Supplement 1: Utah Review: Clinical studies: ROBIS

ROBIS	Study Eligibility Criteria	Identification and Selection of Studies	Data Collection and Study Appraisal	Synthesis and Findings	**Over-all RoB**
Utah Review: Clinical studies	High	High	High	High	**High**

Summary: overall risk of bias is High

Identifying concerns with the review process

Domain 1: Concerns regarding specification of study eligibility criteria

1.1	**Did the review adhere to pre-defined objectives and eligibility criteria?** Response: No. No pre-specified protocol has been provided. The text indicates that 230 studies were initially identified as eligible, with all study designs and outcomes considered eligible. However, only 90 studies were later classified as "eligible" following a post hoc prioritization of "high-priority" study types and outcomes, which was explained by resource limitations rather than methodological justification.	N
1.2	**Were the eligibility criteria appropriate for the review question?** Response: No. The review questions addressed both benefits and harms, and the initial search strategy allowed for all outcomes and study types. However, eligibility was later operationally redefined post hoc to exclude studies of non–"high-priority" outcomes, such as fertility, desistance, and regret. These outcomes are central to assessing benefits and harms, and desistance and regret were also described as "pointedly of interest" to the commissioning body. In addition, excluding studies published before 2010 is inconsistent with the goal of understanding long-term outcomes of hormonal interventions.	N

1.3 Were eligibility criteria unambiguous? Response: Probably no. The original eligibility criteria were later replaced with a post hoc set of criteria, and neither set was clearly defined. This redefinition created uncertainty about what constituted an "eligible" study. The report also describes the evidence base inconsistently: it first refers to analyzing 134 studies encompassing 28,056 pediatric patients (p. 44), later states that 230 studies encompassing the same 28,056 patients were identified (p. 90), and elsewhere indicates that only about 40% of these studies (either 90 studies (p. 34) or 89 studies (p. 90) of 230) were ultimately considered "relevant" and analyzed.	PN
1.4 Were all restrictions in eligibility criteria based on study characteristics appropriate? Response: Probably no. Initially, no exclusions based on study type or outcome were reported, leading to the inclusion of all study types from surveys to case reports, and raising serious questions about what qualifies as relevant evidence and how these sources align with the stated research objectives. Once faced with an excessive number of studies to analyze, several restrictions were applied to limit the number of studies undergoing data extraction. These restrictions were feasibility-driven rather than based on study characteristics aligned with the review question and resulted in a number of arbitrary decisions that detracted from the ability to answer research questions.	PN
1.5 Were any restrictions in eligibility criteria based on sources of information appropriate? Response: No. The exclusion of studies published before 2010 prevented comprehensive identification of long-term outcomes. The researchers applied post hoc judgements regarding what constituted "relevant" studies, effectively redefining eligibility after the search had been conducted.	N

- **High concern:** A fundamental requirement of any systematic review is a clearly defined scope, including pre-specified eligibility criteria for study inclusion. This review does not meet that standard. Eligibility criteria were modified post hoc through outcome and comparison prioritization and feasibility-driven restrictions without a pre-specified protocol, creating a substantial likelihood that relevant studies addressing parts of the stated review question were not fully included in the analytic evidence base.

Domain 2: Concerns regarding identification and selection of studies

2.1 Did the search include an appropriate range of databases/electronic sources for published and unpublished reports? Response: Probably no. Searches were limited to MEDLINE, Embase, and ClinicalTrials.gov, without additional specialist databases (e.g., PsycINFO, CENTRAL, CINAHL). Embase conference records were explicitly excluded.	PN
2.2 Were methods additional to database searching used to identify relevant reports? Response: Yes. Additional records were identified through reference list checks, trial registries, and expert recommendations.	Y
2.3 Were the terms and structure of the search strategy likely to retrieve as many eligible studies as possible? Response: Yes. The search strategy used comprehensive and appropriate terms.	Y
2.4 Were restrictions based on date, publication format, or language appropriate? Response: No. The searches were conducted between March and June 2023, with ClinicalTrials.gov searched in September 2023, and the report was published in May 2025, meaning the search was nearly two years out of date at publication. In addition, the Embase strategies excluded conference abstracts and reviews, a publication-format restriction that ROBIS guidance notes is rarely appropriate.	N
2.5 Were efforts made to minimize errors in selection of studies? Response: Yes. Title/abstract screening and full-text screening were conducted in duplicate in Covidence, with disagreements resolved by consensus or a third reviewer.	Y

