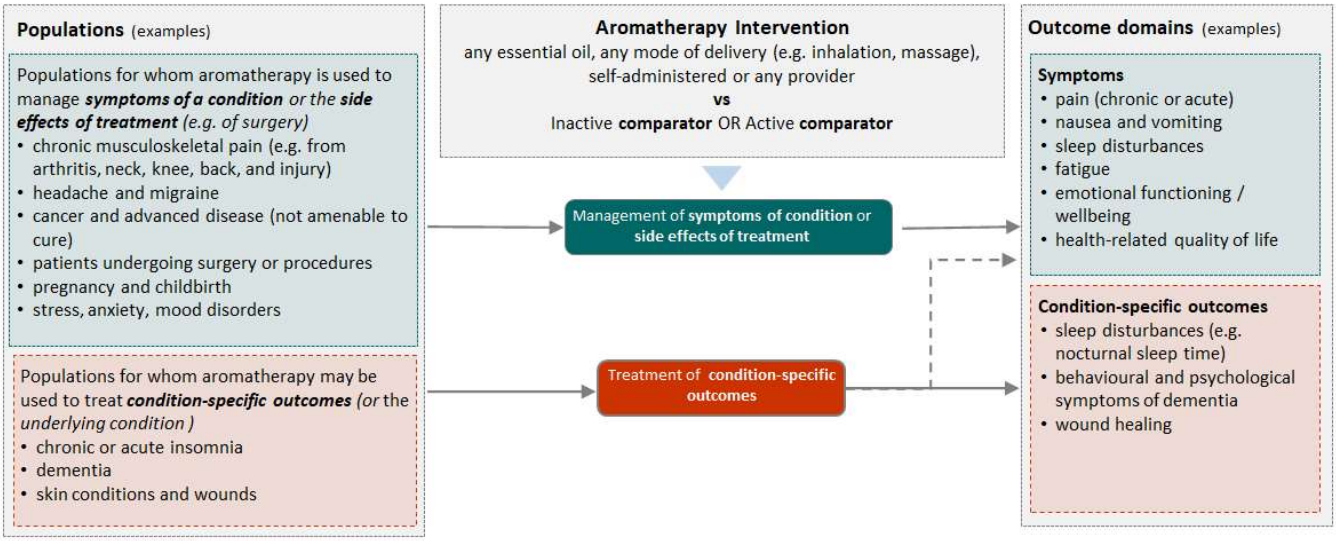

Fig A1.1 | Initial analytic framework for the review showing example population groups and outcome domains for the Evidence Synthesis. The framework was informed by research on the outcomes (and underlying conditions) for which aromatherapy is commonly sought or prescribed in Australia, a scoping search of studies evaluating aromatherapy, the wider literature on aromatherapy, and consideration of frameworks for classifying disease and outcomes [3, 4].

A1.1 Criteria for considering studies for this review

A1.1.1 Types of studies

We included randomised controlled trials (RCTs) (including individually and cluster randomised, and cross-over trials).

Controlled trials in which the allocation sequence did not include a truly random element, was predictable, or was not adequately concealed from investigators were eligible as long as there was an attempt to have some kind of 'randomisation' to groups. Examples included studies that used methods for sequence generation based on alternation, dates (of birth or attendance at a clinic) and patient record numbers [5].

We excluded:

- Non-randomised studies of interventions (NRSIs).
- Studies described as 'randomised trials' or 'controlled clinical trials', but in which decisions about the allocation of participants to treatment groups were (1) made by clinicians or participants, or (2) based on the availability of the intervention. These studies lack any 'attempt' at randomisation and, as such, are likely to be at high risk of selection bias whereby participants may be selected into groups based on factors that are prognostic of outcomes (which may introduce confounding). For the purpose of the review, these studies were considered to be non-randomised studies and excluded.
- Studies for which available reports had not been peer reviewed (grey literature, including theses).

The decision to exclude non-randomised studies was informed by scanning results from a scoping search of the Cochrane Central Register of Controlled Trials (CENTRAL) (see A2.1.1), and results of a more limited search of PubMed using a resource on the National Institute of Health National Centre for Complementary and Integrative Health website (https://www.nccih.nih.gov/health/providers/litreviews). The scoping search of CENTRAL retrieved in excess of 500 potentially eligible trials, from which we anticipated a high proportion (100-200) would meet eligibility criteria for the review. Given the likely size and breadth of the evidence base, and the proposed structure for the synthesis, we considered that any effect of aromatherapy on health outcomes should be detectable from randomised trials. The inclusion of non-randomised studies was unlikely to increase certainty of the results from a body of randomised trial evidence of this size, or alter the conclusions of the review.

Date and language restrictions.

There were no restrictions on publication date.

Potentially eligible studies published in languages other than English were eligible for the review but not eligible for synthesis. In accordance with the protocol, these studies were included in the list of studies 'Awaiting classification' and coded according to whether they were likely to be eligible or whether eligibility could not be determined (see A3.1). The impact of excluding these studies was considered in the assessment of bias due to missing results and the certainty of evidence (see B1.6 and B2.5).

A1.1.2 Types of participants

Studies involving participants with any disease, medical condition, injury, or preclinical condition were eligible for the review. This included healthy participants with clearly-identified risk factors (evident from study eligibility criteria or baseline data). There were no restrictions on age.

We expected that studies would include participants that fall within broad population groups as indicated in the initial framework Figure A1.1. The population groups were based on ICD-11 codes, and encompass conditions identified in aromatherapy literature and the PRACI survey as often treated by aromatherapists [6, 7]. Decisions about which populations to include in the evidence synthesis and how these populations would be grouped for synthesis were made through the prioritisation process (see Appendix A5) and reported in the final framework (see Appendix A6).

Excluded populations. Healthy populations seeking health improvement.

Studies that included both healthy participants and participants eligible for the review were to be included if separate data were available or a majority of participants met the review eligibility criteria [8]. No such studies were included.

A1.1.3 Types of interventions

For the purpose of this review, aromatherapy is defined as "Administration of aromatherapy oils by inhalation, diluted topical use and massage" [2].

Except for the specific exclusions below, aromatherapy treatments were eligible irrespective of the type of essential oil, carrier or dispersant, mode of delivery or route of administration, whether self-administered or provided by a practitioner, the training or qualifications of the practitioner, and the dose and duration of treatment. More details about each of these intervention features is provided under data extraction (see B1).

Excluded therapies. In line with the recommendations from aromatherapy professional associations in Australia and internationally [9-12], we excluded interventions in which the essential oil was

- ingested or administered internally (e.g. oral, vaginal, rectal or other internal routes of administration),
- applied undiluted to the skin, or
- considered unsafe for therapeutic use in humans.

Comparisons

1. Aromatherapy (delivered by any mode, including massage) *versus* any inactive comparator (placebo/sham, no intervention, wait list control, usual care).

2. Aromatherapy delivered by massage *versus* massage alone (this comparison was included to isolate the effects of aromatherapy)

Where a study includes multiple arms, with at least one eligible comparator (e.g. a placebo control arm), we include all eligible comparison(s). These comparisons form the basis of separate syntheses (meta-analyses) for each outcome domain with studies grouped within by population group (see Figure A6.1).

For the evidence inventory, we included studies that compared

3. Aromatherapy (delivered by any mode) *versus* other active comparators (including gold-standard treatments, pharmacology and non-pharmacology interventions except other natural therapies).

Excluded comparisons. In line with the main review objective, which is to examine the effects of aromatherapy rather than the comparative effects of different aromatherapy treatments, we excluded head-to-head comparisons of aromatherapy. For example:

- another essential oil or preparation of an essential oil (e.g. lavender versus ginger),
- a different dilution or dose of the same essential oil,
- a different carrier of the same essential oil,
- a different mode of delivery of the same essential oil (e.g. two different modes of inhalation; inhalation versus massage),
- where the person administering the therapy has a different qualification, specialisation or skill level (e.g. aromatherapists versus other health professional; this includes comparisons of self-administration versus administration by a practitioner),
- or combinations of the above.

A1.1.4 Types of outcomes

We considered for inclusion in the review any outcome that aligned with the reasons why aromatherapy is sought by patients and prescribed by practitioners. In principle, this could include any patient-important outcome that helps elucidate the effects of aromatherapy on an underlying condition or its symptoms, recovery, rehabilitation, or prevention of disease among people with specific risk factors or pre-conditions. Example outcome domains were

shown in the initial analytic framework to illustrate the breadth of outcomes likely to be relevant across a wide range of conditions (Figure A1.1). The outcome domains were based on ICD-11 codes and the COMET outcome taxonomy [3, 4]. These systems provide a widely agreed and understood structure for categorising different outcomes.

Studies were included in the review irrespective of the outcome(s) measured, but the synthesis was limited to outcomes considered to be critical or important for each population group. Outcomes for inclusion in the synthesis were determined through the prioritisation process described in Appendix A5.

The outcome domains determined to be critical or important for the synthesis were as follows (see Appendix A6 and Figure A6.1 for details).

- Pain
- Nausea and vomiting
- Sleep quality
- Fatigue
- Emotional functioning and mental health
- Health-related quality of life
- Physical function

From each study, we selected only one outcome per outcome domain for data extraction (results), risk of bias assessment and inclusion in the synthesis. In selecting outcomes for synthesis, we considered the outcome measure, timing of outcome measurement and data reported as follows.

Outcome measures. For each of these outcome domains, we considered for inclusion any measure of the outcome. Where studies reported multiple outcomes within an outcome domain, we used a population-specific hierarchy of outcomes measures to select the most relevant and valid outcome. The hierarchy of measures was proposed by the review team and agreed through the prioritisation process.

Outcome timing. Where trials reported outcomes measured at multiple timepoints, we selected the first measurement taken after the end of the aromatherapy intervention period (i.e. if aromatherapy was administered five times over a week, we took the first measure after the fifth administration).

Data reported

- When authors reported results for both change scores (change from baseline) and post-intervention (final) values, we selected results for final values.
- If data for the preferred measure was incompletely reported or uninterpretable, we selected another measure.

Excluded outcomes

- experience of care (e.g. satisfaction),
- safety,
- quality, and
- economic outcomes.

Appendix A2. Search methods for identification of studies

A2.1 Electronic searches

The primary source of studies was the Cochrane Central Register of Controlled Trials (CENTRAL), the most comprehensive source of published and unpublished reports of randomised trials. Most CENTRAL records are derived from regular searches of bibliographic databases, such as MEDLINE, Embase and CINAHL. Records from clinical trial registers (ClinicalTrials.gov and WHO International Clinical Trials Registry Platform) and the specialised registers maintained by Cochrane groups also make up a substantial proportion of records in CENTRAL.

As part of Cochrane's centralised search service, the major bibliographic databases and trials registers are searched monthly and, using a combination of automation and crowd screening, records deemed to be reports of randomised trials are added to CENTRAL [13]. Given the large volume of studies we anticipated would be eligible, we limited the search to CENTRAL, with supplementary searches of PubMed, AMED and Emcare, knowing that together these sources would capture a very high proportion of all relevant studies.

The search strategy for CENTRAL included the key thesaurus terms and textwords for aromatherapy, as well as more peripheral terms, such as essential oils (see Appendix A4). The most commonly used essential oils were included as textwords in their own right. This list of oils was compiled from (1) studies included in the overview of aromatherapy for the 2015 Review [35], and (2) the broader aromatherapy literature [14-21]. To ensure no commonly used essential oils were missing from the list, we examined a sample of 272 abstracts from a PubMed Clinical Query for aromatherapy (Category: 'Therapy', Scope: 'Narrow').

Since there is a lag between when records are processed by Cochrane and when they appear in CENTRAL, we ran a search in PubMed for records added since January 2021. In addition, to ensure we included records available in PubMed but which are not indexed in MEDLINE, we searched PubMed for all years, limited to the non-MEDLINE subset (see Appendix A4). We also searched AMED (Allied and Complementary Medicine) and Emcare via Ovid as these databases are not ones that Cochrane searches centrally.

Searches were run on 20 August 2021 and were not limited by language, year of publication or publication status.

A2.2 Searching other resources

Studies provided by the public and key stakeholders (via the Department), NTREAP and NTWC were deduplicated against the records retrieved by the search and screened for eligibility.

All randomised trials included in the 2015 evidence evaluation for aromatherapy were cross-checked against records retrieved by the search and considered for inclusion.

We searched PubMed for retracted publications, expressions of concern and published errata, as well as the Retraction Watch database.

