Methods

To ensure completeness and feasibility of the evidence review, we used an approach in which we prioritized the types of studies according to the design that was more likely to provide the best available evidence. First, we searched for systematic reviews of the literature. Second, we appraised all existing systematic reviews to select the most trustworthy (highest methodological quality, most up-to-date, most applicable) from which to draw conclusions. Third, we used the information presented in the systematic reviews to abstract information regarding the effects of the interventions of interest. Fourth, we assessed the certainty of the evidence (also known as quality of the evidence) abstracted from the selected systematic reviews. We planned to search for primary studies if systematic reviews were not found.

Information sources: We searched for existing systematic reviews in:

1. Epistemonikos (https://www.epistemonikos.org), an electronic database that focuses on systematic reviews. We used a comprehensive search strategy based on the population, using the terms "gender dysphoria", "gender identity disorder" and "transgender". We conducted this search on April 23, 2022.
2. OVID Medline. We used a search strategy based on the population and the interventions of interest, as well as an adaptation of a filter for systematic reviews from the Health Information Research Unit at McMaster University. We conducted this search on April 23, 2022.
3. Grey literature: we conducted a manual search in the websites of specific health agencies: National Institutes for Health and Care Excellence (NICE), Agency for Healthcare Research and Quality (AHRQ), Canada's Drug and Health Technology Agency (CADTH), and the website from the Society for Evidence-Based Gender Medicine (SEGM). We conducted these searches between April 27-30, 2022.

We used no date limits for the searches, but we did limit to systematic reviews published in English. Search strategies are available in Appendix 1.

Eligibility criteria: We included systematic reviews, which we defined as:

1. Reviews in which the authors searched for studies to include in at least one electronic database, and in which there were eligibility criteria for including studies and a methodology for assessing and synthesizing the evidence, or
2. Reviews in which the authors searched for studies to include in at least one electronic database, and although there was no description of eligibility criteria or methodology, the presentation of the results strongly suggested that the authors used systematic methods (e.g. flow chart depicting study selection, tables with the same information from all included studies, synthesis of data at the outcome level).

We screened systematic reviews using the following criteria for inclusion:

- ***Type of participants:*** Young individuals (< 25 years old) with a diagnosis of gender dysphoria/gender identity disorder. We included reviews in which authors used any label and diagnostic criteria for this condition. We included reviews in which the participants in the reported studies were older if it was the only evidence available for a specific question. We excluded reviews with mixed populations (i.e. with and without gender dysphoria) in which people without gender dysphoria constituted more than 20% of the total sample.

- ***Type of Interventions*:** Puberty blockers, cross-sex hormones, gender affirming surgeries. We included any type of puberty blockers and cross-sex hormones, provided with any regimen. We included the following surgeries: phalloplasty, vaginoplasty, and chest surgery (mastectomy or breast implants/augmentation)

- ***Type of comparison:*** When the systematic reviews included comparative studies, the comparator of interest was no intervention. Participants could have received psychotherapy or counselling as a cointervention (in both groups).

- ***Type of outcomes:*** Gender dysphoria, mental health outcomes (depression and anxiety), quality of life, suicidal ideation, suicide, adverse effects (for puberty blockers and cross-sex hormones only), and satisfaction, complications, reoperation, and regret (for surgeries only). We included any length of follow-up. We excluded surrogate outcomes such as blood pressure, bone mineral density, kidney or liver function test values, etc.

- ***Type of studies included in the systematic reviews*:** Any clinical study (studies in which the researchers recruited and measured outcomes in humans) regardless of study design. This included randomized clinical trials, comparative observational studies, and case series. Because we could not quantify effect measures, incidence, or prevalence, we excluded case reports.

We excluded systematic reviews published only in abstract format, and those that we could not retrieve in full text (no access through the McMaster University library, or open access online).

Selection process: The two reviewers screened all titles and abstracts independently and in duplicate, followed by screening of full texts of potentially eligible systematic reviews independently and in duplicate, using the systematic review online application Covidence (https://www.covidence.org). We solved disagreements by consensus.

To select the most useful systematic reviews among all of those that met the eligibility criteria, we used the following prioritization criteria:

1. Date of publication: we prioritized systematic reviews published within the last 3 years (2020-2022)
2. Match between eligibility criteria of the review and the question of interest: we prioritized reviews in which the authors specifically included the population, intervention, comparison, and outcomes of interest for this evidence review
3. Outcome data available: we prioritized systematic reviews in which the authors report outcome data
4. Methodological quality: we used a modified version of the items in AMSTAR 2.[1] We modified the items to ensure assessment of methodological rather than reporting quality (Table 1). We rated each systematic review as having high, moderate, low, or critically low methodological quality, according to the guidance from the developers of the tool.[1]

Pl. Trial Ex. 283B

We reached consensus on critical items that determined this rating (Table 1). We prioritized selection of systematic reviews with highest methodological quality.

For surgical interventions, in addition, we prioritized systematic reviews that covered all gender affirming surgeries (instead of focusing on a specific type of surgery).

We selected a systematic review specifically for each of the outcomes of interest. In other words, we chose the best systematic review to inform each outcome. Each systematic review, however, could inform more than one outcome.