- **High concern.** Searches were limited to MEDLINE, Embase, and ClinicalTrials.gov, without additional specialist databases (e.g., PsycINFO, CENTRAL, CINAHL), and Embase conference records were explicitly excluded, a restriction that can directly remove eligible evidence. In addition, nearly two years had elapsed between the literature searches and publication of the report, leaving the search out of date. As a result, the review cannot be assumed to represent the full body of relevant evidence.

Domain 3: Concerns regarding methods used to collect data and appraise studies

3.1 Were efforts made to minimize error in data collection? Response: No. Data extraction was not conducted independently in duplicate. Spot-checking of extraction tables identified multiple errors in a randomly selected sample, indicating that the procedures used were insufficient to minimize data collection error.	N
3.2 Were sufficient study characteristics available for both review authors and readers to interpret the results? Response: Probably no. Although study details are presented in appendices, extensive redaction of information (e.g., study locations) limits the ability to assess study context, relevance, generalizability, and external validity. These redactions obscure contextual details that are part of the public scientific record and do not protect participant privacy.	PN
3.3 Were all relevant study results collected for use in the synthesis? Response: No. Of the 230 eligible primary clinical studies, only 90 underwent data extraction and risk of bias assessment (Figure 1.3), while the remainder were assigned to "bibliography only." The review also states that no formal evidence synthesis was performed and instead presents only summaries of findings.	N
3.4 Was risk of bias (or methodological quality) formally assessed using appropriate criteria? Response: Yes. Risk of bias was assessed using established tools appropriate to study design: the Newcastle–Ottawa Scale (including an adaptation for cross-sectional studies) for observational studies, and the NIH before–after (pre–post) tool for longitudinal studies without control groups.	Y
3.5 Were efforts made to minimize error in risk of bias assessment? Response: Probably no. The review describes the tools used and how items were rated but does not clearly report independent duplicate risk of bias assessment or formal second-reviewer checking. It also states that the systematic review assessments were conducted by a single assessor, suggesting that the same approach was likely used for the clinical studies. There are examples of inflated risk of bias ratings that did not appropriately identify sources of bias (e.g., Chen et al., 2023; Tordoff et al., 2022), and instances where uncontrolled subgroup comparisons (e.g., males vs females) were treated as if they represented control groups.	PN

High concern: The review does not clearly report that risk of bias assessments were conducted independently by two reviewers. The study list for risk of bias (RoB) assessment was narrowed without clear methodological justification, and multiple apparent errors in the assessments, with ratings higher than warranted by the study designs, suggest that the RoB process may not have been applied consistently or appropriately.

Domain 4: Concerns regarding the synthesis and findings

4.1 Did the synthesis include all studies that it should? Response: No. The review describes results without conducting a formal structured synthesis. Section I.4.7.4, which addresses outcomes from comparisons of TGNB adolescents with other TGNB subgroups, illustrates this limitation. For example, the statement that "rates of depression and suicidal thoughts/self-harm tended to be lower among hormonally treated transgender youth compared to untreated transgender individuals" is presented without specifying the treatments involved, the number of studies or participants included, the outcome measures used, or the magnitude of the reported effects. Without a structured synthesis method, it is unclear whether all eligible studies were systematically incorporated into the conclusions.	N
4.2 Were all predefined analyses followed or departures explained? Response: No. The review reports that no formal synthesis was conducted, and no protocol or predefined synthesis-analysis plan is available against which adherence or departures could be assessed.	N
4.3 Was the synthesis appropriate given the nature and similarity in the research questions, study designs and outcomes across included studies? Response: No. The review reports that no formal synthesis was conducted and that conclusions reflect narrative interpretation of individual studies. These interpretations include claims about a "consensus of the evidence" without accounting for heterogeneity in research questions, study designs, and outcomes across the included studies.	N
4.4 Was between-studies variation (heterogeneity) minimal or addressed in the synthesis? Response: No. The review reports that no formal synthesis was conducted. As a result, between-study variation was not formally assessed (e.g., through subgroup analyses, sensitivity analyses, or structured qualitative exploration), and heterogeneity was not systematically addressed in the conclusions.	N

