Utah Review: Clinical studies: ROBIS

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

Summary: overall risk of bias is High.

Domain 1: Concerns regarding specification of study eligibility criteria

The review applied broad eligibility criteria, including experimental, observational, and descriptive study designs. Although descriptive studies were initially eligible, data extraction was restricted to longitudinal pre–post designs due to time and resource constraints, with other descriptive studies listed in the bibliography only. Outcomes were not pre-specified. No registered or published protocol was identified, making it unclear whether the objectives and eligibility criteria were finalized a priori.

1.1 Did the review adhere to pre-defined objectives and eligibility criteria? Response: No. There was no protocol to evaluate. This review combined puberty blockers and cross-sex hormones as "gender affirming hormonal therapy," and failed to separately assess and summarize the evidence on each. The reviewers mentioned “high-priority” and other outcomes. It is unclear when and how the “high-priority outcomes” were determined, and the decision was made to include only “high-priority” outcomes. Moreover, some outcome measures (such as Long-term outcomes, and Persistence, desistance, and regret) appear to have been added *after* the search and study screening work had been completed, which is not appropriate.	N
1.2 Were the eligibility criteria appropriate for the review question? Response: The eligibility criteria were unclear, especially regarding outcomes and study types. The authors mentioned 134 eligible primary clinical studies (Section 1.4.5.1) or 230 primary clinical studies (Section 1.6.0). Nevertheless, only 90 had data extraction and risk of bias assessed (Figure 1.3). It is unclear how many studies this review included.	N
1.3 Were eligibility criteria unambiguous? Response: See response to Question 1.2.	PN
1.4 Were all restrictions in eligibility criteria based on study characteristics appropriate? Response: Including all study types led to an overwhelming amount of information, subsequently creating difficulties in synthesizing the findings effectively. For example, Table 1.8 shows the inclusion of everything from surveys to case reports, raising serious questions about what qualifies as relevant evidence and how these sources align with the stated research objectives.	PN

1.5 Were any restrictions in eligibility criteria based on sources of information appropriate? Response: Probably yes.	Y

- **High concern:** A fundamental requirement of any systematic review is a clearly defined scope, including pre-specified eligibility criteria for the inclusion of studies. This review fails to meet that standard. Eligibility criteria were modified 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. This review mentioned including 134/230 primary clinical studies involving over 28,000 pediatric TGNB patients from diverse global settings yet provides no transparent rationale for how such a wide range of study types—all with vastly different designs, populations, and methodological quality—fit within a coherent evidence framework. For example, Table 1.8 shows the inclusion of everything from surveys to case reports, raising serious questions about what qualifies as relevant evidence and how these sources align with the stated research objectives. Without clear inclusion criteria, it is impossible to determine whether this is a focused synthesis of meaningful clinical data or an indiscriminate collection of TGNB-related publications. This lack of scope not only undermines methodological rigor but also compromises the interpretability and reliability of the review's conclusions.

Domain 2: Concerns regarding identification and selection of studies

The review conducted electronic searches of Ovid MEDLINE, Embase, and ClinicalTrials.gov, supplemented by reference list checks and expert recommendations. Full search strategies were provided, using both controlled vocabulary and free-text terms. Title/abstract and full-text screening were conducted in duplicate using Covidence, with disagreements resolved by consensus or by a third reviewer. The search was limited to studies published from 2010 onward, and Embase conference abstracts were excluded.

2.1 Did the search include an appropriate range of databases/electronic sources for published and unpublished reports? Response: Probably no. Although a more extensive search was planned, the reviewers did not have enough time to carry it out, and consequently, not all relevant databases were searched. "We planned a comprehensive search of at least 2 standard bibliographic medical databases, including Medline and Embase. While we also considered options to search Cochrane CENTRAL, PsycInfo, and ClinicalTrials.gov if time allowed, and if needed, but as of the date of this report, we have conducted bibliographic database searches only in Ovid Medline, Embase, and ClinicalTrials.gov."	PN
2.2 Were methods additional to database searching used to identify relevant reports? Response: This review included scans of reference lists.	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 terms seem to be comprehensive and sensible.	Y
2.4 Were restrictions based on date, publication format, or language appropriate? Response: No, the search was outdated when the report was published (March and June 2023, with ClinicalTrials.gov searched in September 2023).	N
2.5 Were efforts made to minimise errors in selection of studies? Response: Yes, this review included duplicate screening and resolution of discrepancies.	Y

