Supplement to Treatment for Pediatric Gender Dysphoria

Review of Evidence and Best Practices

Peer Reviews and Replies

Department of Health and Human Services

November 19, 2025

Supplement to Treatment for Pediatric Gender Dysphoria

Review of Evidence and Best Practices

Peer Reviews and Replies

Department of Health and Human Services

November 19, 2025

Suggested Citation:

U.S. Department of Health and Human Services (HHS), *Supplement to* Treatment for Pediatric Gender Dysphoria: Review of Evidence and Best Practices. *Peer Reviews and Replies*. Washington, DC: HHS, November 2025.

Introduction

Treatment for Pediatric Gender Dysphoria: Review of Evidence and Best Practices ("the Review") was published by the U.S. Department of Health and Human Services (HHS) on May 1, 2025; revisions were made on May 15 (see Errata). Following post-publication peer reviews, the Review and the separate Appendix 4 were revised further.

Individual replies to seven solicited peer reviews, together with a reply to two unsolicited peer reviews (Dowshen et al., 2025; Rider et al., 2025), follow. In addition to the changes to the Review and Appendix 4 explicitly noted in the replies, we have made further minor corrections and improvements to clarity and readability. These include typographical fixes, small alterations of wording and formatting, incorporation of some publications that appeared after the Review was first published, web archive links for items in the bibliography, and the addition of a table of contents to Appendix 4.

In an effort to solicit and incorporate feedback from major medical organizations that have expressed support for pediatric medical transition, HHS invited the American Academy of Pediatrics (AAP), the American Psychiatric Association (APA), and the Endocrine Society to participate in the peer review process. All three groups have criticized the Review, with the AAP condemning it in an official statement within hours of its release.[1] Unfortunately, the AAP and the Endocrine Society refused HHS's offer to participate. We are grateful to the APA for accepting the invitation.

[1] American Academy of Pediatrics (2025). See also American Psychiatric Association (2025) and Kellner & Bascom (2025).

Contributors[2]

Evgenia Abbruzzese, Society for Evidence-Based Gender Medicine

Alex Byrne, PhD, Massachusetts Institute of Technology

Farr Curlin, MD, Duke University

Moti Gorin, PhD, MBE, Colorado State University

Kristopher Kaliebe, MD, DFAACAP, University of South Florida

Michael K. Laidlaw, MD, Michael K. Laidlaw MD, Inc.

Kathleen McDeavitt, MD, Baylor College of Medicine

Leor Sapir, PhD, Manhattan Institute for Policy Research

Yuan Zhang, PhD, Evidence Bridge

Contributor biographies

Evgenia Abbruzzese is a healthcare researcher. She has led analytics efforts for a major insurer to identify low-value care and overuse of invasive and non-beneficial interventions, and also led a venture-backed healthcare startup to help patients with medically unexplained symptoms and health anxiety. Abbruzzese has published on topics in pediatric gender medicine, including the original Dutch research underpinning the practice of pediatric medical transition.

Alex Byrne is a professor in the Department of Linguistics and Philosophy at the Massachusetts Institute of Technology. He holds a PhD in philosophy from Princeton University. Byrne has published on gender identity and related topics and is on the editorial board of the *Archives of Sexual Behavior*.

Farr Curlin is an internist and professor at Duke University School of Medicine. He holds an MD from the University of North Carolina at Chapel Hill and completed residency at the University of Chicago, including a MacLean Center Fellowship in Clinical Medical Ethics. He is board-certified in internal medicine. Curlin has published and lectured on medical ethics and the practice of medicine, and is Co-Director of the Theology, Medicine and Culture Initiative at Duke.

Moti Gorin is an associate professor in the Department of Philosophy at Colorado State University. He holds a PhD in philosophy from Rice University and was a post-doctoral Fellow in Advanced Biomedical Ethics at the University of Pennsylvania. Gorin has published on topics at the intersection of pediatric gender medicine and medical ethics.

[2] Contributor affiliations are listed for identification purposes only.

Kristopher Kaliebe is a psychiatrist and professor at the University of South Florida Morsani College of Medicine. He holds an MD from St. George's University School of Medicine and is board-certified in general psychiatry, forensic psychiatry, and child and adolescent psychiatry. Kaliebe treats young people with gender-related distress and is a distinguished fellow of the American Academy of Child and Adolescent Psychiatry.

Michael K. Laidlaw is an endocrinologist in private practice. He holds an MD from the University of California Keck School of Medicine where he went on to complete his residency and fellowship training and was involved in medical research. Laidlaw is board-certified in endocrinology, diabetes, and metabolism, and has published on gender dysphoria and related topics.

Kathleen McDeavitt is a psychiatrist and associate professor at Baylor College of Medicine. She holds an MD from the University of North Carolina at Chapel Hill and completed psychiatry residency at Baylor College of Medicine. McDeavitt is board-certified in general psychiatry and has published on topics related to the evidence base and clinical guidelines in pediatric gender medicine.

Leor Sapir is a senior fellow at the Manhattan Institute for Policy Research. He holds a PhD in political science from Boston College and was a post-doctoral fellow at the Program on Constitutional Government at Harvard University. Sapir has published on pediatric gender medicine, public policy, and related topics.

Yuan Zhang is a Canadian researcher and the founder of Evidence Bridge, which supports evidence-based health policy and decision making. He holds a PhD in health research methodology from McMaster University. Zhang has published on evidence-based medicine, guideline development, patient values and preferences, and health economics.

Disclosures of Interest (as of November 14, 2025)

Evgenia Abbruzzese has received payments for legal consultations related to pediatric gender medicine.

Farr Curlin has received payments for expert testimony and honoraria for speaking engagements related to pediatric gender medicine.

Kristopher Kaliebe has received payment for expert testimony related to pediatric gender medicine as well as an honorarium to attend a conference related to pediatric gender medicine.

Michael Laidlaw has received payment for expert testimony related to gender dysphoria as well as an honorarium to attend a conference related to pediatric gender medicine.

Leor Sapir has received payments for legal consultations and honoraria for speaking engagements related to pediatric gender medicine.

No other disclosures were reported.

Peer reviews

American Psychiatric Association

AMERICAN
PSYCHIATRIC
ASSOCIATION
800 Maine Avenue, S.W.
Suite 900
Washington, D.C. 20024

To: [redacted]
Management and Program Analyst
Office on Women's Health
Office of the Assistant Secretary for Health
U.S. Department of Health and Human Services

Date: July 17, 2025, Updated September 26, 2025

Re: Request for APA to be a reviewer for the HHS Report: "Treatment for Pediatric Gender Dysphoria: Review of Evidence and Best Practices."

The APA appreciates the opportunity to be a peer reviewer for the Health and Human Services Department (HHS) Report: "Treatment for Pediatric Gender Dysphoria: Review of Evidence and Best Practices." Our conclusions are that while the HHS Report purports to be a thorough, evidence-based assessment of gender-affirming care for transgender youth, its underlying methodology lacks sufficient transparency and clarity for its findings to be taken at face value. Key elements including literature selection criteria, analytical frameworks, and justification for excluding other studies, and key findings in studies on which the Report relies, are either underexplained or absent. As a result, the Report's claims fall short of the standard of methodological rigor that should be considered a prerequisite for policy guidance in clinical care.

Below are some specific comments on the Report's methodology:

- With one exception, the authors of the report are not identified. Transparency regarding authorship is essential to the integrity of scientific and policy analysis because it allows readers to assess the expertise of contributors, evaluate their qualifications in relevant fields, and identify potential conflicts of interest or ideological commitments.
- The Report fails to clearly articulate how the studies were selected, what criteria governed their inclusion or exclusion, or how their quality was assessed. This lack of

methodological clarity is particularly concerning given the Report's critique of other systematic reviews.

- The Report fails to address the risk of confirmation bias, a critical oversight. Confirmation bias refers to the cognitive inclination to favor information that affirms one's existing beliefs while discounting or overlooking evidence that challenges them.

Such confirmation bias may exist where, as here:

- The Report fails to take into consideration conclusions of the Cass Review that do not support the Report's outcome. For example, the Cass Review at p. 21 notes that improved access, expert, holistic, comprehensive and individualized assessment, as well as treatment of co-occurring mental health conditions are essential (all of which are consistent with current guidelines) and that while gender-affirming medical interventions are not appropriate for all transgender youth, "for some, the best outcome will be transition."