Table 1: Items used to rate the methodological quality of the eligible systematic reviews

AMSTAR Item	Modification to measure methodological quality
1. Did the research questions and inclusion criteria for the review include the components of PICO?	Does the review have a clear question and are the eligibility criteria for studies consistent with the question?
2. Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report justify any significant deviations from the protocol?	No modification needed
3. Did the review authors explain their selection of the study designs for inclusion in the review?	No modification needed
4. Did the review authors use a comprehensive literature search strategy?	Did the authors search in at least 2 electronic databases, using a reproducible search strategy?
5. Did the review authors perform study selection in duplicate?	No modification needed
6. Did the review authors perform data extraction in duplicate?	No modification needed
7. Did the review authors provide a list of excluded studies and justify the exclusions?	No modification needed
8. Did the review authors describe the included studies in adequate detail?	No modification needed
9. Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?	No modification needed
10. Did the review authors report on the sources of funding for the studies included in the review?	Did the review authors consider conflicts of interest and how they may have affected the results of the primary studies?
11. If meta-analysis was performed, did the review authors use appropriate methods for statistical combination of results?	Was the synthesis of evidence done appropriately? (outcome level, appropriate meta analysis or narrative synthesis)
12. If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?	Did authors use subgroup or sensitivity analysis to assess the effect of risk of bias in meta-analytic results? Likely not applicable to most cases
13. Did the review authors account for RoB in primary studies when interpreting/discussing the results of the review?	Did the review authors incorporate an assessment of risk of bias at the outcome level when drawing conclusions?
14. Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?	Did the review authors incorporate an assessment of heterogeneity at the outcome level when drawing conclusions?

15. If they performed quantitative synthesis did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?	Did the authors address publication bias? (regardless of whether synthesis was using a meta-analysis or narrative)
16. Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?	Did the authors report conflicts of interest and did they manage any existing conflict of interest appropriately?

Shaded items were items considered critical.

Data abstraction: We abstracted outcome data from each of the systematic reviews. To ensure feasibility, we used the data as reported by the authors of the review and did not re-abstract data from the primary studies. One reviewer abstracted the data and a second reviewer checked the data for accuracy.

Data synthesis: Using the systematic reviews prioritized, we synthesized the evidence at the outcome level. Because of the higher likelihood of it resulting in higher certainty of evidence (details below) for each outcome, when there was comparative data (i.e. comparison of outcomes between an untreated and a treated group) and non-comparative data (i.e. changes from before to after treatment in one group, or only outcomes after treatment), we prioritized comparative data.

We prioritized numerical results (i.e. magnitudes of effect) and reported estimates and their 95% confidence intervals (CIs). When results were not reported in that way, we calculated the estimates and CIs when systematic review authors provided sufficient information. When necessary, we assumed moderate correlation coefficients for the changes between baseline and follow up (coefficient= 0.4). When this information was not available we reported narratively the effect estimates and ranges.

When a specific study reported the same outcome measured by more than one scale, we chose the scale presented first. We highlighted situations when the results obtained with other scales were importantly different.

When the same outcome was reported by more than one study but we could not pool the results, we created narrative syntheses.

Certainty of evidence: For each outcome, we assessed the certainty of the evidence (also known as quality of the evidence) using the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach.[2] The certainty of evidence can be rated as high, moderate, low, or very low (Table 2). For effects of interventions, the certainty of the evidence started as high and could be rated down due to serious concerns about risk of bias, inconsistency, indirectness, imprecision, and publication bias. For inferences about the effect of using a treatment versus no treatment, when there was no comparison group, we assessed risk of bias as very serious and rated down the certainty of the evidence 2 levels by default. We used the same principles when assessing the certainty of the evidence in estimates of prevalence or rates, but did not judge risk of bias as resulting in very serious concerns due to lack of a comparison group. For all assessments, we used the information presented by the authors of the systematic review (e.g. assessments of risk of bias of the included studies, effect estimates from studies).

Table 2: GRADE levels of certainty of the evidence

Certainty level	Definition
High ⊕⊕⊕⊕	We are very confident that the true result (effect estimate/ prevalence/ mean, etc.) lies close to that of the estimate of the result
Moderate ⊕⊕⊕◯	We are moderately confident in the result: the true result is likely to be close to the estimate of the result, bur there is a possibility that it is substantially different
Low ⊕⊕◯◯	Our confidence in the result is limited: the true result may be substantially different from the estimate of the result
Very low ⊕◯◯◯	We have very little confidence in the result: the true result is likely to be substantially different from the estimate of the result

Presentation of results: We created GRADE Summary of Findings tables in which we describe the evidence available for each of the outcomes, and the certainty of the evidence. These tables contain the following information:

- Outcomes: measurement method (including scales, if applicable) and follow-up
- Estimates of effect: absolute and relative estimates of effect, and their corresponding 95% CIs.
- Number of studies and participants providing evidence for the outcome
- GRADE certainty of the evidence, with a link to detailed explanations (provided at the bottom of the table) of why the certainty of the evidence was rated at a specific level
- A narrative statement about what happens with the outcome, based on the estimate of effect and certainty of evidence.

References

1. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *Bmj* 2017;358:j4008. doi: 10.1136/bmj.j4008 [published Online First: 2017/09/25]
2. Blashem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of the evidence. *Journal of clinical epidemiology* 2011;64:401-06.