We did not examine the reference lists of included studies to identify additional trials (i.e. backward citation searching), nor did we conduct forwards citation searching (i.e. looking for studies that have cited included studies). Empirical studies assessing the value of reference checking (backward citation searching) as part of the systematic review process indicate that it is most useful for areas that are difficult to search electronically (new technologies, cross-disciplinary topics, complex interventions) or for which review authors aim to locate grey literature [22]. Forward citation searching is much less common in systematic reviews [23] and of questionable value [24]. Conducting forward citation searching for the large volume of aromatherapy studies we included in this review would have generated thousands of additional records to screen, with little evidence that we would identify unique studies. This would have resulted in significant time and cost implications [25]. Given the volume of included studies, it is unlikely that any studies missed through citation searching would impact the findings of the review.

Appendix A3. Methods for selecting studies

A3.1 Selection of studies

Records from CENTRAL, PubMed, AMED and Emcare were imported into EndNote and duplicates removed. All remaining records were imported into Covidence for screening. Records submitted through the Department, NTREAP or the Committee were first deduplicated against these records, with the remaining unique records screened to confirm their eligibility (inclusion decisions were recorded for duplicate and non-duplicate records).

We piloted guidance for title and abstract screening on a sample of 50 records to ensure the eligibility criteria were applied consistently by three reviewers (SB, MM, SM). We amended the screening guidance (but not the eligibility criteria) to enhance consistency.

The trial register records retrieved from CENTRAL (i.e. from ClinicalTrials.gov and WHO ICTRP) were used to identify matching records for included studies, and a subset of the unmatched records (500) was screened to ascertain the likely number of potentially eligible trials for which there was no full text report. This was done to ascertain whether it was feasible to screen registry records to determine the number of ongoing studies and analyse records for missing results. (Some trial register records for which the source in CENTRAL was not given as ClinicalTrials.gov or WHO ICTRP were included in the records screened.) A decision was made not to screen the trial register records or the full registry entry in consultation with NHMRC given the high volume of studies eligible for the review and that an analysis of registry records would contribute little additional information.

All records were reviewed independently by two reviewers at both the title and abstract screening and full-text review stages in Covidence. Disagreements at either stage of screening were resolved by consensus among members of the review team, and advice from NTWC regarding inclusion was not required.

Protocols for studies confirmed as meeting the eligibility criteria, but for which results were not available in a published report, were included on the evidence Inventory.

While screening full-text study reports in Covidence, we extracted the trial register and registry record number (if reported) into notes in Covidence. On completion of study report screening, the list of included studies was imported into Excel, as well as registry record search results. Code was written in Excel to match any registry record details in the included study notes (e.g. registry record number) with the registry records search results.

The following categories of studies were included in a list of 'studies awaiting classification':

- Studies that were only published as abstracts or for which a full report was not available (i.e. we did not seek further information from study authors to confirm eligibility).
- Studies for which a full report was available but the report was incomplete or ambiguous such that eligibility based on one or more PICO criteria or study design could not be confirmed.
- Studies confirmed as likely to be eligible, but for which no English language translation of the full-text publication was available.
- Studies for which eligibility could not be confirmed following translation of the title and abstract using Google translate (Figure A3.1.1)
- Studies for which there were concerns about data that could not be resolved from full report(s) (e.g. where there were important discrepancies in study characteristics or data reported across multiple publications from the same study).

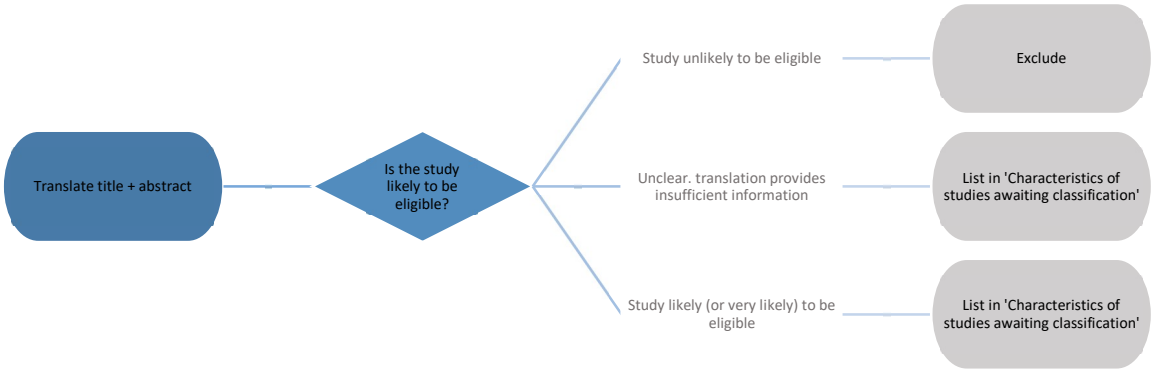

Fig. A3.1.1 | Flowchart showing handling of studies in languages other than English (reproduced from NHMRC framework for natural therapies systematic reviews [26]).

Studies that did not meet the eligibility criteria were excluded and the reason for exclusion was recorded at full-text screening. Inclusion decisions were checked at data extraction, and for any studies identified as ineligible at this stage, the decision and exclusion reason were recorded in Covidence. These studies are included in a 'Characteristics of excluded studies' table in which the reason for exclusion is reported.

The search and study selection steps are summarised in the PRISMA flow diagram in Appendix A7.

For studies that originated from the call for evidence, NTREAP, or the Committee, we recorded and reported exclusion decisions irrespective of whether the study was excluded during title and abstract screening or full text review. We documented the flow of these studies through the review in the PRISMA flow chart and in Appendix C2.

Dealing with duplicate and companion publications

Multiple publications to the same study (e.g. protocols, trial registry entries, trial reports) were identified and linked at the study selection stage in Covidence. Identification and linking of multiple reports were also checked at data extraction in REDCap [27, 28]. Each study was given a unique identifier and all linked records are cited in the final report. Records were matched using trial registry numbers. Where these were not available, we considered author names, trial name, trial location(s), number of participants, baseline characteristics and PICO.

Dealing with multiple study IDs

If multiple study reports resulted in the same study ID (Author Surname, Year) and were reporting the same study, the study ID for index report was given the suffix '.1' after the Year (e.g. Ziyaeifard 2017.1), and the study ID for the secondary report was given the suffix '.2.' (e.g. Ziyaeifard 2017.2).

If multiple study reports resulted in the same study ID (Author Surname, Year) and were reporting different studies, the study IDs for each study were given the suffix 'a', 'b', etc after the Year (e.g. Ebrahimi 2021a, Ebrahimi 2021b) to differentiate them.

Appendix A4. Results of the search

The search retrieved 4609 records. After 812 duplicates were removed in EndNote and a further 250 duplicates removed in Covidence, 3547 records were screened at title/abstract (see Table). The search strategies for each database are given below. The PRISMA flow diagram in Appendix A7 summarises inclusion decisions following title/abstract screening.

Table. Summary of sources searched and records retrieved

Source	Records retrieved	Duplicates removed	Records screened
CENTRAL	2239 (excluding 1133 records from ClinicalTrials.gov or WHO ICTRP register)	56 from EndNote; 53 from Covidence	2130
PubMed	175 added since January 2021 and 434 from PubMed-not-MEDLINE subset	140 from EndNote; 49 from Covidence	420
AMED	235	109 from EndNote; 28 from Covidence	98
Emcare	1526	507 from EndNote; 120 from Covidence	899
TOTAL	4609	1062	3547

2015 evidence evaluation for aromatherapy

The 2015 overview of 20 systematic reviews investigating the effects of aromatherapy comprised 45 unique aromatherapy trials (41 randomised trials and 4 controlled trials). Thirty-eight of the 41 RCTs were retrieved by our search and were screened along with all the other records. We retrieved the full text of the remaining three RCTs and excluded them for the following reasons: not an essential oil; trial reported in a PhD thesis; trial in a healthy population with laboratory-induced stress.

Public submissions

There were 134 records received from the public and key stakeholders (via the Department), NTREAP and NTWC. Twenty six of the 134 were unique records; all 26 were systematic reviews and therefore excluded (see Appendix C2). Eligibility decisions were for all 134 records are reported in Appendix C2.

Retractions and published errata

The PubMed search for aromatherapy was combined with the following search string, across all years: (Expression of Concern[PT] OR Corrected and Republished Article[PT] OR Published Erratum[PT] OR Retracted Publication[PT] OR Retraction of Publication[PT]).

One included study (Choi 2016.1) had a published erratum – in the abstract "geranium" should be changed to "frankincense." This error was accounted for at the data extraction stage.

One retracted study and four published errata were excluded at title/abstract screening, and one published erratum (Heydari 2019) is for a study awaiting classification (correction of author affiliation). The Retraction Watch database included the one retracted study, but no others.

Search strategies

Cochrane Central Register of Controlled Trials (Issue 8 of 12, August 2021) via Cochrane Library

#	Search strategy	Results
#1	MeSH descriptor: [Aromatherapy] explode all trees	256
#2	MeSH descriptor: [Oils, Volatile] explode all trees	416
#3	((aromather* or aroma or aromatic or ((essential or inhal* or diffus* or massag* or bergamot or cedar or chamomile or camomile or eucalyptus or frankincense or geranium or ginger or lavender or lemon or mandarin or marjoram or orange or peppermint or rose or rosemary or “tea tree” or “ti tree” or melaleuca or valerian) near (oil* or scent*)))):ti,ab,kw (Word variations have been searched)	3354
#4	#1 OR #2 OR #3	3404
#5	(clinicaltrials.gov):so	214657
#6	(www.who.int):so	159205
#7	#5 OR #6	373862
#8	#4 NOT #7	2272*
#9	#4 AND #7	1133

* 2239 in CENTRAL; 33 in CDSR

PubMed 20 August 2021

Limited to records added since 1 January 2021	Results
(Aromatherapy[Mesh] OR Oils, Volatile[Mesh] OR aromather* OR aroma OR aromatic OR ((essential OR inhal* OR diffus* OR massag* OR bergamot OR cedar OR chamomile OR camomile OR eucalyptus OR frankincense OR geranium OR ginger OR lavender OR lemon OR mandarin OR marjoram OR orange OR peppermint OR rose OR rosemary OR "tea tree" OR "ti tree" OR melaleuca OR valerian) AND (oil OR oils OR scent*))) AND ((Clinical Trial[PT] OR trial[TI] OR randomi* OR randomly OR placebo) NOT systematic[SB]) AND 01/01/2021:3000[EDAT]	175

Limited to Pubmed-not-MEDLINE subset, all years	Results
((aromather* OR aroma OR aromatic OR ((essential OR inhal* OR diffus* OR massag* OR bergamot OR cedar OR chamomile OR camomile OR eucalyptus OR frankincense OR geranium OR ginger OR lavender OR lemon OR mandarin OR marjoram OR orange OR peppermint OR rose OR rosemary OR "tea tree" OR "ti tree" OR melaleuca OR valerian) AND (oil OR oils OR scent*))) AND ((trial[TI] OR randomi* OR randomly OR placebo) NOT systematic[SB])) AND pubmednotmedline[SB]	434

AMED : allied and complementary medicine (Ovid) <1985 to August 2021>

#	Search strategy	Results

1	exp Aroma therapy/	688
2	Oils volatile/	1000
3	(aromather$ or aroma or aromatic or ((essential or inhal$ or diffus$ or massag$ or bergamot or cedar or chamomile or camomile or eucalyptus or frankincense or geranium or ginger or lavender or lemon or mandarin or marjoram or orange or peppermint or rose or rosemary or “tea tree” or “ti tree” or melaleuca or valerian) adj6 (oil$ or scent$))).af.	2527
4	exp Clinical trials/	4781
5	(trial or random$ or placebo).af.	27239
6	(1 or 2 or 3) and (4 or 5)	235

Ovid Emcare <1995 to 2021 Week 32>

#	Search strategy	Results
1	exp Aromatherapy/	1410
2	(aromather$ or aroma or aromatic or ((essential or inhal$ or diffus$ or massag$ or bergamot or cedar or chamomile or camomile or eucalyptus or frankincense or geranium or ginger or lavender or lemon or mandarin or marjoram or orange or peppermint or rose or rosemary or “tea tree” or “ti tree” or melaleuca or valerian) adj6 (oil$ or scent$))).af.	16069
3	exp Clinical Trial/	429897
4	trial.ti. or (randomi$ or randomly or placebo).af.	550847
5	(1 or 2) and (3 or 4)	1719
6	(review or meta-analysis).ti.	231651
7	5 not 6	1526

Appendix A5. Prioritisation process: methods used to refine the questions addressed in the synthesis

Decisions about the final synthesis questions and criteria for including studies in each synthesis were made through the prioritisation process in Figure A. The process was designed to minimise bias in the selection of results for inclusion in the synthesis while ensuring coverage populations and outcomes most relevant to the Australian context.