- There is no indication that key stakeholders - namely, transgender individuals, their families, and clinicians - were consulted or that their perspectives were considered. A comprehensive review of best practices would include input from recipients (both those for whom treatment was beneficial, and those for whom it was not), families, and providers of the treatments under evaluation.

- While the Report is clear about the potential harms of intervening medically, it does not apply any kind of rational scrutiny to potential harms that have been associated with withholding intervention, including higher rates of depression, anxiety, suicidality, and social withdrawal.

- The Report draws heavily from the Cass Review which itself has been criticized by experts for its methodological flaws and biases. See, e.g., https://law.yale.edu/sites/default/files/documents/integrity-project_cass-response.pdf; https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-025-02581-7

Below are additional studies and reports for review and consideration:

Chen D et al., Psychosocial functioning in transgender youth after 2 years of hormones. N Engl J Med 2023; 388;240-250

de Vries, A. L. C., Steensma, T. D., Doreleijers, T. A. H., & Cohen-Kettenis, P. T. (2011). Puberty suppression in adolescents with gender identity disorder: A prospective follow-up study. *Journal of Sexual Medicine, 8*(8), 2276–2283. https://doi.org/10.1111/j.1743-6109.2011.02316.x

de Vries, A.L.C, et al., Young Adult Psychological Outcome After Puberty Suppression And Gender Reassignment, 134(4) PEDIATRICS 696–704 (2014), https://pubmed.ncbi.nlm.nih.gov/ 25201798.

Green AE et al. Association of gender-affirming hormone therapy with depression, thoughts of suicide, and attempted suicide among transgender and nonbinary youth. J Adolesc Health 2022; 70(4):643-649

Hughto, J. M. W., Gunn, H. A., Rood, B. A., & Pantalone, D. W. (2020). Social and medical gender affirmation experiences are inversely associated with mental health problems in a US nonprobability sample of transgender adults. *Archives of Sexual Behavior, 49*(7), 2635–2647. https://doi.org/10.1007/s10508-020-01655-5

LaFleur J, Heath L, Gonzalez V., et al. Gender-affirming medical treatments for pediatric patients with gender dysphoria. A report of the University of Utah College of Pharmacy Drug Regimen Review Center (DRRC). Salt Lake City, UT: University of Utah: 2024 https://le.utah.gov/AgencyRP/reportingDetail.jsp?rid=636

Luke R. Allen et al., Well-Being and Suicidality Among Transgender Youth After Gender Affirming Hormones, 7(3) CLINICAL PRAC. PEDIATRIC PSYCH. 302 (2019), https://psycnet.apa.org/record/2019-52280-009

Murad, M. Hassaan, et al., Hormonal Therapy and Sex Reassignment: A Systematic Review and Meta-Analysis of Quality of Life and Psychosocial Outcomes, 72(2) CLINICAL ENDOCRINOLOGY 214 (Feb. 2010), https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2265.2009.03625.x

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Olsavksy AL et al. Associations among gender-affirming hormonal interventions, social support, and transgender adolescents' mental health. J Adolesc Health 2023; 72(6):860-868

Rosenthal, Stephen M, Challenges in the Care of Transgender and Gender-Diverse Youth: An Endocrinologist's View, 17(10) NATURE REV. ENDOCRINOLOGY 581, 586 (Oct. 2021), https://pubmed.ncbi.nlm.nih.gov/34376826

Taylor, J., Mitchell, A., Hall, R., Heathcote, C., Langton, T., Fraser, L., & Hewitt, C. E. (2024). Interventions to suppress puberty in adolescents experiencing gender dysphoria or incongruence: A systematic review. *Archives of Disease in Childhood, 109*(Suppl 2), s33–s47. https://doi.org/10.1136/archdischild-2023-326669

Tordoff, D. M., Wanta, J. W., Collin, A., Stepney, C., & Inwards-Breland, D. J. (2022). Mental health outcomes in transgender and nonbinary youths receiving gender-affirming care.

JAMA Network Open, 5(2), e220978.
https://doi.org/10.1001/jamanetworkopen.2022.0978

Turban, J. L., King, D., Carswell, J. M., & Keuroghlian, A. S. (2020). Pubertal suppression for transgender youth and risk of suicidal ideation. *Pediatrics, 145*(2), e20191725.
https://doi.org/10.1542/peds.2019-1725

Turban, JL, et al., Access To Gender-Affirming Hormones During Adolescence and Mental Health Outcomes Among Transgender Adults, J. PLOS ONE (2022), https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0261039

van der Miesen AIR, Steensma TD, de Vries ALC, Bos H, Popma A.J Adolesc Health. 2020 Jun; Psychological Functioning in Transgender Adolescents Before and After Gender-Affirmative Care Compared with Cisgender General Population Peers. 66(6):699-704. doi: 10.1016/j.jadohealth.2019.12.018. Epub 2020 Apr 6. PMID: 32273193

Dr. Johan C. Bester

Peer Review: Treatment for Pediatric Gender Dysphoria (HHS Report), May 1, 2025
Reviewer: Johan C Bester, MBChB, PhD, HEC-C, Associate Dean for Pre-clerkship Curriculum, Professor of Family and Community Medicine, Professor of Healthcare Ethics, Saint Louis University School of Medicine
Submission of review: July 3, 2025

Peer reviewer background:
I hold a medical degree and practiced in family medicine and emergency medicine settings for 12 years. I then made a full-time transition to bioethics and academia. I completed a PhD in Applied Ethics, with a focus on ethical issues related to measles vaccination. My scholarly work has focused on pediatric ethics and ethical issues in vaccination. I have published extensively on the ethics of medical decision-making in children and adolescents. I teach in the medical curriculum, and have designed and taught courses in medical ethics, epidemiology, and evidence-based medicine.

Peer review focus:
I have read the entire report (Treatment for Pediatric Gender Dysphoria, May 1, 2025, hereafter called "the Review"), with a focus on the evidence review, methods, and ethical analysis. I will provide general comments on the Review and its findings, and then focus specific attention on the methods of the evidence review and the ethical analysis.

Overall impression:
The Review is a review of the evidence and ethical analysis of interventions offered to minors (adolescents and children) who have gender dysphoria. This is an important and timely work. It is well written, methodologically rigorous, and makes a significant contribution to the discussion on this topic. I will point out some areas where I would recommend minor improvements or further analysis. These serve as recommendations for improving the work, and do not affect the overall findings of the Review. What is here is thorough, compelling, and well done. The main findings and recommendations of the Review is consistent with the findings and recommendations of other high-level evidence reviews and analyses that have been published on this topic.

Main findings and conclusions of the review:
1. There is no compelling evidence of benefit for gender transition interventions offered to minors with gender dysphoria. The evidence base is of low quality. What evidence there is does not demonstrate any clear benefit resulting from social transition, puberty blockers, cross-sex hormones, or gender modification surgery.

2. There are significant concerns about the potential harms from gender transition interventions offered to minors with gender dysphoria, and in particular puberty blockers, cross-sex hormones, and gender modification surgery. The best available evidence for potential harms comes from what is known about these interventions from use in other settings and from basic science knowledge. There is a lack of robust studies that have investigated these harms in the setting of gender transition in minors.
3. The Clinical Practice Guidelines that have been most influential and have informed practice related to gender dysphoria in minors in the United States are of low quality. In particular, the WPATH Standards of Care and the guidelines developed by the Endocrine Society are beset by problems that make them unsuitable for use. These include conflicts of interest, methodological problems, and misapplication or misrepresentation of available evidence.
4. Given the lack of demonstrable benefit and concern about potential harms, the use of puberty blockers, cross-sex hormones, and gender transition surgery in minors with gender dysphoria cannot be ethically justified.
5. There is no compelling evidence from studies that psychotherapy as treatment for gender dysphoria is beneficial. However, psychotherapy is generally shown to be beneficial for the treatment of psychiatric distress and various mental illnesses, and is thought to carry low risk of harm.