4.5 Were the findings robust, e.g. as demonstrated through funnel plot or sensitivity analyses? Response: No. The review does not report any analyses to assess the robustness of the findings (e.g., sensitivity analyses or other robustness checks).	N
4.6 Were biases in primary studies minimal or addressed in the synthesis? Response: No. Although the review assessed risk of bias, ratings appear inflated and do not adequately account for important sources of bias. The review also reports that no formal synthesis was conducted and that conclusions reflect the authors' interpretation of individual studies. Without a structured synthesis framework, the risk of bias assessments were not incorporated into the findings (e.g., through weighting, stratification, sensitivity analyses, or restricting conclusions to low risk of bias evidence). As a result, important primary-study biases were not addressed when drawing conclusions, including statements that "the consensus of the evidence supports that the treatments are effective in terms of mental health, psychosocial outcomes, and the induction of body changes consistent with the affirmed gender in pediatric GD patients."	N

High concern: The authors state that "we were not contracted to include a synthesis of the evidence that we found." However, systematic reviews are a type of evidence synthesis and require transparent methods for integrating findings across studies and incorporating risk of bias assessments into conclusions about the evidence. In this review, the process by which evidence was combined and translated into findings lacks transparency and methodological safeguards, creating a substantial risk that the conclusions do not reflect the totality or quality of the underlying evidence.

Judging risk of bias

A. **Did the interpretation of findings address all of the concerns identified in the Phase 2 assessment?** Response: No. The review authors failed to address the limitations in the review process, such as feasibility-driven restrictions and prioritization, and incomplete data extraction and risk of bias assessment. Most importantly, the review did not conduct a formal evidence synthesis. As a result, risk of bias assessments were not incorporated into the presentation of results or the formulation of conclusions, and no framework (e.g., GRADE) was applied to assess the certainty of the evidence. Consequently, the level of confidence that can be placed in the review's conclusions cannot be determined.	N
B. Was the relevance of identified studies to the review's research question appropriately considered?	N

	Response: No. Although study characteristics and risk of bias assessments were reported, the review does not demonstrate a structured assessment of the applicability of included studies to the research question when interpreting findings. Conclusions regarding effectiveness and safety are presented broadly without systematic consideration of study design limitations, duration of follow-up, indirectness, or heterogeneity, increasing the risk that issues of relevance were not adequately incorporated into the interpretation.	
C.	**Did the reviewers avoid emphasizing results on the basis of their statistical significance?** Response: Probably no. In the absence of a formal synthesis, and with conclusions presented as broad support for effectiveness and safety across multiple domains, the interpretation risks emphasizing favorable findings without a transparent and balanced accounting of mixed or null results across outcomes and studies. The review does not demonstrate a structured approach to avoid selective emphasis based on statistical significance.	PN

Risk of bias in the review: High

Across Domains 1-4, serious methodological concerns were identified, including post hoc, feasibility-driven restrictions, and "high-priority" filtering in eligibility and extraction (with many potentially relevant studies assigned to bibliography-only), likely incomplete identification of eligible evidence (limited database coverage and exclusion of Embase conference records), and documented data-extraction errors. The review explicitly conducted no formal synthesis, did not assess heterogeneity, and did not integrate risk of bias assessments into its conclusions through a structured method.

A systematic review shouldn't just compile information; its aim is to synthesize (with or without meta-analysis) the evidence in a clear, transparent, and reproducible manner to facilitate informed decision-making. In this case, the review presents conclusions about effectiveness and safety that are not transparently traceable to a reproducible synthesis of the totality and quality of the underlying evidence.