In brief

- We screened studies against the review eligibility criteria and collated information about the populations and outcomes addressed in the eligible studies.
- The information was de-identified and presented in aggregate form so that it was not possible to identify the studies (no bibliographic information, titles etc). No information was provided about the number of studies, number of participants, methodological quality of studies or results.
- NTWC/NTREAP prioritised populations and outcomes (below) and confirmed the synthesis structure.

Prioritisation of populations for inclusion in the synthesis

Studies involving any population were eligible for the review (except for the specific exclusions listed in A1.1.2), however a provision was made in the protocol to limit the populations (conditions) for inclusion in the synthesis if the number of eligible studies was unmanageable. Because of the large number of eligible studies, NTWC/NTREAP reviewed and accepted a proposal to exclude some populations from the synthesis that were not identified through the PRACI survey as commonly treated by practitioners in Australia (see A6 for exclusions).

Prioritisation and selection of outcomes for the synthesis

To prioritise the most important outcomes for this review we did the following.

- We compiled a list of population-specific outcomes from included studies and example outcome measures.
- Outcomes in the list were categorised by the outcome domains and population groups in the initial framework Figure A1.1. Outcomes that fell outside the proposed outcome domain were also listed.
- NTWC was asked to indicate whether each of the listed outcome domains (or specific outcomes) was critical, important or of limited importance for understanding the effects of aromatherapy on each population group. Only critical and important outcomes were considered in the synthesis.

Outcome selection. From each study, we selected only one outcome per outcome domain for data extraction (results), risk of bias assessment and inclusion in the summary and synthesis.

For each outcome domain, we anticipated that there would be considerable multiplicity of results arising as follows.

(1) Across studies, both the specific outcomes and methods used to measure each outcome would vary.
(2) Within studies, results may be reported for multiple outcomes within a domain (e.g. pain intensity overall, pain on walking), multiple measures (e.g. visual analogue scale; overall and subscale scores from the Western Ontario and McMaster Universities Osteoarthritis Index), at multiple timepoint, or combinations of all three.

We addressed this by using the following approach to select one outcome per domain from each study and by using the standardised mean difference (SMD) to quantify the effects of aromatherapy. Using the SMD enables results to be combined for meta-analysis irrespective of the measure used, thus ensuring that any study that reports an outcome within a domain can be included in the analysis (see B1.2 and B2.1).

- An initial hierarchy of population-specific outcomes and measures was presented to NTWC for discussion and approval (e.g. a hierarchy of pain outcomes and measures for osteoarthritis).
- Where possible, the outcome hierarchy was based on those used in published Cochrane reviews, systematic reviews of measures that provide evidence of the relevance and validity of measures, and core outcome sets.
- We also sought advice on the most relevant time point for outcome measurement.

Appendix A6. Final framework: synthesis questions and criteria for including studies in each synthesis

The final analytic framework for the review and the evidence synthesis as agreed through the prioritisation process is presented in Figure A6.1. Panel A shows the final list of populations and outcome domains eligible for the evidence synthesis. There is a meta-analysis for each outcome domain with population groups within as listed.

Population groups included in the synthesis

Some refinements were made to the populations listed in the initial framework (Panel A). We separated acute conditions and indications from chronic and longer-term conditions, to provide greater clarity about which outcomes were relevant. For example, for people with osteoarthritis undergoing knee replacement surgery, the population was categorised as 'surgery' rather than 'chronic' if treatment was focused on outcomes in the acute perioperative period rather than longer-terms outcomes. In turn, health-related quality of life, fatigue and physical function were considered relevant only to populations with chronic or longer-term conditions receiving aromatherapy treatment over weeks or longer (not days) and where outcomes were measured in a time-frame likely to detect meaningful improvement (i.e. generally 4 weeks or more from commencement of aromatherapy).

Population groups excluded from the synthesis

Given the number of studies included in the review, agreement was reached through the prioritisation process to exclude studies of aromatherapy for the treatment of skin conditions (22 studies), skin infections, infestations or wounds (20 studies), and substance withdrawal (2 studies) from the synthesis (Panel B). Outcomes specific to these population groups were also excluded from the synthesis (Panel B, blue boxes). This was necessary to ensure that the review was manageable. The characteristics of studies excluded from the synthesis on this basis are reported in Appendix E3, and a list of references for the studies and reasons for exclusion are in Appendix C3.

Prioritised outcomes

The outcome domains specified in the initial analytic framework were endorsed, and the outcomes relevant to each population groups were agreed with some refinement to the presentation in the initial framework.

- An additional outcome domain was added (physical function).
- Fatigue, health-related quality of life and physical function would be considered for chronic and longer-term conditions only.
- The outcomes listed in Panel B, white box, were not prioritised for any population. Characteristics of studies among populations eligible for the synthesis that only measured these ineligible outcomes (as determined from included reports) are reported on the evidence Inventory (Appendix E3).

Outcome measures. A hierarchy of outcome measures was also agreed for population groups. These hierarchies were used to select an outcome when a study reported multiple measures (results) for an outcome domain (e.g. subscale scores from a measure).

Timepoints. Where trials reported outcomes measured at multiple timepoints, we selected the first measurement taken after the end of the aromatherapy intervention period (i.e. if aromatherapy was administered five times over a week, we took the first measure after the fifth administration).

Panel A. Evidence synthesis

Populations	Comparisons	Outcome domains (number refers to section of report)	Population groups in each meta-analysis (number of trials and participants for Comparison 1; italics indicate that there were no studies for Comparison 2)
Acute conditions or indications Surgery Procedures (e.g. haemodialysis, biopsy, phlebotomy, dressing removal) Hospitalisation Labour and childbirth Acute musculoskeletal pain (e.g. injury) Headache (acute) Other acute pain (e.g. dysmenorrhea) Sleep disruption Mental distress (e.g. signs or symptoms of anxiety) **Chronic or longer-term conditions** Cancer and advanced disease (not amenable to cure) Chronic musculoskeletal conditions (e.g. arthritis, neck, knee and back pain) Migraine or headache (chronic or episodic) Other chronic conditions (e.g. other chronic pain, diabetes, allergic rhinitis) Chronic insomnia Mental disorders (e.g. diagnosed depression, anxiety) Dementia – behaviour change (e.g. agitation) Menopause Pregnancy and postnatal period	**Comparison 1.** Aromatherapy (any mode) versus Inactive control (no intervention, placebo, usual care) **Comparison 2.** Aromatherapy (massage) versus Inactive massage control (comparable to aromatherapy arm)	4.2 Pain	• Surgery (acute postoperative) (20 trials, 1597 participants) • Procedures (during or after) (29 trials, 2322 participants) • Labour and childbirth (9 trials, 1239 participants) • *Acute musculoskeletal conditions (1 trial, 60 participants)* • *Migraine or headache (1 trial, 141 participants)* • Other acute pain (9 trials, 855 participants) • *Cancer and advanced disease (2 trials, 338 participants)* • Chronic musculoskeletal conditions (7 trials, 347 participants) • Other chronic pain (4 trials, 294 participants)
		4.3 Nausea and vomiting	• *Surgery (acute postoperative) (10 trials, 982 participants)* • *Procedures (N&V during or after) (1 trial, 2322 participants)* • *Cancer and advanced disease (8 trials, 738 participants)* • *Pregnancy (4 trials, 271 participants)*
		4.4 Sleep quality	• *Surgery (acute postoperative) (3 trials, 227 participants)* • *Hospitalisation (not for surgery) (8 trials, 498 participants)* • *Sleep disruption (primary diagnosis or co-morbidity of a chronic or longer-term condition; 5 trials, 378 participants)* • *Cancer and advanced disease (3 trials, 163 participants)* • *Chronic insomnia (3 trials, 131 participants)* • *Dementia (0 trials)*
		4.5 Fatigue	• *Cancer and advanced disease (3 trials, 398 participants)* • Chronic musculoskeletal conditions (1 trials, 34 participants) • *Pregnancy (1 trials, 89 participants)* • Other chronic conditions (including migraine, insomnia and pain-conditions; 5 trials, 378 participants)
		4.6 Emotional functioning and mental health	• Surgery (perioperative anxiety) (17 trials, 1428 participants) • *Procedures (periprocedural anxiety) (33 trials, 2854 participants)* • Hospitalisation (12 trials, 1030 participants) • *Labour and childbirth (5 trials, 484 participants)* • Mental distress (primary diagnosis or co-morbidity of a chronic or longer-term condition; 5 trials, 440 participants) • Cancer and advanced disease (7 trials, 275 participants) • *Mental disorders (primary diagnosis or co-morbidity of a chronic or longer-term condition; 0 trials)* • Dementia (7 trials, 521 participants)
		4.7 Health-related quality of life	• Cancer and advanced disease (3 trials, 527 participants) • Other chronic and longer-term conditions (11 trials, 521 participants)
		4.8 Physical function	• *Cancer and advanced disease (1 trial, 60 participants)* • Chronic musculoskeletal conditions (6 trials, 313 participants) • *Migraine or headache (chronic or episodic; 0 trials)* • Other chronic conditions (including insomnia; 3 trials, 154 participants)

Panel B. Populations and outcomes excluded from the evidence synthesis (reported in evidence inventory)

Populations	Outcome domains (white panel applies to eligible and ineligible populations)
Skin conditions (e.g. eczema, acne, pruritis, psoriasis; 22 studies)	Severity, symptoms or flare of skin condition
Skin infections, infestations or wounds (20 studies)	Severity, signs or symptoms of skin infection or infestation (e.g. wound or ulcer size)
Neonates experiencing substance withdrawal (1 study) Substance use rehabilitation (1 study)	• Mortality or survival • Symptoms not covered by eligible domains • Cognitive function • Activities of daily living (except if measure is suitable for HR-QoL or physical function domains) • Physiological function, signs and symptoms (e.g. blood pressure, heart rate) • Anthropometric measures (e.g. weight, BMI)

Fig A6.1 | Final analytic framework for the review as agreed through the prioritisation process (Appendix A5). Panel A, column 1 lists population groups eligible for the synthesis, column 2 the outcome domains that form the basis of meta-analyses, and column 3 the populations included for each analysis. Panel B, blue boxes show populations and associated outcomes excluded from the synthesis to limit the size of the review. The white box shows outcomes that were not prioritised as critical or important for any eligible population group. Studies that only reported one or more of these ineligible outcomes are in the evidence inventory. (reproduced from main report).

Appendix A7. Summary of inclusion decisions based on the final framework

The flow of studies through the review is summarised in Figure A7.1, the PRISMA flowchart. Inclusions for each synthesis and the evidence inventory are reported in Figure A6.1 and described in the main report.