Discussion of the main findings and contributions of this review:
The main findings and conclusions of the review are correct. It affirms what we know from other reviews and from direct analysis of available studies on this topic. Here is what we know. The prevalence of gender dysphoria has increased among children and adolescents in the last 15 to 20 years. Reasons for this are not quite understood but are likely multifactorial. The epidemiology of gender dysphoria has changed. By far the majority of cases will resolve without any intervention or treatment. There is no compelling evidence of benefit from gender transition treatments for resolution of dysphoria or for management of co-morbidities. The treatments used for gender transition are generally not reversible, have long-term implications, and carry the potential for harms.

Two things are made clear by these facts. First, that current practice in the United States of offering these interventions to minors with gender dysphoria ought not continue. Second, that much research is needed to clarify questions related to present-day gender dysphoria in the United States, inclusive of possible causes, the natural course, and potential treatments. Despite these facts, re-affirmed by the Review, clinical practice in the United States has proceeded in the opposite direction than would be expected. Practice and public discourse have continued as if there is compelling evidence for the benefit of gender transition treatments, as if there is a low risk of harm, and as if it is settled that these interventions decrease mortality and morbidity. Further research about harms, other treatment modalities, causes of

gender dysphoria, and natural course is not only not happening, but met with hostility and suspicion.

Various explanations are possible, given that context in the United States surrounding these matters. First, practitioners and supporters of gender medicine appear to have an *a priori* commitment to the provision of gender transition treatments to minors. What came first was certainty that these treatments are beneficial and medically necessary. What came second was seeking for evidence that would justify this proposition after the fact. When no evidence was forthcoming, shifting justifications for treatment emerged. Ultimately, a commitment to the provision of such interventions became a sort of litmus test, where one's support for children with gender dysphoria and one's credibility as an expert in this arena was contingent on a rock-solid commitment to gender transition interventions. Second, the issue became seen as a civil rights issue and became a matter of politics and activism, rather than a medical issue. Any attempt to ask questions about these treatment approaches was seen as actions of oppression and led to vilification. Practitioners and members of the public were afraid to ask questions, afraid to speak up.

Given this background, it becomes clear how important the Review is. What the Cass review did in the UK, the Review does in the United States. Firstly, it places the issue of treatment for gender dysphoria back where it belongs. This is a medical issue, that should be approached with the usual standards that govern medical treatments for minors and the usual evidence-based approach essential to the justification of medical treatments. Secondly, it highlights the importance of practice change in the United States, the need for the medical profession to self-regulate, and the need for the State to intervene to protect children should the medical profession fail to do so. Thirdly, it raises questions about the way in which medical professional organizations, medical institutions, and professionals were swayed by considerations outside of evidence and the good of their patient in this particular issue. It is a call to the medical profession and its institutions to re-evaluate itself. Are we really as evidence-based as we think? Are we practicing as we should? Are we the safe-guard for patient interests we profess to be? Can we interpret evidence and clinical practice guidelines as the public would expect expert physicians to be able to do? If so, why were we swayed to support interventions for which no evidence exists that may risk harm? Fourthly, the Review, together with the Cass review, removes the stifling atmosphere that suppressed dialogue, debate, and reasoned inquiry into matters related to gender dysphoria and treatment in minors. It recenters the issue as a medical one, where medical professionals and researchers with different beliefs and views can discuss and dialogue with a focus to seeking the best treatments and outcomes for their patients. It is critical that we have open dialogue about these matters, in a way that fosters public trust and patient trust. The Review goes a long way to re-setting the public atmosphere that would allow such open discussion, and would remove the fear, the retribution, and the self-censoring. This is also why it is helpful that the Review includes sections on the history of gender dysphoria and related treatments,

the role of WPATH, and the social and clinical environment in the United States. The Review is clear about the context within it is written, and seeks to provide insight into that context for readers.

There is no doubt that the Review would be attacked to discredit it by some of those who are in opposition to the contributions and conclusions of the Review. I would expect that, like the Cass review, the Review may be painted by some as ideological, transphobic, and ill-conceived. So, one must look at the Review carefully to see if there is any evidence of political bias, ideology, or transphobia. On reading it, there are no indications that the Review has such components or influences. It is clear that the Review is focused on the medical issue of whether these treatments may benefit minors with gender dysphoria, not larger social issues or ideological questions related to gender. The primary issue that the Review is focused on is the good of minors with gender dysphoria. The Review is agnostic as to different views on gender and gender identity. The Review takes it for granted that there are people who identify as transgender, live out their transgender identity, and that the rights of such people to live their lives as they see fit should be respected. The Review takes into account various views on sexuality and sexual orientation, and finds the middle ground focused on evidence and the ethical considerations central to medicine that would appeal to those of different persuasions. In fact, one big contribution of the review is to move the discussion around these topics towards more neutral ground and neutral language such as one would find in general when studying medical treatments or medical questions. The Review does not appear to be transphobic, does not take sides in social disagreements, and does not advocate for specific policies. It does an excellent job of placing the questions around gender dysphoria back on a neutral footing, where the usual standards that pertain to medical treatments can be applied to the issue.

Comments on methodology and evidence review:

1. The review of evidence for benefit of gender transition interventions:

The Review conducts a Systematic Review of Systematic Reviews. In chapters 5 and 6, there appears a description of methodology that shows inclusion criteria, search criteria, a flow diagram to account for studies identified, included, and excluded. The Review makes use of an appropriate tool to assess for bias and analyzes the included studies appropriately. Moreover, the Review directly engages with an analysis of the observational studies that have been central to previous claims of purported benefit from gender transition treatments in minors. These methods all appear sound, and one can be confident that this systematic review is reproducible, consistent with the methods of systematic reviews of this kind, and has identified the best available evidence related to the clinical questions.

2. The review of evidence for harm of gender transition interventions:

The Review analyzes studies included in the systematic review and shows how they fall short in monitoring for potential harms. Consistent with best practice in evidence-based medicine, the Review then seeks the best available evidence, which is what is known about these interventions in general when used for other purposes, and what is known about these interventions from basic and clinical science considerations. The Review identifies a set of harms that are likely to occur with great certainty, and a set of additional potential harms that may occur with use of these interventions. The methods here are appropriate, and the conclusion that clinicians should be weary of the potential for harm with these interventions is sound.

3. The review of Clinical Practice Guidelines:

The Review makes use of appropriate methods to analyze prominent CPGs that have influenced medical practice related to gender dysphoria. CPGs must meet certain criteria to attain to trustworthiness, inclusive of being based on a systematic review, meeting criteria for formation of the group that creates the guideline, avoiding and mitigating conflicts of interest, and linkage of recommendations to strength of evidence using a recognized method such as GRADE. The Review appropriately analyzes available CPGs using these criteria, and demonstrates which CPGs meet the quality standard attaining to trustworthiness, and which do not. We can have confidence in the findings here, based on the methods.

4. The review of evidence related to psychotherapy as treatment for gender dysphoria:

The final chapter of the Review is focused on psychotherapy. The evidence review here draws on the systematic review in Chapter 5, and it concludes that there is a lack of evidence for psychotherapy in minors with gender dysphoria because it has not been adequately studied in this context. However, there is evidence that psychotherapy can be useful for managing co-morbidities that often accompany gender dysphoria, such as depression or anxiety. This seems reasonable, however one would need to proceed with caution. We cannot just assume that because psychotherapy benefits minors with mental illness but don't have gender dysphoria, that psychotherapy would have the same benefits and risks for the treatment of distress related to gender dysphoria. The history of medicine is full of stories of interventions that seemed reasonable based on inferences such as these, where further studies demonstrated lack of benefit or potential harms. While it seems common sense to say that clinicians should try psychotherapy for these patients, and it seems the risk for harm is not high, one should not strongly endorse psychotherapy as a treatment modality for gender dysphoria with the given evidence base. A strong recommendation here should be that further studies of psychotherapy in the context of gender dysphoria are needed, given the lack of evidence.