High concern is warranted because, despite duplicate screening and detailed search strategies, the identification methods were not sufficiently comprehensive for a broad review question. Searches were restricted 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. When the report was released, about two years had passed, and the literature search was already outdated. Given the breadth of eligible outcomes and study designs, these limitations make it likely that some eligible studies were not identified, so 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

Data were extracted using Excel into evidence tables that recorded study design, participant characteristics, interventions and comparators, outcomes, study period, and follow-up information. Extraction was conducted by members of the review team, but not independently in duplicate. Risk of bias was formally assessed using design-appropriate instruments: the Newcastle–Ottawa Scale (including an adapted version for cross-sectional studies) for observational studies, and the NIH before–after (pre–post) tool for longitudinal studies without control groups. The review does not clearly report that risk-of-bias assessments were conducted independently by two reviewers.

3.1 Were efforts made to minimise error in data collection? Response: No. This review did not properly extract information from primary studies. According to the review, "We included all eligible case reports and case series if they addressed the population and interventions of interest. Initially we had planned to extract data for all descriptive study designs, including case reports that contained a statement about IRB (or other ethics board) review and approval/exemption, or specified that informed consent was obtained from study subjects for research or publication. However, due to time constraints and the number of relevant publications, only longitudinal, pre-post descriptive studies underwent data collection." The reviewers changed their plan on data extraction, citing time constraints. This should not be considered a sufficient justification to deliver a suboptimal scientific report.	N

3.2 Were sufficient study characteristics available for both review authors and readers to interpret the results? Response: No. The report engages in what appears to be performative virtue signaling, notably through its extensive redaction of information that is already publicly available in the original academic publications—such as study locations, years, and patient characteristics. These redactions do not protect participant privacy, as the data are part of the public scientific record, but instead obscure important contextual details necessary for evaluating study relevance, generalizability, and external validity. This practice creates a false impression of ethical rigor while, in reality, it undermines transparency and scientific clarity.	N
3.3 Were all relevant study results collected for use in the synthesis? Response: No. There are 219 studies determined to be "relevant" and included in the bibliography (appendix I.F). In other places, the authors mentioned 134 eligible primary clinical studies (Section 1.4.5.1) or 230 primary clinical studies (Section 1.6.0). Nevertheless, only 90 had data extraction and risk of bias assessed (Figure 1.3). Inconsistencies aside, not all eligible studies were extracted.	N
3.4 Was risk of bias (or methodological quality) formally assessed using appropriate criteria? Response: No. For primary studies, the authors used the Newcastle-Ottawa Scale for cohort, case-control, and cross-sectional studies, but the NIH Quality Assessment Tool for Before-After studies. The authors could have used Newcastle-Ottawa Scale for Before-After studies and presented the risk of bias assessment results in a clearer and more consistent manner. The necessity of using a different risk of bias tool was not justified. Additionally, the Utah review had a tendency to provide inflated assessments to study limitations compared with systematic reviews by researchers from University of York or McMaster university.	N
3.5 Were efforts made to minimise error in risk of bias assessment? Response: Probably no. See response to Question 3.4. There are noticeable disagreements between this review and the systematic reviews by University of York researchers. For example, the study by Tordoff et al, 2022 was rated by this review as having a "moderate" risk of bias (6/9) using the Newcastle-Ottawa scale (NOS). The study received full points for sample representativeness (4/4), cohort comparability (1/1), and also scored a point on follow-up adequacy (p. 570). Yet these ratings are difficult to reconcile with the study's design. Follow-up was only 12 months—widely insufficient to assess longer-term outcomes of cross-sex hormones. The analytic sample was also highly selected (64 of 169 eligible participants, or <40%, remained at 12 months), and the "untreated" comparator shrank to just seven cases, who were inherently atypical as the only clinic patients still untransitioned at 12 months despite having access to treatment. Data extraction and, it seems, risk of bias assessment were performed by a single author.	PN

High concern: Data extraction was not conducted independently in duplicate and spot-checking identified extraction errors. In addition, extraction was restricted for feasibility reasons, with many studies assigned to "bibliography only" or excluded from full extraction, indicating that not all relevant results were collected for synthesis. For risk of bias assessment, the Utah review tend to report inflated assessments

compared with other systematic reviews. The rationale for choosing NIH Quality Assessment Tool for Before-After studies was unclear.