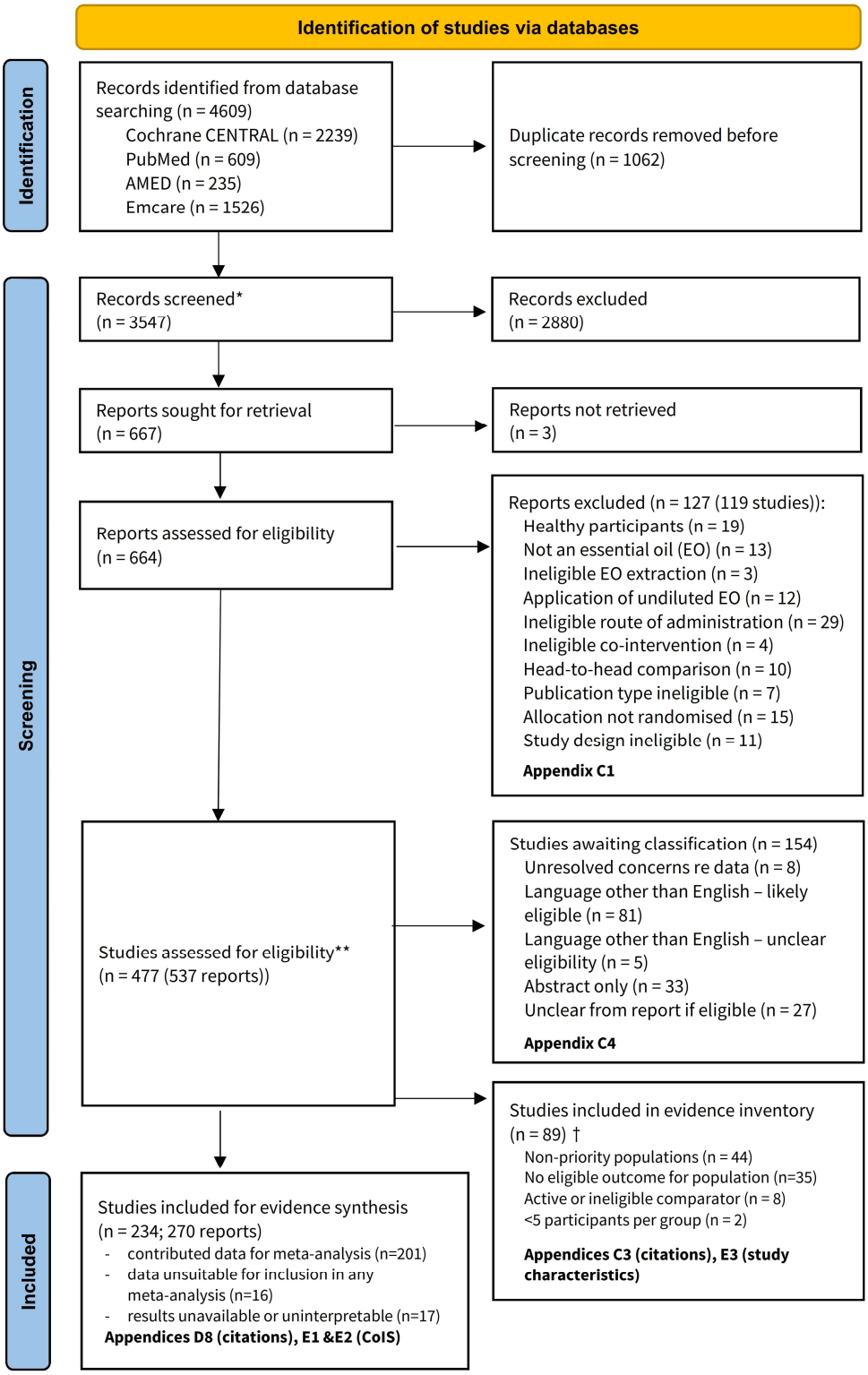

Fig. A7.1 | PRISMA diagram showing the flow of studies through the review (reproduced from main report Fig. 4.1.1). * In addition to records from the search, 134 public submissions were received and screened, of which 26 were unique records. All 26 were systematic reviews and therefore excluded (see Appendix C2). ** Studies are the unit of interest in the review. For each study there may be multiple reports. † Exclusion of these studies from synthesis was agreed through the prioritisation process (Fig 3.5.1; Methods appendix A5, A6). CoIS: characteristics of included studies.

Appendix B. Data collection, analysis and interpretation of findings

B1 Data extraction and management

Study data were collected and managed using REDCap electronic data capture tools hosted at Monash University [27, 28]. The form for extracting results data was developed by the review biostatistician (JM). Four authors (MM, SB, AS and SM) pre-tested the data extraction and coding form on 3-5 studies (as needed to achieve a consistent understanding of data fields and coding), purposefully selected from the included studies to cover a diversity of data types. All four authors discussed the coding after one author (MM) had reviewed the extracted and coded data on study characteristics for completeness, accuracy and consistency. Quantitative data was reviewed with the review biostatistician (ST or JM). Revisions to the data extraction form and guidance were made as required to maximise the quality and consistency of data collection across all data extractors. Data extractors were trained in the use of the form, and initial coding and extraction was reviewed with feedback provided prior to continuing with further studies. Frequently asked questions were logged with responses shared with all extractors to promote concordance.

To streamline the allocation of studies for analysis and selection of outcomes (when multiple results were reported for a domain), we implemented a two-step process for data extraction. In the first step, studies were triaged by a senior author (MM or SB). For each study we coded population groups, outcome domains and comparisons, and allocated the study to analyses according to the analytic framework for the review. We listed all outcomes measured and selected the outcomes for inclusion in the synthesis according to our pre-specified decision rules. The triage process included confirmation of study eligibility and basic checks of methodology (e.g. confirming that a trial met the minimum requirements for randomisation). Studies that were eligible for the review, but not the synthesis, were assigned to the evidence inventory at triage.

For each included study, one review author (KB, IF, PN, AS, ST) then extracted study characteristics and quantitative data using the data extraction and coding form. A second author (MM) independently verified the data. All queries related to the quantitative data were referred to a biostatistician, who also extracted more complex data and that from crossover trials. Discrepancies were resolved through discussion with a senior author (SB, JM) if agreement could not be reached or for more complex scenarios.

Where available, we extracted information relating to the characteristics of included studies and results as follows.

1. Study identifiers and characteristics of the study design
    - Study references (multiple publications arising from the same study were matched to an index reference; code as index paper, protocol, registry entry, results paper 1, 2, …)
    - Study name, location (country), enrolment dates (not reported by most studies), and trial registration number
    - Study design (categorised as ‘individually randomised’, ‘cluster randomised’, ‘crossover’, or ‘other’); whether clustering was likely to arise because of the way aromatherapy was delivered (e.g. at a regular clinic such as for haemodialysis; this information was used to determine which risk of bias tool to use for assessment).
    - Funding sources and funder involvement in study, financial and non-financial interests declared by investigators, potential conflicts (reviewer judgment), ethics approval.

2. Characteristics of each intervention group (including comparator groups)
    - Characteristics of the intervention covering domains of the Template for Intervention Description and Replication (TIDieR) checklist [29]
    - Aromatherapy intervention goal (coded, for example: relieve surgery-related side effects, treat underlying condition, prevent a condition among people with risk factors)
    - Coding of comparators (e.g. inactive – placebo, inactive – no intervention, inactive control – massage)
    - Number of participants: randomised to each group, at follow up for selected outcome, and included in analysis and reasons for loss to follow-up

3. Characteristics of participants

- Participant eligibility criteria (verbatim; precis of key criteria to characterise population)
- Participant characteristics: age (e.g. mean, median, range), sex
- Population group: coded using categories specified in the final analytic framework for the review (e.g. chronic musculoskeletal pain, headache or migraine, cancer and advanced disease (not amenable to cure), surgery, procedures, pregnancy, labour and childbirth, chronic insomnia, sleep disturbance, dementia, mental distress)
- Condition: specific underlying condition as described in study (e.g. haematological tumours; rheumatoid arthritis), including information about severity (if relevant) and closest ICD-11 code.
- Treatment/procedure: applied to studies in which aromatherapy was administered for the relief of symptoms or side effects of a treatment or procedure for an underlying condition (e.g. radiotherapy; bone marrow biopsy). Could include pharmacological treatment (e.g. chemotherapy), surgical, diagnostic or other procedures (as described in study).
- Other characteristics of importance within the context of each study

4. Outcomes assessed and results
    - Outcomes measured (list of all outcomes categorised as 'eligible' or 'ineligible' and categorised according to the final analytic framework; measures used for each)
    - For outcomes selected for inclusion in the summary and synthesis of results:
        - Outcome domain: categorised according to the outcome domains specified in the final analytic framework for the review (e.g. pain, nausea and vomiting, sleep quality, fatigue, emotional functioning and mental health, health-related quality of life, physical function)
        - Outcome as described in the included study (verbatim or precis)
        - Measurement method (e.g. WOMAC; overall score and pain, function and stiffness subscales), information required to interpret the measure (scale range and direction, minimally important difference) and timing of outcome measurement (exact timing; described in relation to timing of aromatherapy (e.g. immediately after end of aromatherapy intervention period) and other treatment (4, 8 and 12 hrs post-surgery)
        - Results including: summary statistics by group (means and standard deviations, or number of events for cognitive outcomes that have been dichotomised, and sample size), estimates of intervention effect (e.g. mean differences or adjusted mean differences), confidence intervals, t-values, p-values, or risk ratios/odds ratios for binary outcomes.
        - Data required to support risk of bias judgements (see Assessment of risk of bias of included studies) [30]

B1.1 Assessment of risk of bias of included studies

B1.1.1 Assessment of risk of bias in RCTs

We assessed the risk of bias in included studies using the revised Cochrane 'Risk of Bias' tool (RoB 2) for randomised trials [5, 30] for each outcome included in the synthesis. For cluster trials and cross-over trials, we used the variant of the RoB 2 tool specific for the design [31]. We also used the cluster trial RoB 2 tool for studies in which clustering effects were likely (e.g. those where our assessment was based on the effect of assignment to the intervention).

RoB 2 addresses five domains:

- bias arising from the randomisation process;
- bias due to deviations from intended interventions;
- bias due to missing outcome data;
- bias in measurement of the outcome;
- bias in selection of the reported result.

To promote concordance, the assessment was piloted by four review authors (SB, MM, SM, AS) on three studies across a range of scenarios. Based on this, review-specific guidance was developed for other assessors (KB, IF, PN, AS). One review author (KB, IF, PN, AS) then applied the tool to the selected results from each study following the RoB 2 guidance

[5], and a second author (SB) checked assessments. In the initial phases, areas of uncertainty were discussed and frequently asked questions were logged with responses shared with all extractors to promote concordance. Advice was sought from the lead reviewer (SB) where there was uncertainty, and the review biostatistician (JM) for more complex scenario (such as those arising in studies with clustering or more complex analyses). Supporting information and justifications for judgements for each domain (low, some concerns, high risk of bias) was recorded. We derived an overall summary of the risk of bias from each assessment, following the algorithm in the RoB 2 guidance as implemented in the Excel assessment tool [5].

When multiple effects of the intervention using different approaches were presented in the trial report, we selected one effect for inclusion in the meta-analysis and for risk of bias assessment. The selected effect was chosen according to the following hierarchy, which orders the approaches from (likely) least to most biased for estimating the *effect of assignment to the intervention*: 1. the effect that corresponds to a full intention-to-treat analysis, where missing data have been multiply imputed, or a model-based approach has been used (e.g. likelihood-based analysis, inverse-probability weighting); 2. the effect corresponding to an analysis that adheres to intention-to-treat principles except that the missing outcome data are excluded; 3. the effect that corresponds to a full intention-to-treat analysis, where missing data have been imputed using methods that treat the imputed data as if they were observed (e.g. last observation carried forward, mean imputation, regression imputation, stochastic imputation); or 4. the effect that corresponds to an 'as-treated' or 'per-protocol' analysis, or an analysis from which eligible trial participants were excluded [5, 30]. The effect used in the assessment was recorded in the data extraction form.

B1.2 Measures of treatment effect

We anticipated that many of the outcomes would be continuous (e.g. pain, anxiety), and that varying measurement instruments would be used to measure the same underlying construct across the studies. For this reason, we quantified the effects of aromatherapy using the standardised mean difference (SMD) (implementing the Hedges' adjusted *g* version). In trials where a continuous measure had been dichotomised (e.g. a continuous pain scale is dichotomised into improvement or no improvement) and analysed as binary outcomes, we re-expressed reported, or calculated, odds ratios as SMDs [32]. We did not report any of our meta-analysis results as dichotomous outcomes.

B1.2.1 Interpretation of treatment effects

Given the wide range of conditions, outcomes and measurement methods reported in the studies included in this review, it was not possible to specify thresholds for interpreting the size of the effect for each outcome measure. We planned to use Cohen's guiding rules for interpreting SMDs where 0.2 represents a small effect, 0.5 a moderate effect, and 0.8 a large effect [33]. In practice, our interpretation was based on whether there was an important effect or not [34, 35], with an SMD of 0.2 standard units set as the threshold for an important difference. If the SMD fell within the pre-specified range of -0.2 to 0.2 (i.e. within both thresholds), the effect of aromatherapy was considered to be no different from control. An SMD above 0.2 or below -0.2 was interpreted as an important effect. We opted to use the most intuitive interpretation of effect estimates for each outcome, so positive values indicate benefit for some outcomes (an increase in sleep quality, health-related quality of life, and physical function) and harm for other outcomes (an increase in pain, nausea and vomiting, anxiety or agitation, and fatigue). Because we were concerned that bias may be leading to exaggerated effect sizes, we chose not to describe the size of effect (i.e. we did not interpret effects as small, moderate or large).