5. Areas for improvement in methodology related to evidence review:

The Review would benefit significantly from making clear who wrote the review, how the writers were selected, what the specialties and areas of focus of those writers are,

whether a methodologist was included in the review and writing group, what conflicts of interest exist, and how those conflicts of interest were mitigated. The absence of this information leaves a gap in the methodological assessment of the Review. Like with the CPGs that were found to be of low quality, it is important that readers can assess these variables when reading and interpreting the Review.
Based on the level of evidence related to psychotherapy in the Review, I would recommend that the Review make a stronger suggestion for further studies of psychotherapy as a treatment method for gender dysphoria. This seems to be a key insight that should be highlighted.

Comments on the Ethical Analysis (chapter 13):
1. Informed consent
The Review considers the issue of informed consent, and rightly highlights that there is controversy about whether minors can provide consent for gender transition interventions.
It then examines briefly arguments for and against the idea that minors provide consent. The Review then briefly discusses that information is usually not shared in a way that allows for a full informed consent. The Review does not reach a full conclusion on these matters, but instead then pivots to a discussion of the risk-benefit profile of these interventions, citing the primacy of benefit and risk in medical decision-making and in pediatrics specifically.
What is here is fine, but I would really have liked to see some expanded argumentation and analysis of the issue of consent for these interventions. There has been among some proponents of gender transition the explicit or implicit view advanced that minors should provide consent for these procedures themselves, and that any desire for gender modification is sufficient to authorize these procedures. It has struck me for a while now that the pressure to allow minors to lead decision-making in this area departs markedly from how medical decision-making for minors is usually done. Usually, parents are decision-makers for minors, and together with clinicians make decisions that serve the best interests of the minor. The minor engages and gives assent in an age-appropriate fashion but is not ultimately the authorizer or decision-maker of treatment. There are exceptions, for instance in the areas of mental health treatments or treatment for sexually transmitted infections, where adolescents can authorize treatment themselves. This is usually based on the interests of the minor and the public good, and not on arguments from a minor's supposed autonomy. The whole way in which decisions are made for minors have in mind to protect them from harmful decisions, to advance their interests, and to hold their future autonomy and health in trust. The interests of the minor are primary in medical decision-making for the minor. It is not the minor's wishes, desires, refusals, or impressions about their own health that is primary. However, some advocates for gender transition want decision-making for minors in the context of gender transition interventions to work in exactly the opposite way: minors should be the drivers of decision-making, and it is the desires and wishes of minors that primarily authorize the provision of treatment. Given this, and the prominence of these views now within gender medicine, it would

be good to see a thorough analysis of consent for these procedures, over and above the question of benefit. In essence, I'd like to see more description of how medical decisions are usually made for minors, the ethical reasons why this is so, and then how decisionmaking and consent procedures differ in the case of gender transition. A central question is whether minors have the capacity to consent for these treatments. Based on what we know about adolescent development, decision-making capacity, and how to best make medical decisions for minors, my view is that the answer would be that minors cannot be asked to consent for these treatments, nor lead the decision-making around them.

The Review does summarize well how the practice of gender medicine in the US has fallen below the standards of what is required of informed consent. A valid consent process requires a number of things, including full disclosure of relevant information, and voluntariness. The Review demonstrates that full disclosure has not happened in many gender treatment practice settings in the United States, and the true state of lack of benefit and potential harms have generally been obscured when consent is sought from parents and minors. Further, language has been employed on a routine basis that undermines the voluntariness of consent or permission for treatment from parents. The phrase "You can either have a dead daughter or a live son" has been used on routine basis by practitioners of gender transition to push doubting parents to provide consent for gender transition. For one thing, this phrase obscures the truth – there is no evidence that gender transition is lifesaving or that gender dysphoria inevitably leads to death. But more importantly, this is a coercive phrase that places irresistible pressure on parents to acquiesce to gender transition. It removes the voluntariness of the consent process. What loving parent can resist anything if told that the alternative is that their child will be dead? Most parents would pay any cost to save their child's life. So egregiously does this phrase manipulate the care and concern of parents for their children, so much does it bypass rational reasoning, that it can be seen as a coercive lever used by providers of gender transition to force parents into complying. This is no valid informed consent; yet treatment proceeds as if it is. This is a serious ethical violation, something that should cause great concern. The phrase "You can either have a dead daughter or a live son" should live in infamy in the annals of medical ethics in perpetuity. It is dishonest, untrue, coercive, and undermines sound medical decision-making for vulnerable minors.

2. Benefit/risk analysis

The Review spends much time analyzing the implications of benefit and risk of harms related to gender transition interventions in minors. The Review rightly argues that informed consent and shared decision-making are necessary but are not sufficient to justify medical treatments. There must be evidence of benefit that outweighs risk of harm. There is no obligation on medical professionals to offer non-beneficial treatments, and there is no patient right to demand non-beneficial treatments. Further, medical professionals ought not offer treatments where there is uncertainty of benefit while there is risk of harm, or where harms outweigh benefits. Thus, the

question of benefits and risk of harm are primary to the ethical analysis of any treatment, and certainly gender transition treatments.

This argument is correct, and the analysis here by the Review is strong. Given the lack of evidence for benefit, the potential for harms, and the long-term implications of gender transition treatments, there is no ethical basis for offering these treatments to minors. Whether these treatments are offered or not does not depend on one's ideological view of gender, on whether one wishes they worked, on support for a struggling minor, on civil rights, on concern for transgender people, or any other consideration. It simply hinges on this: is there evidence for benefit? Does the expected benefit outweigh the potential harms? Any intervention that cannot clear these bars cannot be offered as treatment to patients.

The arguments here can be made even stronger by reference to the ethical standards that govern medical decision-making for minors. In minors, the primary standard that governs medical decision-making is best interests. When faced with a treatment choice, the various options and interventions should be weighed against the child's various interests. For each possible intervention, careful consideration should be given to how the intervention would advance the minor's interests, and how the intervention would set back the child's interests. This should be compared to the effect on interests of doing nothing. In the end, parents and clinicians should choose the course of action that has the highest likelihood to advance the various interest of the child. When making decisions for children with gender dysphoria, we must recognize that there is no clear evidence that gender modification would benefit the minor. There are risks of harms that would set back the minor's welfare. A large majority of cases of gender dysphoria resolve without any treatment, with good long-term outcomes. Given these facts, in the vast majority of cases it would seem the best interest of the child would not be served by medical interventions aimed at gender transition.

3. Alternative clinical rationales

The Review considers the shifting justifications for gender transition treatments that have been presented. At first, puberty blockers were justified as a "pause" that gives minors time to think; now it is clear the puberty blockers are an entry to further gender transition. Next, gender transition interventions were justified as being life-saving, beneficial, and medically necessary to prevent suicide and improve mental health. When it became clear that there was no evidence for this position, the justification shifted again. We see now the emergence of a justification that sees the provision of medical gender transition as the fulfillment of patient wishes and desires, as meeting the embodiment goals of the individual. The Review rightly argues that this is not in keeping with the usual standards of justification for medical treatments. More importantly, though, is that this mode of justification depends on an individual view of what is good for the self, of a set of long-term goals and autonomous action that is more suited to adults living in a liberal society than to minors who are still developing

their identity and view of the good. Making decisions with long term, high-stakes implications of this sort without significant protections and guardrails is beyond the capacity and developmental stage of minors, and therefore fails as a justification for the provision of these treatments in minors.

4. Justice

The arguments from justice add to the overall analysis. I would just add a slight amendment here. The Review pulls the principle of justice from the Belmont report, which is focused on research, and then says it also can apply to medical care. There is no need to proceed in this way. Justice is one of the principles of medical ethics in the principlism approach of Beauchamp and Childress, and is widely recognized as one of the ethical principles central to good medical practice. The justice arguments offered here are central to medical ethics, and therefore are applicable in a consideration of the ethical status of interventions for gender transition.

5. Summary

Overall, I find the ethical analysis compelling and on point. It draws on arguments relevant to medical ethics, and proceeds with analysis that is thorough and to the point. There are some areas where I wish the Review would have gone further in its analysis, particularly around informed consent. I also wish there was more in here about how decisions are usually made for children, and the ethical guardrails in place to protect the interests of children. But even without these, the analysis here is strong.