Domain 4: Concerns regarding the synthesis and findings

The review explicitly states that it was not contracted to perform a synthesis and that no formal synthesis was conducted; instead, conclusions reflect the authors' interpretation of individual studies alongside risk-of-bias assessments and evidence tables.

4.1 Did the synthesis include all studies that it should? Response: No evidence synthesis. Rather than synthesizing findings across studies per outcome, the authors merely describe individual study results without aggregating data or drawing meaningful conclusions. Section I.4.7.4, which aims to address outcomes from comparisons of TGNB adolescents to other TGNB subgroups, is an example to illustrate this fundamental failure to conduct a proper evidence synthesis. When the review claimed that "rates of depression and suicidal thoughts/self-harm tended to be lower among hormonally treated transgender youth compared to untreated transgender individuals" there is no indication of which treatments were used, how many studies were included, how many participants were analyzed, which outcome measures or scales were employed, or how large the reported effects were.	N
4.2 Were all predefined analyses followed or departures explained? Response: No evidence synthesis. The review failed to conduct proper evidence synthesis.	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 evidence synthesis. The review failed to address the nature and similarity across included studies.	N
4.4 Was between-studies variation (heterogeneity) minimal or addressed in the synthesis? Response: No evidence synthesis. The review failed to address heterogeneity issue.	N
4.5 Were the findings robust, e.g. as demonstrated through funnel plot or sensitivity analyses? Response: No evidence synthesis. The review failed to assess the robustness of the findings. Lack of evidence synthesis did not stop the authors concluding "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
4.6 Were biases in primary studies minimal or addressed in the synthesis? Response: No evidence synthesis. The review failed to address the risk of bias in primary studies.	N

High concern: The authors claimed "we were not contracted to include a synthesis of the evidence that we found," which is not justified. By definition, systematic reviews are a type of "evidence synthesis." The review explicitly states that no formal synthesis was conducted and that conclusions reflect authors' interpretation of individual studies. In the absence of a predefined quantitative or structured qualitative

synthesis framework, there was no systematic integration of all eligible studies, no assessment of between-study heterogeneity, no sensitivity or robustness analyses, and no formal incorporation of risk-of-bias judgments into the conclusions. As a result, the process by which evidence was combined and translated into findings lacks transparency and methodological safeguards, creating a substantial risk that the conclusions may not reflect the totality or quality of the underlying evidence.

Phase 3

A. Did the interpretation of findings address all of the concerns identified in the Phase 2 assessment? Response: The review authors failed to address the limitations in the review process, such as incomplete data extraction and risk of bias assessment. The most serious concern is this review did not conduct a proper evidence synthesis. Consequently, this review did not take the risk of bias into account when presenting their results or formulating their conclusions, and they failed to apply the GRADE or equivalent framework to assess the quality (certainty) of evidence. There is no way to determine how confident we should be in their conclusions.	N
B. Was the relevance of identified studies to the review's research question appropriately considered? Response: No. This review did not consider the relevance or applicability of the study findings.	N
C. Did the reviewers avoid emphasizing results on the basis of their statistical significance? Response: Probably no. The review authors did not include their own pooled analysis. But when the review authors described results of included studies, they emphasized results based on statistical significance.	PN

Risk of bias in the review: High

Across Domains 1–4, serious methodological concerns were identified, including 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, reproducible, and efficient manner to facilitate informed healthcare decision-making. The review authors failed the fundamental purpose of conducting a systematic review. Despite these limitations, it presents broad affirmative claims regarding effectiveness and safety that are not transparently traceable to a reproducible synthesis of the totality and quality of the underlying evidence.