B1.3 Unit of analysis issues

In this review, unit of analysis issues arose from non-standard designs (cluster trials, cross-over trials) and from trials with more than two eligible intervention groups (arms). In the following we outline the methods that were used for making adjustments when necessary. Any adjustments were indicated on the forest plots and documented (e.g. assumed intra-cluster correlation and average cluster size). Studies for which we were unable to make the necessary adjustments due to missing information are listed in Appendix E4.

For cluster randomised trials that had not appropriately accounted for correlation in observations within clusters, we attempted a re-analysis. We did this by inflating the variance of the intervention estimates by a design effect (DEFF). The

DEFF is calculated from two quantities – an intra-cluster correlation (ICC) and the average cluster size. Estimates of ICC were imputed from other cluster trials included in the review, where possible, or by using external estimates from empirical research (e.g. Bell 2013 [36]). The average cluster size was calculated from reported information in the trial.

For cross-over trials where an appropriate paired analysis was not available, we attempted to approximate a paired analysis by imputing missing statistics (e.g. correlation). Estimates of the missing statistics were imputed from other cross-over trials included in the review, where possible, or by using external estimates from empirical research (e.g. Balk 2012 [37]).

For trials where more than one comparison from the same trial is eligible for inclusion in the same meta-analysis (e.g. lavender oil, ginger oil, different dose of the same oil, multiple inactive control groups), we combined intervention groups, where it made sense to do so; otherwise, we reduced the sample size so that the same participants did not contribute more than once.

B1.4 Dealing with missing data

As planned in the protocol, we did not contact trial authors to obtain missing information (e.g. study characteristics, description of conduct of the trial) or aggregate level statistics (e.g. missing standard deviations). However, we attempted to calculate statistics necessary for meta-analysis using algebraic manipulation of reported statistics (e.g. computing the standard error for the treatment effect from a reported p-value). When standard deviations could not be calculated from available statistics, but interquartile ranges or ranges were reported, we used the formula in Wan et al [38] to estimate approximate standard deviations. When neither of the above methods were possible, we imputed the standard deviation using the average standard deviation across trials included in the same meta-analysis that used the same measurement tool in a similar population. When means were missing, but medians are reported, we used the formula in Wan et al [38] to estimate approximate means. Studies for which we calculated or imputed statistics are annotated in forest plots, and the impact of these decisions was explored in sensitivity analyses (see B2.4 sensitivity analyses and Appendix D for results). Studies for which we could not calculate or impute the statistics required for inclusion in the meta-analysis are listed in Appendix E4 with reasons for why the results could not be included.

We planned to deal with missing outcome data within the primary trials through sensitivity analyses, where trials judged to be at a high risk of bias or some concerns would be excluded; however, this was not possible because none of the trials included in the review were at low risk of bias (see B2 Data synthesis)[1]. Risk of bias 'due to missing outcome data' was considered within the overall bias judgement for each trial.

B1.5 Assessment of heterogeneity

We assessed statistical heterogeneity of the intervention effects visually by inspecting the overlap of confidence intervals on the forest plots, through formal tests for heterogeneity using the χ^2 test (using a significance level of α=0.1), and quantified heterogeneity using the I^2 statistic [39]. When there was evidence of heterogeneity, we judged its importance by considering where the point estimates for studies lay in relation to the threshold for an important difference (all on one side, indicating similar interpretations across the studies, or not).

B1.6 Assessment of biases due to missing results

We planned to use a framework for assessing risk of bias due to missing results in which an assessment is made for each meta-analysis regarding the risk and potential impact of missing results from studies in which we knew an outcome was measured but not reported (termed 'known-unknowns') and the risk of other missing studies or results (termed 'unknown-unknowns') [40]. The assessment of 'known-unknowns' involves assessment of whether trials meeting the

[1] In the protocol we reported that we would conduct sensitivity analyses excluding trials judged at high or unclear risk of bias. The terminology "Unclear risk of bias" has been replaced in ROB2 with "some concerns". The approach described here is consistent with the protocol in that the sensitivity analyses were to be restricted to studies at low risk of bias.

inclusion criteria for a particular meta-analysis have missing results through examination of the publication's methods section, trial registry entry (if available), and trial protocol (if available). In practice, the assessment of 'known-unknowns' was not feasible due to the large number of included studies and additional studies in trial registers. This assessment was therefore limited to examining the potential impact of studies for which data could not be included in the meta-analysis and those in languages other than English that were judged as being likely to meet the eligibility criteria for each synthesis (see A1.1.1 Types of studies; A3.1 Selection of studies). For the former, we made an assessment as to whether the missing result was potentially due the result itself (e.g. 'not statistically significant'), and whether inclusion of the result could lead to a notable change in the meta-analysis (e.g. if the missing result is from a large trial). We also considered whether there was evidence of selective non-reporting of results from the assessment of 'unknown unknowns' which would mean the synthesis result would already be downgraded for publication bias.

In assessing 'unknown-unknowns', we judged whether the trials not identified were likely to have results eligible for inclusion (i.e. for the outcome domain 'pain', is it likely that missing studies would have been eligible for inclusion in the overall analysis or for particular conditions). We used contour enhanced funnel plots to examine whether there was evidence of small study effects [41]. We also undertook sensitivity analyses to compare the combined effect estimated from the random-effects model (primary analysis) with that estimated from a fixed (common) effect model. If there was evidence of funnel plot asymmetry, and the random-effects estimate was importantly larger than the fixed-effect estimate, with no explanation for the difference (e.g. differences in populations or intensity of the delivery of intervention between small and large trials, differences in risk of bias between small and large trials), then we downgraded for 'suspected' reporting (publication) bias.

B2 Data synthesis

B2.1 Meta-analysis

Separate comparisons were set up based on outcome domains agreed in the final framework (see Figure A6.1 Appendix A6). These comparisons were stratified by the population groups in the final framework. This approach to structuring the meta-analysis yielded an overall estimate of the effect of aromatherapy for the outcome (review objectives 1 and 2), as well as estimates within each population group (review objective 3 and 4). Subgroup analysis by population group was used to examine whether these population groups explained any observed statistical heterogeneity in the intervention effects (see Subgroup analysis).

We combined the effects using a random effects meta-analysis model, since we expected and found there to be clinical and methodological diversity across the trials that may contribute to statistical heterogeneity. These analyses used the restricted maximum likelihood estimator (REML) of between trial heterogeneity variance and the Hartung-Knapp-Sidik-Jonkman confidence interval method. Analyses were conducted in Stata Statistical Software [42].

Forest plots were used to visually depict the intervention effect estimates and their confidence intervals. Forest plots are stratified by condition and risk of bias (within population group).

B2.2 Summary and synthesis when meta-analysis is not possible

Studies that were eligible for the evidence synthesis but could not be included in meta-analyses, are included in the characteristics of included studies table (Appendix E1). These studies are counted as 'missing results' rather than included in a summary or other synthesis (i.e. the result was judged to be uninterpretable or there were major concerns about the integrity of the data). Details of the syntheses for which each of the studies was eligible are tabulated, together with the number of participants and the reason why data are missing (Appendix E4). We do not report the results from these studies (if available) because of concerns about the validity of the data and because the individual studies are unlikely to change the findings from the meta-analyses (i.e. we were able to include the majority of studies in meta-analyses). Nor did we assess risk of bias because bias (under- or over-estimating the effect) is only relevant if results are included in a meta-analysis or reported. The reasons why these studies were not included in the analysis do not relate to bias (i.e. incomplete reporting of effects and their variances, errors in reporting or analysis of data, no information to interpret), so a risk of bias assessment would not characterise the problems with these studies.

B2.3 Subgroup analysis and investigation of heterogeneity

We undertook a subgroup analysis to examine whether population group explains any observed statistical heterogeneity in the intervention effects, using the pre-defined groups specified in the final framework (see Figure A6.1 for population groups in each meta-analysis). In addition, for the comparison aromatherapy versus inactive comparator, we examined whether mode of delivery (massage or 'other' including inhalation and topical application) explained any observed statistical heterogeneity in the intervention effects (Results in Appendix D).

B2.4 Sensitivity analyses

We undertook and report sensitivity analyses examining if the meta-analysis estimates were robust to the:

- *meta-analysis model*. In addition to fitting a random-effects model, we fitted fixed effect models. The analysis was undertaken to investigate the impact of any small-study effects.
- *assumptions made to enable inclusion of results in the meta-analysis,* specifically (1) transforming or imputing statistics, and (2) including change scores (change from baseline) when post-intervention (final) values (and their standard deviations) were unavailable.

Results of the sensitivity analyses were tabulated, including the meta-analysis estimate (and its confidence interval), along with details of the original and sensitivity analysis assumptions (Appendix D).

We also planned to undertake a sensitivity analyses examining if the meta-analysis estimates were robust to inclusion of trials judged to be at an overall high risk of bias or some concern. We planned to exclude trials judged to be at an overall high risk of bias or some concerns; however, there were no trials judged to be at low risk of bias in the review, so these sensitivity analyses could not be performed.

B2.5 Summary of findings tables and assessment of certainty of the body of evidence

We prepared GRADE summary of findings tables for each of the main comparisons, reporting results for critical and important outcome domains (up to seven). For each result, one author (SB) used the GRADE approach to assess our confidence in where the effect lies relative to our threshold for a small effect (the certainty of evidence) (see Measures of treatment effect). In accordance with detailed GRADE guidance [35, 43, 44], an overall GRADE of high, moderate, low or very low certainty is reported for each result based on whether there are serious, very serious or no concerns in relation to each of the following domains [34].

1. **Risk of bias**. We assessed the overall risk of bias across all studies contributing to each synthesised result. All studies were rated at high risk of bias or some concerns (i.e. contributed 100% of the weight in all meta-analyses). As such, it was not possible to perform sensitivity analyses to examine whether removing studies at high risk of bias or some concerns changed the direction or size of effect estimate importantly (a reduction in benefit or an increase in harm being most concerning) (see Sensitivity analyses). We therefore rated down all results for risk of bias. Where the majority of studies were at high risk of bias, we rated down for very serious concerns.
2. **Imprecision**. We judged imprecision by examining where the 95% confidence interval for each pooled effect estimate lay in relation to our threshold for an important effect (an SMD of -0.2 or 0.2; see Measures of treatment effect). Where the confidence interval crossed a threshold leading to different interpretations (e.g. interpretation of the upper bound of the interval was 'an important effect' and the lower bound 'little or no effect'), we considered rating down for imprecision. If the extent to which the confidence interval crossed the threshold was modest, and the interpretation was consistent with the point estimate, we did not rate down (e.g., if the upper bound of the confidence interval was an SMD of -0.15 and the point estimate -0.50). We rated down for serious imprecision if the confidence interval crossed one threshold (important benefit or important harm) and the interpretation of either the upper or lower bound of the interval was different from the point estimate (e.g. if the upper bound of the confidence interval was an SMD of 0.40 indicating an important increase in pain, and the point estimate was -0.15 indicating an unimportant reduction in pain). We rated down for very serious imprecision if the confidence interval

crossed two thresholds (important benefit and important harm). In line with GRADE guidance, we considered the likely impact of inconsistency when rating imprecision since inconsistency can contribute to imprecision [45, 46].

3. **Inconsistency**. We assessed whether there was important, unexplained inconsistency in results across studies considering the overlap of confidence intervals (non-overlap indicating potentially important differences in direction or size of effect), statistical measures that quantify and test for heterogeneity (I^2 statistic, χ^2 test), and where the point estimates lie in relation to the threshold for an important effect (if all to one side of a threshold, we were less concerned). To enhance our interpretation of whether inconsistency is important, we also examined the prediction interval, considering whether it included values that lead to a different conclusion than an assessment based on the confidence interval [47]. Where there was evidence of importantly inconsistent results, we considered whether the results of subgroup analyses provided a credible explanation for the inconsistency (see Assessment of heterogeneity; specifically, the population subgroups and whether aromatherapy was delivered by massage or not). Where inconsistency was not explained, we rated down. Where a result was based on a single study, inconsistency was not rated [45].
4. **Indirectness.** We assessed whether there are important differences between the characteristics of studies included in each synthesis and the question we were seeking to address, such that the effects observed may not apply to our question (i.e. the applicability of the evidence). For example, differences between the interventions delivered and aromatherapy practice in Australia that are likely to influence the size of effect. Where results came from a single small study, we were concerned that similar effects might not be observed in the population of interest more generally, and rated down for serious indirectness. Where the included studies addressed only part of the population of interest (e.g. the only form of acute pain was dysmenorrhea), we did not rate down for indirectness. Instead, we specified the population from which data came when interpreting results and indicated uncertainty for the population group more generally.
5. **Publication bias**. Our judgement of publication bias was based on assessment of bias due to missing results, primarily from interpretation of contour enhanced funnel plots (see Assessment of biases due to missing results). In these assessments, we also considered the potential impact of excluding studies in languages other than English and of data that could not be included in the meta-analyses.
6. **Upgrading domains** (large effect size, dose response gradient, opposing plausible residual confounding). While, in principle, these domains apply to randomised trials, there is no precedent for rating up the evidence from randomised trials, and we did not have reason to apply them in this review.