Professor Karleen Gribble

WESTERN SYDNEY
UNIVERSITY

School of Nursing and Midwifery
Building ER, Parramatta Campus
+61 (0) 431 118485
k.gribble@westernsydney.edu.au

11 July 2025

Thank you for giving me the opportunity to review the US Department of Health and Human Services publication, 'Treatment for Pediatric Gender Dysphoria: Review of Evidence and Best Practices.' I provide the following comments based upon my expertise in relation to language, breastfeeding and detransition.

Language

It is noted in the introductory section of the executive summary of the review that,

The understandable desire to avoid language that may cause discomfort to patients has, in some cases, given rise to modes of communication that lack scientific grounding, that presuppose answers to unresolved ethical controversies, and that risk misleading patients and families. This Review uses scientifically accurate and neutral terminology throughout.

I commend this approach and the explanation provided for this reasoning in Chapter 2. Reasoning provided for rejection of terms such as 'sex assigned at birth,' use of 'gender identity' rather than just 'gender' and use of sexed language generally in the Review is well argued. I recommend that it would be worthwhile to add text on the risks of using terminology suggesting that people can change their sex. These risks include providing encouragement for people to change their sex markers in their health records (with associated adverse health consequences for individuals e.g. [1, 2]) and for incorrect recording of sex generally with resultant corruption of statistics, as described by Sullivan et al. [3]. In line with this, I would suggest that the terms 'male-to-female' and 'female-to-male' in the Review be replaced with other descriptors and that the term 'sex reassignment surgery' be reconsidered as these terms suggest that sex can be changed.

I also note that the Review does not make it clear that not everyone applies the concept of gender identity to themselves. Amongst those who appear more commonly to overtly reject personal application of gender identity are women who see it as regressive[4] and detransitioners who believe they were harmed by this concept[5]. The review would benefit from some text noting the non-universality of gender identity.

Chest masculinisation surgery

In considering the risks and benefits of treatments for paediatric gender dysphoria the Review notes,

To discharge their duties of nonmaleficence and beneficence, clinicians must ensure, insofar as reasonably possible, that any interventions they offer to patients have clinically favorable risk/benefit profiles relative to the set of available alternatives, which includes doing nothing.

The Review also states that,

The claims made here about the probability and magnitude of harms and benefits are grounded in the best available evidence. Sometimes, the probabilities are known with a high degree of certainty. For example, the probability that mastectomy will lead to an inability to breastfeed is 1.0 or close to it… As for the nature of medical benefits and harms and their relative weights, the Review's working assumptions cohere with common moral intuition, standard medical judgment as revealed in medical diagnostic criteria, and the outcomes of interest to clinicians and researchers, as well as the law. For example, the analysis would conclude that a minor improvement in depressive symptoms does count as a benefit but that such a benefit, even if assured, does not outweigh moderate or even low but nonnegligible risks of infertility or serious sexual dysfunction, loss of breastfeeding function, or lifelong medical dependency, which the Review considers harms.

Finally, the Review states,

We can be certain in the ordinary sense of "certain" that these interventions cause harm, even if we do not have "high certainty" evidence in the technical sense employed in evidence based medicine (EBM).41 We do not need results from RCTs to be certain that removing an adolescent's breasts will eliminate or substantially impair capacity for breastfeeding.

I commend the authors for taking this approach. It has been frustrating to see systematic reviews of the evidence of interventions not consider known outcomes simply because those undertaking the primary research have not included them. However, despite the statements quoted above, neither the body of the Review nor the overview of the systematic reviews includes harm in terms of inability to breastfeed in the analysis of findings. I would suggest that this be addressed. Added content should note that implications of chest masculinisation surgery may include psychological distress[5] and adverse health outcomes for children (including increased risk of necrotising enterocolitis, infections, SIDS and impaired development) and mothers (including increased risk of ovarian cancer and type 2 diabetes)[6].

Misinformation about the ability to breastfeed after chest masculinisation surgery is widespread and health websites and academic publications publish content that is generally unreasonably optimistic. It may be helpful to provide a citation explaining the nature of chest masculinisation surgery to make it clear why breastfeeding is prevented[7].

Breast binding

The Review does not discuss breast binding. Breast binding, usually referred to as 'chest binding', is considered to be a part of social transition (much like changing hairstyle or clothing) rather than being a physical intervention. However, it is very much a physical intervention. Breast binding is

supported by WPATH[8]. Breast binding appears common and is often undertaken without medical oversight[9]. Symptoms associated with breast binding include back and chest pain, shortness of breath and, although unusual, rib fracture[10]. Breast binding also deforms the breasts themselves, particularly the case for girls with larger breasts. The image in Figure 1 of Sood et al.[11] shows the deformation of the breasts of a 14-year-old girl due to binding. Her breasts are similar in appearance to those of an elderly woman but are not as deformed as descriptions I have heard from parents regarding the impact of breast binding on their daughter's anatomy.

Also anecdotally, it seems that the unattractive appearance of breasts that have been protractedly bound is also a motivation for seeking chest masculinisation surgery. The changes to breast structure and appearance caused by breast binding are not reversible.

Breast binding is analogous in some ways to the practice of breast ironing, a traditional practice in West Africa whereby the breasts of pubertal girls are flattened with hard objects such as a stone to discourage male sexual attention[12]. Breast ironing is considered to be a form of sex-based violence and child abuse[12], including by the United Nations[13]. The UK's Metropolitan Police notes that binders can also be used for breast ironing[14] raising the question of why breast binding as a part of a West African tradition is child abuse or self-harm but breast binding to support a transgender identification is apparently not? The impact of breast ironing on breast function has been little researched but reportedly is connected to ongoing breast pain, difficulties breastfeeding and low milk supply[15,16].

It is outside my area of expertise, however I would also note that 'genital tucking' is also considered to be a part of social transition for boys and that very young children even are being supported in this practice. This is also a physical intervention and also has potential for adverse consequences[8]. This should also be discussed in the Review.

I would encourage the authors to consider reconceptualising breast binding and genital tucking in the Review as a physical intervention rather than as part of social transition.

Detransition

The Review notes that individuals who desist in their transgender identification may experience regret as a result of gender identity-related medical interventions. Regret associated with chest masculinisation surgery is not mentioned but should be added since: 1) this is the most common surgical procedure for minors with gender dysphoria, 2) WPATH and other guidelines omit to recommend that the impact of this surgery on breastfeeding be discussed with those considering it, 3) proponents of this surgery regularly falsely state that this surgery can be reversed, 4) emergence of acute regret may occur many years after surgery as there may be decades between surgery and a woman giving birth and being unable to breastfeed[5].

Karleen Gribble PhD, BRurSc(Hons)

Adjunct Professor, School of Nursing and Midwifery

1. Stroumsa D, Roberts Elizabeth FS, Kinnear H, Harris LH. The power and limits of classification: 32-year-old man with abdominal pain. *New England Journal of Medicine.* 2019;380(20):1885-1888.
2. Whitley CT, Greene DN. Transgender man being evaluated for a kidney transplant. *Clinical Chemistry.* 2017;63(11):1680-1683.
3. Sullivan A, Murray Blackburn Mackenzie, Webb K. *Review of data, statistics and research on sex and gender*: University College London; 2025.
4. Munzer M, Jameson N, Harris A, Curran C, Dinsdale N, Gribble K. Sex and gender identity: data collection and language considerations for human research ethics committees and researchers. *Journal of Academic Ethics.* 2025.
5. Gribble KD, Bewley S, Dahlen HG. Breastfeeding grief after chest masculinisation mastectomy and detransition: a case report with lessons about unanticipated harm. *Frontiers in Global Women's Health.* 2023;4.
6. Victora CG, Bahl R, Barros AJD, et al. Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. *Lancet.* 2016;387(10017):475-490.
7. Gribble K. Letter to the Editor for 'The lactation and chestfeeding/breastfeeding information, care and support needs of trans and non-binary parents: An integrative literature review'. *New Zealand College of Midwives Journal.* 2024;60:246004
8. Coleman E, Radix AE, Bouman WP, et al. Standards of care for the health of transgender and gender diverse people, Version 8. *International Journal of Transgender Health.* 2022/08/19 2022;23(Supplementary 1):S1-S259.
9. Julian JM, Salvetti B, Held JI, Murray PM, Lara-Rojas L, Olson-Kennedy J. The impact of chest binding in transgender and gender diverse youth and young adults. *Journal of Adolescent Health.* 2021/06/01/ 2021;68(6):1129-1134.
10. Peitzmeier SM, Silberholz J, Gardner IH, Weinand J, Acevedo K. Time to first onset of chest binding–related symptoms in transgender youth. *Pediatrics.* 2021;147(3):e20200728.
11. Sood R, Jordan SW, Chen D, Chappell AG, Gangopadhyay N, Corcoran JF. Mastectomy and chest masculinization in transmasculine minors: a case series and analysis by ethical principles. *Ann Plast Surg.* 2021;86(2):142-145.
12. Falana TC. Breast ironing: a rape of the girl-child's personality integrity and sexual autonomy. *Social Sciences, Humanities and Education Journal (SHE Journal).* 2020;1(3):93-102.
13. UN Women. Breast ironing. Available at: https://www.endvawnow.org/en/articles/609-breast-ironing.html.
14. Metropolitan Police. Breast ironing (flattening). Available at: https://www.met.police.uk/advice/advice-and-information/caa/child-abuse/breast-ironing-flattening/
15. Simpson H. Raising awareness of breast ironing practices and prevention. *Practicing Midwife.* 2018;21(5):38-41.
16. Fotabong M, Obajimi G, Lawal T, Morhason-Bello I. Prevalence, awareness and adverse outcomes of breast ironing among Cameroonian women in Buea Health District. *Medical Journal of Zambia.* 2022;49(4).

Dr. Richard J. Santen

Peer Review of Treatment for Pediatric Gender Dysphoria

Dr. Richard J. Santen, Emeritus Professor of Endocrinology, University of Virginia School of Medicine

I have reviewed the DHHS document and find that the summary of data and detailed discussions reasonably reflect an overview of the information currently available and its interpretation. The "Umbrella Review" of multiple systematic reviews is particularly helpful as it covers an extensive volume of data and provides an assessment of the level of validity of each review. In examining the tables in detail, I believe that the overall assessment of these studies was scientifically sound. I verified that the criteria for such an overview were being met. My assessment also allows the conclusion that Chapter #7 also contains scientifically valid information. However, I believe that one area of the DHHS document has not received the emphasis that is essential. That is the issue whether gender-affirming hormone therapy (i.e., puberty blockers and cross-sex hormone therapy) is experimental or accepted practice. I have reviewed the concepts underlying the definition of experimental therapy in this critique and suggest that a specific section be added to address this. In my opinion, whether or not gender-affirming hormone therapy is experimental or not is the most important issue underlining all of the current controversy. Another important issue is the concept of "stacking" of the membership of committees developing clinical practice guidelines. I will first address these two issues and then I will then make specific comments about several additional, but lesser important issues.

Experimental therapy versus accepted practice: This issue requires thoughtful and in-depth discussion. As this document is not meant to make recommendations, I suggest that the topic be extensively discussed with attention to the pros and cons of this issue. As a basis for my comments, I have used the Gemini artificial intelligence (AI) platform to define the criteria to determine what is experimental therapy. Having read extensively about this topic and having conducted substantial clinical research in my career, I believe that this definition of experimental research is as complete and valid as I have found elsewhere. I quote this below and then use this definition to apply to comments in this manuscript.

Experimental therapy: Definition

Overview: Determining whether a therapy for a patient is considered experimental is a complex process with significant implications for patient care, ethics, and regulation. There is not a single universal definition, but it generally refers to treatments that are not yet recognized by the professional medical community as effective, safe, and proven for this specific condition for which they are being used. Considered below is a breakdown of the definition and key considerations.

Definition of experimental therapy: Therapy is to be typically considered experimental if:

Lack of established efficacy and safety: There is insufficient scientific evidence, for example from well-designed clinical trials, to definitively prove its effectiveness and safety for the intended use. This is often because it is a new, unknown or a rarely used intervention.

Deviation from standard of care: It does not align with the usual clinical practice supported by a consensus of medical practitioners for the specific condition.

Undergoing or awaiting research: It is currently being studied in clinical trials, or it has not yet undergone the necessary rigorous testing or to gain widespread acceptance.

Off-label use in certain context: While off label use, for example using an approved drug for a purpose not specifically approved by regulatory bodies, can sometimes be considered standard practice based on emerging evidence, it can also be considered experimental if the evidence for its new use is limited or speculative.

Not approved by regulatory bodies: In many countries, therapies are considered experimental until they receive approval from regulatory agencies like the U.S. Food and Drug Administration for specific indication.

Practical considerations

Monitoring and follow-up: Will the patient be monitored for safety and effectiveness? What resources are available for ongoing care and management of potential side effects?

Resource availability: Does the healthcare institution have the necessary expertise, equipment, and support staff to administer and manage the experimental therapy safely?

Cost: The financial burden of experimental treatments can be substantial for patients in healthcare systems.

Summary

In summary, classifying a therapy as experimental hinges on the level of robust scientific evidence for its safety and efficacy in the particular context. The decision to use a therapy requires a collaborative and transparent discussion between the patient, their family, and the medical team, ensuring comprehensive informed consent, rigorous ethical considerations, and adherence to relative regulatory frameworks.

Review of data from the text of the DHHS document about experimental therapy: In this paragraph, I have extracted specific statements regarding the experimental nature of gender-affirming hormone therapy. The Finnish and Swedish guidelines and the Cass Review

consider gender affirming care to be experimental. I will review the data supporting this conclusion. Finnish--- Page 145, line 6 …. "use of hormones should be limited to nationally overseen research or exceptional circumstances". Page 145 line 16…. "Following an SR, Finnish authorities concluded that the body of evidence supporting puberty blockers and cross -sex hormones for youth is inconclusive. Importantly, the guidelines explicitly state that "in the light of available evidence, gender reassignment of a minor is an experimental practice." With respect to Swedish guidelines, Page 147 , line 21…. "Medical and surgical interventions are subject to equally rigorous restrictions. Treatment with puberty blockers is confined to the context of clinical research. Until such research protocols receive ethics board approval, puberty blockers may be administered only in exceptional cases under the updated guidelines. Similarly, the use of cross-sex hormones (testosterone or estrogen), is permitted solely within research studies." In response to the Cass Review, page 149 line 13…., NHS England introduced major policy changes . "Puberty blockers are no longer routinely commissioned due to insufficient evidence regarding their long term safely and effectiveness. Instead, they will only be accessible through a structured clinical research trial which is currently being designed." Page 23, line22 …. "The reality is that we have no good evidence on the long-term outcomes of interventions to manage gender-related distress". As an unrelated comment, it is of interest that from 2014, puberty blockers moved from a research- only protocol to being available through routine clinical practice even though the evidence had not changed. (see page 52).

In marked contrast, the Clinical Practice Guidelines of the Endocrine Society and WPATH consider gender-affirming care, puberty blockers and cross-sex hormone therapy to be standards of care and supported by evidence. The discussion in these two guidelines highlights the fact that gender-affirming care and cross-sex hormone therapy have been used over the past 30 years and are practiced in multiple countries. Twenty-five scientific societies have approved of this approach.

One should note that the issue about the experimental nature of the gender-affirming care approach has only been raised in the past 5 years. The Swedish practitioners first raised concerns about this approach and the Finnish followed. After legal issues arose in the UK, the NHS commissioned the CASS report. These resulted in a high degree of controversy among stakeholders and recent emphasis on examining all of the issues involved.

The DHHS document, in my opinion, needs to specifically address the components of the definition of experimental medicine and how current studies relate to this definition. In the section below, I will examine each of the criteria for determining if a therapy is experimental and comment how current data are congruent with these criteria or conflicting.

1. Medical and scientific considerations:

<u>Existing evidence:</u> What preclinical data, early phase clinical trial results, or anecdotal evidence exist? Is there any indication of potential benefit?