Using GRADE decision rules, we derived a GRADE for the certainty of evidence for each result included in the summary of findings table [44]. A result from a body of evidence comprised of randomised trials begins as 'high' certainty evidence (score=4), and can be rated down (-1, -2 or -3) for serious, very serious or extremely serious concerns on any GRADE domain that reduces confidence that aromatherapy has an important effect (as determined by the pre-specified thresholds) [43, 44, 48].

Summary of findings tables were prepared using the GRADEpro GDT software [35]. The tables include:

- estimates of the effects of aromatherapy reported as standardised mean differences
- the overall GRADE (rating of certainty) and an explanation of the reason(s) for rating down (or borderline decisions) [49].
- the study design(s), number of studies and number of participants contributing data
- a plain language statement interpreting the evidence for each comparison and outcome, following GRADE guidance for writing informative statements (see B2.6 interpretation of findings) [50].

We present the certainty of evidence in summary of findings tables using one of four levels with the following symbols and interpretations.

Certainty	GRADE interpretation	Implications
High (⨁⨁⨁⨁)	we are very confident that the true effect lies close to that of the estimate of the effect	further research is very unlikely to change the confidence in the estimate of effect
Moderate (⨁⨁⨁⊖)	we are moderately confident in the effect estimate: the true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different.	further research is likely to have an important impact in the confidence in the estimate of effect
Low (⨁⨁⊖⊖)	our confidence in the effect estimate is limited: the true effect may be substantially different from the estimate of the effect.	further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate
Very low (⨁⊖⊖⊖)	we have very little confidence in the effect estimate: the true effect is likely to be substantially different from the estimate of effect.	any estimate of effect is very uncertain

B2.6 Interpretation of findings (evidence statements)

When interpreting results, we followed GRADE guidance for writing informative statements [50]. All interpretations are based on where the point estimate lies in relation to the pre-specified thresholds for an important effect (an important effect or not) and the direction of effect (beneficial or harmful). The certainty of evidence is communicated by qualifying the interpretation of effect (e.g. 'probably' improves for moderate certainty). For low certainty evidence the interpretation is qualified with the word 'may'. For example, 'Aromatherapy may improve sleep quality' indicates that the point estimate lies above the threshold for important benefit (an SMD >0.2) and that the evidence is of low certainty.

For very low certainty evidence, we do not provide an interpretation of the result except to state 'The evidence is very uncertain about the effect of aromatherapy on outcome'. This is one of two options that GRADE provides for interpreting findings based on very low certainty of evidence: "one option gives the direction of the effect, the other does not" [50]. The decision not to interpret very low certainty results was made independently by the NTWC to ensure a consistent and clear interpretation of findings across Natural Therapy Review reports (see Appendix G).

References for Appendix A and B

1. Brennan SE, McDonald S, Murano M, McKenzie JE: **Effectiveness of aromatherapy for prevention or treatment of disease, medical or preclinical conditions, and injury: protocol for a systematic review and meta-analysis.** *Systematic Reviews* 2022, **11:**148.
2. National Health and Medical Research Council: **Statement of requirement: Evidence evaluations for review of natural therapies (tranche two).** 2020.
3. Dodd S, Clarke M, Becker L, Mavergames C, Fish R, Williamson PR: **A taxonomy has been developed for outcomes in medical research to help improve knowledge discovery.** *Journal of Clinical Epidemiology* 2018, **96:**84-92.
4. World Health Organisation: **International statistical classification of diseases and related health problems (11th ed.) Available from: https://www.who.int/classifications/classification-of-diseases Access date: 15 January 2021** 2019.
5. Higgins JPT, Savovic J, Page MJ, Sterne JAC: **Revised Cochrane risk-of-bias tool for randomized trials (RoB 2). Available from https://www.riskofbias.info/welcome/rob-2-0-tool/current-version-of-rob-2. Access date: 8 February 2021.** 2019.
6. Steel A: **PRACI Study: unpublished data.** 2021.
7. Steel A, Leach M, Wardle J, Sibbritt D, Schloss J, Diezel H, Adams J: **The Australian Complementary Medicine Workforce: A Profile of 1,306 Practitioners from the PRACI Study.** *J Altern Complement Med* 2018, **24:**385-394.
8. McKenzie JE, Brennan SE, Ryan RE, Thomson HJ, Johnson RV, Thomas J: **Chapter 3: Defining the criteria for including studies and how they will be grouped for the synthesis.** In *Cochrane Handbook for Systematic Reviews of Interventions.* 2nd edition. Edited by Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Welch V. Chichester (UK): John Wiley & Sons; 2019
9. Canadian Federation of Aromatherapists (CFA): **About Canadian Federation of Aromatherapists (CFA). Available from: http://www.cfacanada.com/ Access date: 4 February 2021.** 2021.
10. International Aromatherapy and Aromatic Medicine Association: **About the International Aromatherapy and Aromatic Medicine Association (IAAMA). Available from https://www.iaama.org.au Access date: 4 February 2021.** 2021.
11. International Federation of Professional Aromatherapists: **International Federation of Professional Aromatherapists (IFPA). Available from: http://www.ifparoma.org/ Access date: 4 February 2021.** 2021.
12. National Association for Holistic Aromatherapy: **About National Association for Holistic Aromatherapy (NAHA). Available from: https://naha.org Access date: 4 February 2021.** 2021.
13. **How CENTRAL is created. Available from: https://www.cochranelibrary.com/central/central-creation Access date: 7 February 2021**
14. Ball EL, Owen-Booth B, Gray A, Shenkin SD, Hewitt J, McCleery J: **Aromatherapy for dementia.** *Cochrane Database of Systematic Reviews* 2020.
15. Candy B, Armstrong M, Flemming K, Kupeli N, Stone P, Vickerstaff V, Wilkinson S: **The effectiveness of aromatherapy, massage and reflexology in people with palliative care needs: A systematic review.** *Palliat Med* 2020, **34:**179-194.
16. Hines S, Steels E, Chang A, Gibbons K: **Aromatherapy for treatment of postoperative nausea and vomiting.** *Cochrane Database of Systematic Reviews* 2018.
17. Orchard A, van Vuuren S: **Commercial Essential Oils as Potential Antimicrobials to Treat Skin Diseases.** *Evid Based Complement Alternat Med* 2017, **2017:**4517971.
18. PDQ Integrative Alternative Complementary Therapies Editorial Board: **Aromatherapy With Essential Oils (PDQ®): Health Professional Version.** In *PDQ Cancer Information Summaries*. Bethesda (MD): National Cancer Institute (US); 2019.
19. Peterfalvi A, Miko E, Nagy T, Reger B, Simon D, Miseta A, Czéh B, Szereday L: **Much More Than a Pleasant Scent: A Review on Essential Oils Supporting the Immune System.** *Molecules* 2019, **24:**4530.
20. Smith CA, Collins CT, Crowther CA: **Aromatherapy for pain management in labour.** *Cochrane Database of Systematic Reviews* 2011.
21. Stea S, Beraudi A, De Pasquale D: **Essential oils for complementary treatment of surgical patients: state of the art.** *Evid Based Complement Alternat Med* 2014, **2014:**726341.
22. Horsley T, Dingwall O, Sampson M: **Checking reference lists to find additional studies for systematic reviews.** *Cochrane Database Syst Rev* 2011**:**MR000026.
23. Briscoe S, Bethel A, Rogers M: **Conduct and reporting of citation searching in Cochrane systematic reviews: A cross-sectional study.** *Res Synth Methods* 2020, **11:**169-180.
24. Wright K, Golder S, Rodriguez-Lopez R: **Citation searching: a systematic review case study of multiple risk behaviour interventions.** *BMC Med Res Methodol* 2014, **14:**73.

25. Cooper C, Booth A, Britten N, Garside R: **A comparison of results of empirical studies of supplementary search techniques and recommendations in review methodology handbooks: a methodological review.** *Syst Rev* 2017, **6:**234.
26. National Health and Medical Research Council: **Draft framework for protocols for systematic reviews of randomised controlled trials and non-randomised studies of interventions.** 2020.
27. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, McLeod L, Delacqua G, Delacqua F, Kirby J, Duda SN: **The REDCap consortium: Building an international community of software platform partners.** *J Biomed Inform* 2019, **95:**103208.
28. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG: **Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support.** *J Biomed Inform* 2009, **42:**377-381.
29. Hoffmann T, Glasziou P, Barbour V, Macdonald H: **Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide.** *BMJ* 2014, **1687:**1 - 13.
30. Sterne JAC, Savovic J, Page MJ, Elbers RG, Blencowe NS, Boutron I, Cates CJ, Cheng HY, Corbett MS, Eldridge SM, et al: **RoB 2: a revised tool for assessing risk of bias in randomised trials.** *BMJ* 2019, **366:**l4898.
31. Higgins JPT, Eldridge SM, Li T: **Chapter 23: Including variants on randomized trials. Cochrane Handbook for Systematic Reviews of Interventions version 6.1 (updated September 2020). Cochrane. Available from www.training.cochrane.org/handbook.** (Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page M, Welch V eds.); 2020.
32. Chinn S: **A simple method for converting an odds ratio to effect size for use in meta-analysis.** *Stat Med* 2000, **19:**3127-3131.
33. Schünemann HJ, Vist GE, Higgins J, Santesso N, Deeks JJ, Glasziou P, Akl EA, Guyatt GH, On behalf of the Cochrane GRADEing Methods Group: **Chapter 15: Interpreting results and drawing conclusions.** In *Cochrane Handbook for Systematic Reviews of Interventions.* 2nd edition. Edited by Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Welch V. Chichester (UK): John Wiley & Sons; 2019
34. Brennan SE, Johnston RV: **Research Note: Interpreting findings of a systematic review using GRADE methods.** *Journal of Physiotherapy* 2023.
35. Zeng L, Brignardello-Petersen R, Hultcrantz M, Siemieniuk RAC, Santesso N, Traversy G, Izcovich A, Sadeghirad B, Alexander PE, Devji T, et al: **GRADE guidelines 32: GRADE offers guidance on choosing targets of GRADE certainty of evidence ratings.** *J Clin Epidemiol* 2021, **137:**163-175.
36. Bell ML, McKenzie JE: **Designing psycho-oncology randomised trials and cluster randomised trials: variance components and intra-cluster correlation of commonly used psychosocial measures.** *Psychooncology* 2013, **22:**1738-1747.
37. Balk EM, Earley A, Patel K, Trikalinos TA, Dahabreh IJ: **Empirical Assessment of Within-Arm Correlation Imputation in Trials of Continuous Outcomes. Methods Research Report. (Prepared by the Tufts Evidence-based Practice Center under Contract No. 290-2007-10055-I.) AHRQ Publication No. 12(13)-EHC141-EF. Rockville, MD: Agency for Healthcare Research and Quality. November 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm.** In; 2012: *AHRQ Methods for Effective Health Care*].
38. Wan X, Wang W, Liu J, Tong T: **Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range.** *BMC Med Res Methodol* 2014, **14:**135.
39. Higgins JP, Thompson SG: **Quantifying heterogeneity in a meta-analysis.** *Stat Med* 2002, **21:**1539-1558.
40. Page MJ, Higgins JPT, Sterne JA: **Chapter 13: Assessing risk of bias due to missing results in a synthesis. Cochrane Handbook for Systematic Reviews of Interventions version 6.1 (updated September 2020). Cochrane. Available from www.training.cochrane.org/handbook.** (Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page M, Welch V eds.); 2020.
41. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L: **Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry.** *J Clin Epidemiol* 2008, **61:**991-996.
42. StataCorp (2021). Stata Statistical Software: Release 17. StataCorp.
43. Schunemann HJ, Brozek J, Guyatt G, Oxman AD (Eds.): **Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach**. Hamilton, Canada: McMaster University; 2013.
44. Schünemann HJ, Higgins J, Vist GE, Glasziou P, Akl EA, Skoetz N, Guyatt GH, On behalf of the Cochrane GRADEing Methods Group (formerly Applicability and Recommendations Methods Group) and the Cochrane Statistical Methods Group: **Chapter 14: Completing 'Summary of findings' tables and grading the certainty of the evidence.** In *Cochrane Handbook for Systematic Reviews of Interventions.* 2nd edition. Edited by Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Welch V. Chichester (UK): John Wiley & Sons; 2019
45. Guyatt G, Zhao Y, Mayer M, Briel M, Mustafa R, Izcovich A, Hultcrantz M, Iorio A, Alba AC, Foroutan F, et al: **GRADE Guidance 36: Updates to GRADE's approach to addressing inconsistency.** *J Clin Epidemiol* 2023.