The 2017 clinical practice guidelines of the Endocrine Society "suggest" (a recommendation with lesser strength than we recommend) the use of puberty blocks for gender-dysphoric youth. The level of evidence for this conclusion is considered "low" as defined by the GRADE method. WPATH states that "we recommend" various aspects of gender-affirming care but no recommendation states

the level of evidence supporting that recommendation. The ES and WPATH guidelines did not utilize commissioned systematic reviews to support their recommendations about the use of puberty blockers or cross-sex hormone therapy. Page 13 line 23…. In a strong dissenting opinion, the DHHS Umbrella Review states that the level of evidence supporting this form of therapy is very low. In my opinion, this "umbrella analysis" appears to be more credible than the ES and WPATH conclusions.

Deviation from standard of care: According to the ES and WPATH guidelines, the use of puberty blockers and cross-hormone therapy do not deviate from standard of care. The Swedish, Finnish, and UK documents do not agree and state that there is currently no generally accepted standard of care. Accordingly, there is no agreement on this issue.

Undergoing or awaiting research: Gender-affirming hormonal care is currently being studied in clinical trials and all organizations agree that data are needed about the long-term effects of puberty blockers and cross-sex hormone therapy on brain development, fertility, sexual function, bone health and cardiovascular disease. A key additional question to be addressed by research is whether adolescents not treated with puberty blockers or cross-hormone therapy will change their mind about gender non-conformity as they grow older. Two insurance studies, one in Germany and one in the USA suggest that 50-75 % will change their minds as adults but using insurance data is considered by some experts to be scientifically unsound methodology.

Off-label use in certain context: Off label use, for example using an approved drug for a purpose not specifically approved by regulatory bodies, can sometimes be considered standard practice.

This issue is most pertinent for treatments used in pediatric patients where clinical trials are often difficult to accomplish. Pediatricians comment that it is appropriate to utilize drugs which are off label for pediatric patients but approved for adults. Off-label use is legally acceptable in the USA but not in some other countries.

Not approved by regulatory bodies: In many countries, therapies are considered experimental until they receive approval from regulatory agencies like the U.S. Food and Drug Administration for specific indications.

In the countries requiring regulatory body approval, use of puberty blockers and cross-sex hormone therapy would be considered experimental as for example in the UK.

Practical considerations:

Monitoring and follow-up: The DHHS document comments that the strict standards espoused by the Dutch protocol are not being followed now that gender-affirming hormone therapeutic approaches are more commonly used globally. Anecdotal experiences shared with me by pediatric endocrinologists indicate the nurse practitioners in the USA are initiating testosterone therapy for adolescent girls without adequate training for this and without appropriate monitoring and follow up.

Resource availability: Does the healthcare institution have the necessary expertise, equipment, and support staff to administer and manage the experimental therapy safely?

The leading institutions managing adolescents with gender-dysphoria do have the necessary components for high quality care of these patients. However, individual practitioners without adequate training and resources are now more commonly managing patients with gender-dysphoria.

Pros and cons of defining gender-affirming hormonal therapy as experimental:

The DHHS document was designed to evaluate evidence and not establish guidelines. My suggestion as a reviewer is that the pros and cons of this issue be discussed. My analysis indicates that there is currently no agreement whether gender-affirming hormonal therapy is experimental or standard practice. A strong case can be made that it is experimental but this conclusion can reasonably be disagreed upon based on the long-standing experience in clinical practice. The pros of considering it experimental are that initiating treatment will necessitate all of the high standards applied to research studies, namely: informed consent, discussion of known risks and benefits, rigorous monitoring, safety review board oversight, training requirements of the researchers, and long-term follow-up. The cons are that clinical research trials in this area are difficult, particularly RCTs, and that considering the approaches experimental will result in withholding benefit from many adolescents with gender dysphoria.

The second general issue to be discussed is Guideline Stacking: the Endocrine Society published a manuscript on the trustworthiness of guidelines in 2022 (see JCEM 107:129- 2138, 2022) cautioning about the practice of "stacking" of clinical practice guideline (CPG) writing committees. The concept of "stacking", its definition, and its role in guideline development needs to be stated in the DHHS document. Four criteria were proposed by the ES to ensure a trustworthy CPG. (1) to ensure a multidisciplinary CPG, including members with expertise relevant to the topic (2) to encourage panel diversity with factors such as internationality, gender, race/ethnic, and career stage (3) to avoid "stacking" and (4) to ensure adherence to the CGC's conflict of interest/duality of interest policy. "Stacking" was defined as "inappropriately restricting guideline development group membership to those with a particular point of view." The ES CPG in 2017 included 9 of it 10 members who cared for patients with gender-dysphoria and could be considered advocates for this approach. WPATH committee membership also was "stacked" with advocates, a fact confirmed by the legal depositions of Dr. Marci Bowers and Eli Coleman. The basic issue with "stacking" is the problem of intellectual conflicts of interest. Intellectual COIs are defined as "academic activities that create the potential for an attachment to a specific point of view that could unduly affect an individual's judgement about a specific recommendation". According to the Institute of Medicine of Medicine in the USA (now called the National Academy of Medicine) "A person whose work or professional group fundamentally is jeopardized, or enhanced, by a guideline is said to have an intellectual COI". A reasonable assessment of the guideline development committee members of the ES and WPATH would come to the conclusion that nearly all members had intellectual conflicts of interest and that "stacking" was present. A careful reading of both guidelines indicates that these intellectual conflicts of interest were never stated.

The Swedish and Finnish guideline groups did not appear to have intellectual conflicts of interest

nor “stacking”. The Finnish guidelines were written by a group called the “Council for Choices in Health Care in Finland”, a public body that monitors, defines and assesses the Finnish Public Health Services. This group is unlikely to be dominated by advocates for gender-affirming care. The Swedish Guidelines were written by its healthcare Authority, Socialstyrelsen, a group also unlikely to be dominated by advocates. When the Cass Review was being developed, the avoidance of advocates was explicitly stated and “stacking’ and intellectual bias was not an issue.

My assessment is that the guideline committees of ES/WPATH were “stacked” with members with an intellectual conflict of interest and the Finnish and Swedish were not. As the ES/WPATH guidelines conflicted with the Finnish/Swedish, the presence or absence of committee panel “stacking” may have resulted in the markedly conflicting recommendations. This “stacking “ issue should be explicitly stated in the DHHS report. As a side note, the DHHS report does not recommend the ES/WPATH clinical practice guidelines based on an assessment of the criteria for valid guidelines but it does recommend the Finnish and Swedish.

Minor comments:

Page 10 line 7…. The statement “Additionally_the natural history of pediatric gender dysphoria is poorly understood, though existing data suggests it will remit without intervention in most cases.” This statement is ambiguous and conflicts with later statements in the DHHS document. Ambiguous because pre-pubertal gender dysphoria is known to commonly resolve but there are no scientifically sound data on adolescents who have not experienced gender dysphoria in the prepubertal period. The data on insurance reports and decrease in gender dysphoria are generally not considered scientifically sound by experts. My recommendation would be to delete this sentence and cover this issue in more detail later in the document.

Page 119 The ranges of testosterone in women and estradiol in men are too high. For women, testosterone ranges should about 10 to 35 ng/100 ml and estradiol for men from 10 to 40 pg/ml. These values should be inserted.

Page 144. Line 14 on. The comment about anabolic steroid abuse in men should be deleted. The amounts of anabolic steroid that cause the symptoms described are very much higher than the amounts used as cross-sex hormone therapy.

The article by Dr Joanna Olson-Kennedy is now available online in a non-peer reviewed format. The conclusions from this should be cited with the caveat this it is not peer reviewed. Also the New England Journal manuscript (see page 104) which is the NIH funded study should highlight the differences in results between birth assigned males and females as an adjunct to the discussion of the Olson-Kennedy manuscript.

Dr. Jilles Smids

Section of Medical Ethics, Philosophy, and History of Medicine, Erasmus MC, Rotterdam, the Netherlands

Review Chapter 13 of the HHS report on Gender Dysphoria

This review analyses chapter 13 from the HHS report on gender dysphoria (GD), which deals with the ethics of pediatric gender medicine. I would like to start with a disclosure: I provided constructive critical feedback on an early version of chapter 13, and at a much later stage on several other chapters of the HHS report. Accordingly, the question of why one would cooperate with the production of a report commissioned by the Trump administration that had just characterized pediatric medical gender care as 'child mutilation' may, to a lesser extent, also be asked to me. My considerations were that a report would be produced for the HHS anyway, and that it was always better if a good quality analysis would be produced instead of a document written in the same style as the earlier executive order written by the Trump administration. The composition of the team of authors, as far as known by me at that point in time, gave me sufficient confidence that most likely they would write a balanced and evidence-based analysis. So I decided to provide feedback, hoping to help achieving the production of a report with these characteristics. Later in my review, I will provide some overall reflections as to which extent I think the report has succeeded in that respect.