46. Zeng L, Brignardello-Petersen R, Hultcrantz M, Mustafa RA, Murad MH, Iorio A, Traversy G, Akl EA, Mayer M, Schunemann HJ, Guyatt GH: **GRADE Guidance 34: update on rating imprecision using a minimally contextualized approach.** *J Clin Epidemiol* 2022, **150:**216-224.
47. Nikolakopoulou A, Higgins JPT, Papakonstantinou T, Chaimani A, Del Giovane C, Egger M, Salanti G: **CINeMA: An approach for assessing confidence in the results of a network meta-analysis.** *PLoS Med* 2020, **17:**e1003082.
48. Hultcrantz M, Rind D, Akl EA, Treweek S, Mustafa RA, Iorio A, Alper BS, Meerpohl JJ, Murad MH, Ansari MT, et al: **The GRADE Working Group clarifies the construct of certainty of evidence.** *Journal of Clinical Epidemiology* 2017.
49. Santesso N, Carrasco-Labra A, Langendam M, Brignardello-Petersen R, Mustafa RA, Heus P, Lasserson T, Opiyo N, Kunnamo I, Sinclair D, et al: **Improving GRADE Evidence Tables part 3: Guidance for useful GRADE certainty in the evidence judgments through explanatory footnotes.** *J Clin Epidemiol* 2016.
50. Santesso N, Glenton C, Dahm P, Garner P, Akl EA, Alper B, Brignardello-Petersen R, Carrasco-Labra A, De Beer H, Hultcrantz M, et al: **GRADE guidelines: 26. informative statements to communicate the findings of systematic reviews of interventions.** *Journal of Clinical Epidemiology* 2019.

Appendix C. Lists of studies considered for review

Overview of Appendix C – separate file

Appendix C is comprised of four parts (below).

These Appendices report the studies excluded at full text review with reason for exclusion, the public submissions and eligibility decision for each, the studies awaiting classification (including studies in languages other than English), and the studies included on the evidence inventory.

Appendix C1. Citation details of studies from search results excluded
Appendix C2. Citation details of studies from public submissions
Appendix C3. Citation details of studies for studies on the evidence inventory
Appendix C4. Citation details of studies awaiting classification

Appendix D. Extended results and citations for studies included in the evidence synthesis

Overview of Appendix D – separate file

Appendix D is comprised of a single file in which we report results for additional subgroup analyses, sensitivity analyses and analyses to inform the assessment of biases due to missing results from each synthesis.

The Appendix begins with a brief statement about where information about the characteristics of included studies and risk of bias assessments are located. It is then ordered by outcome as per the results section in the main report.

Appendix D also contains the reference list for studies included in the evidence synthesis.

Sections are as follows

D1 Pain
D2 Nausea and vomiting
D3 Sleep
D4 Fatigue
D5 Emotional functioning and mental health
D6 Health-related quality of life
D7 Physical function
D8 Citation details of studies included in the evidence synthesis

Appendix E. Characteristics of studies included in the review

Overview of Appendix E – separate files

Appendix E is comprised of four parts, each in a separate file.

Appendix E1 provides information about the characteristics of each of the studies eligible for the evidence synthesis.

- study ID, location, setting, and study design
- the population eligibility criteria, number of participants randomised, participant characteristics, and ICD-11 codes
- the aromatherapy treatment goal, and details about the aromatherapy intervention(s) and comparator(s)
- a list of all reported outcome(s) categorised according to whether they were eligible or ineligible for the synthesis, the measurement method for each eligible outcome, the timing of outcome measurement, and the outcome(s) selected for inclusion in the synthesis for each outcome domain

Appendix E2 provides information about funding, declaration of interest and ethics approval for each study.

Studies were included in E1 and E2 irrespective of whether they provided data that could be included in the meta-analysis.

Appendix E3 provides details of the characteristics of each of the studies included in the evidence inventory. These studies were eligible for the review, but were excluded from the synthesis following agreement from the NHMRC and NTWC, with input from NTREAP. The reasons why each study was excluded from the synthesis is reported in this file.

Appendix E4 provides a list of studies that were eligible for the evidence synthesis, but for which data could not be included in the meta-analysis. Details of the syntheses for which each of the studies was eligible are tabulated, together with the number of participants and the reason why data are missing.

Appendices are as follows

E1. Characteristics of studies included in the evidence synthesis
E2. Funding sources, potential conflicts of interest and ethics approval for studies included in the evidence synthesis
E3. Characteristics of studies included in the evidence inventory (ineligible for the evidence synthesis)
E4. List of studies eligible for the evidence synthesis with data that could not be included for meta-analysis

Appendix F. Risk of bias assessments

Overview of Appendix F – separate file

Appendix F is a single file containing the full risk of bias assessment for each study that contributed data for meta-analysis.

The Appendix

- begins with information to orient the reader to the content,
- provides the signalling questions for the risk of bias tools, and
- includes additional methods information about how trials with clustering were handled.

Appendix G. Differences between the protocol and the review

Changes from the protocol and methods not implemented

	Section	Planned method	Change	Details (text, rationale or both)
1	A1.1.3	We planned to examine the effects of aromatherapy compared to "gold standard" treatments, in the exceptional circumstance that there were studies at low risk of bias that could be combined in a synthesis.	Active comparators were not included in synthesis	There were no studies at low risk of bias, therefore the criterion as specified in the protocol was not met. Studies with active comparators are reported on the evidence inventory.
2	A2.2 Searching other resources	Where these groups [making submissions] recommend particular systematic reviews, we will examine references for included studies to identify potentially eligible randomised trials.	We did not screen the reference lists of reviews	**Text deleted.** **Rationale**. Our search was comprehensive, limited to randomised trials (which are unlikely to be missed using the search methods employed for this review), and the findings of the review are unlikely to change with the addition of additional trials.
3	A3.1 Selection of studies	We propose to split title and abstract screening into two phases. Phase 1 records (indexed with the thesaurus terms Aromatherapy or Oils volatile or with aromatherapy in the title) will be screened independently by at least two reviewers. Phase 2 (remaining records) will be screened by one reviewer, with a 10% random sample screened by a second reviewer (with further sampling if needed until 80% agreement is achieved).	Two people screened independently	**Revised text**: "All records were reviewed independently by two reviewers at both the title and abstract screening and full-text review stages in Covidence."
4	A3.1 Selection of studies	Studies confirmed as meeting the eligibility criteria, but for which results are not available in a published report, will be included in a list of 'ongoing studies'.	Reported on evidence inventory	Protocols for studies confirmed as meeting the eligibility criteria were reported on the evidence inventory.
5	A3.1 Selection of studies	The following will be included in a list of 'studies awaiting classification': Studies that are only published as abstracts or for which a full report is not available (i.e. we will not seek further information from study authors to confirm eligibility). Studies confirmed as likely to be eligible, but for which no	Additional code for studies awaiting classification	**Text added**. "Studies for which a full report was available but the report was incomplete or ambiguous such that eligibility based on one or more PICO criteria or study design could not be confirmed (i.e. we did not seek further information from study authors to confirm eligibility)"

	Section	Planned method	Change	Details (text, rationale or both)
		English language translation of the full-text publication is available. Studies for which eligibility cannot be confirmed following translation of the title and abstract using Google translate.		
6	A3.1 Selection of studies	Studies that did not meet the eligibility criteria were excluded and the reason for exclusion was recorded at full-text screening. These studies are included in a 'Characteristics of excluded studies' table in which the reason for exclusion is reported.	Additional check of eligibility at data extraction	**Revised text**. "Studies that did not meet the eligibility criteria were excluded and the reason for exclusion was recorded at full-text screening. *Inclusion decisions were checked at data extraction, and for any studies identified as ineligible at this stage, the decision and exclusion reason were recorded in Covidence.* These studies are included in a 'Characteristics of excluded studies' table in which the reason for exclusion is reported."
7	A3.1 Selection of studies	For studies that originated from the call for evidence, NTREAP or NTWC, we will record and report exclusion decisions irrespective of whether the study was excluded during title and abstract screening or full text review. We will document the flow of these studies through the review in the PRISMA flow chart and annotate tables with the source.	We did not annotate tables with source because no additional studies were identified	**Revised text**. "For studies that originated from the call for evidence, NTREAP, or the Committee, we recorded and reported exclusion decisions irrespective of whether the study was excluded during title and abstract screening or full text review. We documented the flow of these studies through the review in the PRISMA flow chart and in Appendix C2."
8	A3.1 Selection of studies	Records were to be matched using trial registry numbers. Where these were not available, we considered author names, trial name, trial location(s) and number of participants.	Additional information was required to match multiple trial records	**Revised text**. "Records were matched using trial registry numbers. Where these were not available, we considered author names, trial name, trial location(s), number of participants, *baseline characteristics and PICO*."
9	A3.1 Selection of studies	We planned to screen all registry records to identify ongoing studies and to conduct an analysis of missing results.	Partial screening only	**Revised text**. The trial register records retrieved from CENTRAL (i.e. from ClinicalTrials.gov and WHO ICTRP) were used to identify matching records for included studies, and a subset of the unmatched records (500) was screened to ascertain the likely number of potentially eligible trials for which there was no full text report. This was done to ascertain whether it was feasible to screen registry records to determine the number of ongoing studies and analyse records for missing results. (Some trial register records for which the source in CENTRAL was not given as ClinicalTrials.gov or WHO ICTRP were included in the records screened.) A decision in consultation with NHMRC was made not to screen the trial register records or the full registry entry given the volume of studies eligible for the

	Section	Planned method	Change	Details (text, rationale or both)
				review and that an analysis of registry records would contribute little additional information.
10	A3.1 Selection of studies	Standard one-step process for screening full text studies was planned.	An additional check of study eligibility was added at data extraction	**Revised text.** Inclusion decisions were checked at data extraction, and for any studies identified as ineligible at this stage, the decision and exclusion reason were recorded in Covidence. **Rationale**. Some studies required more detailed review (often across multiple reports) to confirm eligibility.
11	B1.2 Measure of treatment effect	Where a valid and reliable minimal important difference (MID) is available for a familiar measure of relevance to the population groups in the meta-analysis, we will re-express the SMD in units of the measure and interpret the effect in relation to the MID if feasible to do so.	We did not re-express SMDs in units of a familiar measure	**Rationale**. Due to the diversity of populations, population-specific outcomes, and outcome measures, it was not feasible to re-express the SMDs using a familiar measure.
12	B1.2 Measure of treatment effect	For dichotomous outcomes, we will seek advice from NTWC on interpreting the size of the effect (seeking agreement on a threshold for a small but important difference).	Method not required.	**Rationale.** We did not report dichotomous analyses for any of our meta-analyses. For most of our analyses, all included studies reported continuous outcomes. For some analyses (e.g. nausea and vomiting) the majority of outcomes were continuous but some were dichotomous having been dichotomised from a continuous measure. We re-expressed these results using the SMD, hence there was no need to interpret dichotomous outcomes.
13	B1.2 Measure of treatment effects	We planned to use Cohen's guiding rules for SMDs where 0.2 represents a small effect, 0.5 a moderate effect, and 0.8 a large effect.	We used a single threshold for an important effect (0.2) and did not interpret effect size.	**Revised text (and rationale).** Because we were concerned that bias may be leading to exaggerated effect sizes, we chose not to describe the size of effect (i.e. we did not interpret effects as small, moderate or large) because this is likely to be misleading. **Implications.** This has no implications for the certainty of evidence because our a priori plan was to assess certainty in relation to whether there was an important effect or not (i.e. in relation to a threshold for an important difference of an SMD of 0.2), not our certainty in the magnitude of effect (trivial, small, moderate or large).
14	B1.6 Assessment of bias due to missing results	We planned to undertake a full assessment of 'known-unknowns' to determine whether results are missing from each meta-analysis, by examining the methods section, trial registry entry (if available), and trial protocol (if available) for trials meeting the inclusion criteria for the meta-analysis.	Assessment not done (with some exceptions)	**Revised text (and rationale)**. "In practice, the assessment of 'known-unknowns' was not feasible due to the large number of included studies and additional studies in trial registers. This assessment was therefore limited to examining the potential impact of studies for which data could not be included in the meta-analysis and those in languages other than English that were judged as being likely to meet the eligibility criteria for each synthesis." **Implications**. Likely to be minimal. We downgraded the certainty of evidence for publication bias for most analyses (overall, some subgroups) based on evidence from contour

	Section	Planned method	Change	Details (text, rationale or both)
				enhanced funnel plots and sensitivity analyses, so additional downgrades would not apply.
15	B2.2 Summary and synthesis when meta-analysis is not possible	For a particular comparison, if we are unable to analyse most of the effect estimates (due to incomplete reporting of effects and their variances, variability in the effect measures across the studies), we will consider alternative synthesis methods …	Other synthesis methods not used	**Rationale.** We were able to analyse most of the effect estimates. Concerns about the integrity of data led to a decision not to report results from studies that could not be included. Text in this section has been revised accordingly.
16	B2.4 Sensitivity analysis	Analysis to examine the impact of risk of bias.	Could not be done	**Revised text.** All studies were rated at high risk of bias or some concerns (i.e. contributed 100% of the weight in all meta-analyses). As such, it was not possible to perform sensitivity analyses as per the protocol to examine whether removing studies at high risk of bias or some concerns changed the direction or size of effect estimate importantly (a reduction in benefit or an increase in harm being most concerning)
17	B2.4 Sensitivity analysis	Our stated method was to undertake and report sensitivity analyses in which we excluded "trials judged to be at an overall high or unclear risk of bias."	Terminology updated (not a change to protocol)	**"**Unclear risk of bias" is the terminology used in the original Cochrane ROB tool and protocol. Updated ROB2 terminology replaces this wording with "some concerns".
18	B2.4 Sensitivity analysis	Sensitivity analysis	Additional analysis	We conducted an additional sensitivity analysis, not mentioned in our protocol, to examine if the meta-analysis estimates were robust to "the assumptions made to enable inclusion of results in the meta-analysis, specifically (1) transforming or imputing statistics, and (2) including change scores (change from baseline) when post-intervention (final) values (and their standard deviations) were unavailable."
19	B2.5 GRADE assessment - imprecision	For large effects, we planned to consider whether the sample size meets the optimal information size (based on number of events for binary outcomes; sample size for continuous outcomes).	We did not do this.	We did not consider sample size in judging imprecision for large effects, partly because we interpreted effects only in relation to a threshold for a small important effect (thus, whether the effect was slight or large was not factored in) and partly because concerns about large effects in small studies are driven by concerns about publication bias, which we considered. We were concerned that results from single studies, or where the aggregate sample size was small, may not be similar to what would be observed in the population of interest more generally. For this reason, we considered the potential for indirectness in these circumstances.
20	B2.5 GRADE assessments – risk of bias	As per B2.4 we did not use the term 'some concerns' in the protocol when describing our approach to rating down for risk of bias	Terminology updated (not a change to protocol)	The use of 'some concerns' is consistent with the use of the ROB2 tool. Our approach to GRADE is consistent with that for sensitivity analyses where downgrades of -1 are considered where the majority of studies are rated as

	Section	Planned method	Change	Details (text, rationale or both)
				'some concerns' or studies with the majority of weight in the analysis are rated as 'high risk of bias'. Downgrades of -2 are made where most or all studies are at high risk of bias. Decisions not to rate down in these circumstances would be warranted if sensitivity analyses showed removal of studies at risk of bias did not materially alter the effect estimate.
21	B2.6 Interpretation of findings	Our endorsed protocol stated that we would report "a plain language statement interpreting the evidence for each comparison and outcome, following GRADE guidance for writing informative statements". We did not specify which option would be used for very low certainty evidence (i.e. give the direction of the effect, or limit to a statement that the 'evidence is very uncertain').	After review of our draft report, NTWC advised not to include direction of effect for very low certainty evidence.	Rationale. The Natural Therapies Review reports endorsed prior to submission of our aromatherapy review report did not include direction of effect in evidence statements for very low certainty evidence. To ensure a consistent and clear approach across reviews, the NTWC requested that the same approach be used in the aromatherapy review report.

Appendix H. Response to comments from the Methodological review

Methodological review (or peer review) was conducted to appraise the methodological quality and assess the appropriateness of reporting for this systematic review (including appendices).

For reporting, the methodological review assessed the systematic review against the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) Checklist (2020) and where applicable, the MECIR (Methodological Expectations of Cochrane Intervention Reviews) manual to ensure the systematic review was designed and conducted in accordance with:

- NHMRC's Developing your Guideline module in NHMRC's Guidelines for Guidelines Handbook
- Cochrane Handbook for Systematic Reviews of Interventions (updated 2022)
- GRADE guidance and GRADE working group criteria for determining whether the GRADE approach was used (GRADE handbook).

Assessment included the application of criteria for considering studies for the review and synthesis, search methods, data extraction and analysis, assessment of risk of bias of studies, assessment of the certainty of evidence using GRADE, and the interpretation and summary of findings.

The systematic review (including appendices) has been updated to reflect the amendments suggested by methodological review and NHMRC's Natural Therapies Working Committee, where appropriate. In summary, updates included additional information and/ or clarification of the Plain Language Summary, Executive Summary, Results sections and Appendices, including:

- Additional information about the overall quality of the evidence base added to the summaries and discussion.
- GRADE judgements clarified and confirmed where appropriate.
- Clarifications to the PRISMA diagram.
- Additional information provided on the types of active comparators considered.
- Rewording in various parts of the report to improve clarity.

A detailed record of responses to all comments indicating changes that were made was provided to NHMRC together with the amended Report and Appendices documents.

Appendix I. Abbreviations

Below is a list of abbreviations used in the report. Abbreviations for outcome measures are in a table following the list.

AMED: Allied and Complementary Medicine Database

CENTRAL: Cochrane Central Register of Controlled Trials

CINAHL: Cumulative Index of Nursing and Allied Health Literature

CM: Complementary Medicine

COMET: Core Outcome Measures in Effectiveness Trials

DEFF: design effect

GRADE: Grading of Recommendations, Assessment, Development and Evaluation

IAAMA: International Aromatherapy and Aromatic Medicine Association

ICC: intra-cluster correlation

ICD-11: International Classification of Diseases 11th Revision

ICTRP: International Clinical Trials Registry Platform

MA: Meta-analysis

MeSH: Medical Subject Headings

MID: minimal important difference

NHMRC: National Health and Medical Research Council

NRSI: non-randomised study of interventions

NTREAP: Natural Therapies Review Expert Advisory Panel

NTWC: Natural Therapies Working Committee

PICO: population, intervention, comparator, outcome

PRACI: Practitioner Research and Collaboration Initiative

PRISMA: Preferred Reporting Items for Systematic review and Meta-Analyses

PRISMA-P: Preferred Reporting Items for Systematic review and Meta-Analyses Protocols

RCT: randomised controlled trial

REML: restricted maximum likelihood estimator

RR: risk ratios

SMD: standardised mean difference

TIDieR: Template for Intervention Description and Replication

TGA: Therapeutic Goods Administration

Abbreviations for measures reported in this review

Abbreviation	Measure
	Barthel Index
	Blau QOL scale
	Rhodes Index of Nausea
	Oucher Scales
	Children's Fear Scale
	Distress Thermometer
ADRQL	Alzheimer Disease-related Quality of Life
BAI	Beck Anxiety Inventory
BCTQ	Boston Carpal Tunnel Syndrome Questionnaire
BDI	Beck Depression Inventory
BFI	Brief Fatigue Inventory
BPI-DPN	Brief Pain Inventory for Diabetic Painful Neuropathy
BSPAS	Burn Specific Pain Anxiety Scale
CFS	Chalder Fatigue Scale
CMAI	Cohen Mansfield Agitation Inventory
COMFORT-B	COMFORT behaviour scale
CSDD	Cornell Scale for Depression in Dementia
DAN	Douleur Aiguë du Nouveau-né
DASS-21	Depression Anxiety and Stress Scale 21
DCM	Dementia Care Mapping
DN4	Douleur Neuropathic 4 Questions
EORTC QLQ-C30	European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire
EPDS	Edinburgh Postnatal Depression Scale
EQ‐5D	European Quality of Life with 5 Dimensions
ESASr	Edmonton Symptom Assessment System revised
FACES	Wong-Baker FACES Pain Rating Scale
FIS	Face image scale
FLACC	Face Activity Consolability Scale
FRS	Faces Rating Scale
FSS	Fatigue Severity Scale
GDS-SF	Geriatric Depression Scale Short Form
GHQ-28	General Health Questionnaire [28 items]
GMPI	Geriatric Multidimensional Pain Illness Inventory
HADS	Hospital Anxiety and Depression Scale
HAM-A	Hamilton Anxiety Rating Scale
INVR	Index of Nausea and Retching
ISK	Lequesne Algofunctional Index of Severity for Knee Disease
KDQOL-SF	Kidney Disease Quality of Life - Short Form
KOOS	Knee injury and Osteoarthritis Outcome Score
MDAS	Modified Dental Anxiety Scale
MFI	Multidimensional Fatigue Inventory
MFSI	Multidimensional Fatigue Symptom Inventory
MRS	Menopause Rating Scale
MSIS-29	Multiple Sclerosis Impact Scale
MYMOP2	Measure Yourself Medical Outcome Profile 2
NDI	Neck Disability Index
NePIQoL	Neuropathic Pain Impact on Quality of Life
NIH-CPSI	NIH-Chronic Prostatitis Symptom Index
NIPS	Neonatal Infant Pain Scale
none	McGill Pain Questionnaire

Abbreviation	Measure
NPI	Neuropsychiatric Inventory
NPI-NH	Neuropsychiatric Inventory Nursing Home
NPRS	Numeric Pain Rating Scale
NPSI	Neuropathic Pain Symptom Inventory
NRS	Numerical Rating Scale
OAKQOL	OsteoArthritis of Knee Hip Quality Of Life
PANAS	Positive and Negative Affect Schedule
PAS	Pittsburgh Agitation Scale
PeNAT	Pediatric Nausea Assessment Tool
PFS	Piper Fatigue Scale
PGCARS	Philadelphia Geriatric Centre Affect Rating Scale
PIPP-R	Premature Infant Pain Profile - Revised
PIRS-20	Pittsburgh Insomnia Rating Scale - 20
POMS	Profile of Mood States
PONV Intensity Scale	Postoperative Nausea and Vomiting Intensity Scale
PSQI	Pittsburgh Sleep Quality Index
PSS	Premenstrual Syndrome Scale
PUQE-24	24-hour Pregnancy-Unique Quantification of Emesis
RCSQ	Richards-Campbell Sleep Questionnaire
RFS	Rhoten Fatigue Scale
RMDQ	Roland-Morris Disability Questionnaires
RQLQ	Rhinoconjunctivitis Quality of Life Questionnaire
RSCL	Rotterdam Symptom Checklist
SCID-II	Structured Clinical Interview for DSM Disorders for DSM-IV
SF-36	Short Form Health Survey
SMHSQ	St Mary's Hospital Sleep Quality Questionnaire
SOT	Short Observational Tool [dementia]
SPHERE	Somatic and Psychological Health Report
STAI	State-Trait Anxiety Inventory
STAI + pain indices	State-Trait Anxiety Inventory + pain indices
STAI-6	State-Trait Anxiety Inventory - Short Form
STAI-CH	State-Trait Anxiety Inventory for Children
TPPPS	Toddler Preschooler Postoperative Pain Scale
VAS	Visual Analogue Scale
VAS-A	Visual Analogue Scale for Anxiety
VDS	Visual Descriptive Scale
VRS [pain]	Verbal Rating Scale [pain]
VRS [nausea]	Verbal Rating Scale [nausea]
VSH	Verran and Snyder-Halpern Sleep Scale
WHOQOL-BREF	WHO Quality of Life-BREF
WOMAC	Western Ontario and McMaster Universities Osteoarthritis Index