Chapter 13 argues for the following main theses. First, the commonly held medical ethical principles of beneficence and non-maleficence require sufficient scientific evidence for a favorable risk/benefit profile to justify pediatric medical transition (PMT, as the report calls it). Second, recent attempts to justify PMT on the basis of respect for patient autonomy misconstrue this medical ethical principle, and constitute a radical departure from standard understandings in pediatric gender medicine which take PMT to be justified by its (purported) resulting mental health benefits. Third, the chapter ends with a research-ethical analysis of potential research into PMT that is skeptical of the justification for offering it even in the context of clinical trials.

A strong feature of chapter 13 is its extensive reference to authoritative texts dealing with medical ethics to establish the common understanding just mentioned that doctors ought to offer only treatments that are medically indicated, i.e. for which the benefits reasonably outweigh the harms, and that respect for patient autonomy means that patients have the right to refuse or consent to such treatments offered by the doctor; patient autonomy does not constitute a right to receive treatments on the basis of the patient's wish. The chapter cites the classic textbook by Beauchamp and Childress (2019), a report by the American Academy of Pediatrics' (AAP) Committee on Bioethics (Katz et al., 2016), and a report by the Institute of Medicine's Committee on Quality of Health Care in America (Institute of Medicine (US) Committee on Quality of Health Care in, 2001), and other sources. In that way, the chapter sets the stage for correcting a rather common habit in medical- ethical analyses of framing the ethics of PMT as an inherent tension between the principles of non- maleficence and beneficence on the one hand, and respect for patient autonomy on the other hand. Again, there is no such inherent tension, because when there is no reasonable evidence of a positive risk/benefit profile, patients do not have a claim on receiving PMT, and not offering PMT is not an infringement of their autonomy.

When it comes to the evaluation of PMT, the chapter makes a very strong cumulative argument for the conclusion that, given our current knowledge, a precautionary approach is most warranted:

> The natural history of pediatric GD is poorly understood and decades of research has shown that early-onset GD usually resolves without medical intervention. There is no compelling evidence that the same will not prove true in the case of adolescent-onset symptoms, and limited evidence suggesting it will. And in any case, it is widely acknowledged that clinicians are unable to distinguish patients whose GD will persist from those whose GD will resolve. Further, there are concerns about the role medicalization itself may play in contributing to the persistence of the conditions being treated, and less invasive and less risky interventions are available. Lastly, medical intervention has known and plausible harms, and decades of research conducted by leading academic institutions have failed to produce reliable evidence of medical benefit. (p225).

It is the mutually supportive nature of these individually already weighty considerations that makes this case against the routine offering of PMT so strong. In most medical-ethical analyses, our lack of knowledge of the natural history of GD is given nowhere near sufficient weight (e.g. Allen et al., n.d., p. 8), if it is mentioned at all. Yet, we do not know which percentage of adolescents would have outgrown their GD without PMT, or whose GD would sufficiently have decreased for adolescents no longer to desire PMT. It can be 5 %, but also 50 % or even 80%. Administering such invasive treatment as PMT that results in life-long dependency of medical care and has serious medical risks and harms, when there is such profound uncertainty whether the adolescent even needs it, is simply unacceptable. This lack of knowledge of the natural history applies to those with childhood onset GD (Baron & Dierckxsens, 2022; Byrne, 2024) and even more for those with adolescent onset GD (Kaltiala-Heino et al., 2018). For both categories, the potential for overdiagnosis and harmful overtreatment is very high.

Regarding the worry that puberty blockers lock adolescents into their GD, the very high percentage of them continuing from puberty blockers to cross-sex hormones, more than 95% (Brik et al., 2020; Carmichael et al., 2021) is a reason for grave concern in this respect. This is especially the case because there are plausible mechanisms for such lock-in effects (Cass, 2022): puberty suppression halts bodily and psychosexual development, while sexual and romantic experience may be instrumental in outgrowing GD (Steensma et al., 2011).

Finally, regarding the direct harm/benefit profile, chapter 13 benefits from being able to refer to other chapters of the HHS report that have dealt extensively with these. The umbrella review from chapter 5 concludes that there is nearly exclusively (very) low certainty evidence regarding the harms and benefits of puberty blockers and cross-sex hormones. As reported by Block for *The BMJ*: “Mark Helfand, professor of medicine at Oregon Health Sciences University, said that the overview of systematic reviews was the report’s strongest portion, although it failed to add anything new. “The systematic reviews have consistently found that the primary studies have serious limitations, leaving uncertainty about both benefits and harms”. While I do think that chapter 13 is among the strongest parts of the HHS report, this positive evaluation of the report’s fundament, the systematic reviews, by an independent expert is important.

After referring to chapter 5, chapter 13 provides a convincing further discussion arguing for an important asymmetry between the implications of this uncertainty for harms versus benefits, arguing against those who claim that in such situations of uncertainty patients, their parents and clinicians should decide together. In brief, systematic reviews tend to underestimate the harms, for various reasons helpfully

summarized in chapter 6. Moreover, from basic physiologic evidence, there is reason enough to be very careful despite absence of high certainty evidence of harm. For example, one needs only to consider how seriously infertility is taken in clinical practice to know that this risk is not viewed as a mere hypothetical (Stolk et al., 2023). In such situation, a precautionary approach is indeed required.

So far for the chapter's analysis resulting in the conclusion of an unfavorable risk/benefit profile. Regarding its second main goal, criticizing current attempts to come up with alternative justifications for PMT based on autonomously stated embodiment goals or the desire for gender euphoria, chapter 13 does a very good job as well. It is hard to overestimate how radical the departure is of these attempts from normal clinical practice in other domains of medicine and from established medical ethical understanding. For example, clinicians and researchers from the Amsterdam gender clinic, including one of the founders of the Dutch Approach, Annelou de Vries, surprisingly argue that scientifically established improvement in terms of mental health benefits and GD is not necessary to justify PMT (Oosthoek et al., 2024). It is highly significant that just at the moment that a wide consensus has emerged regarding the weakness of the scientific evidence base for PMT, proponents of the gender affirming care model begin moving the goalposts in a surprisingly open and explicit manner. Chapter 13 provides a rather thorough and comprehensive discussion of these new justifications. Along the way, they also interestingly criticize the 'gender incongruence' diagnosis from ICD-11, even though these criticisms remain rather brief and would need much further development to enable assessing their merits.

Regarding the chapter's third major theme, whether PMT while not qualifying for routine treatment should be provided in a strict research context instead. These research ethical questions are currently increasingly important, as there are several new initiatives for more cohort studies and even RCTs (both in the UK and the Nordic European countries). I am unsure to which extent I find their analysis convincing. In any case, their references to established research ethical principles are valid, e.g. that there should be a reasonable prediction of a possible positive benefit/harm balance. However, I don't think we are in a position to claim that predictions of possible overall benefit are unreasonable even for a small subset of GD adolescents currently undergoing PMT. In any case, research ethical analysis of research in pediatric gender medicine has just begun, and constitutes an urgent research priority. After all, better research is still an important way to overcome the current heated debates and impasse. In this sense, I can understand Helfland's disappointment with the HHS report (Block, 2025). Reflection on appropriate research methods in pediatric gender medicine is urgent (Cf. Van Breukelen, n.d.)

In addition to these three major themes that I singled out, I should point to the chapter's very insightful discussion on the role of regret in ethical evaluation of PMT. It convincingly discusses the reasons why regret, although important, is complicated as an outcome measure and not the key issue in determining the ethicality of PMT. And for those so far missing a discussion of the principle of justice, this is included as well, arguing that adolescents with GD should receive care of the same quality standards as youth with other conditions receive (Cf. Kingdon et al., 2025; Smids, 2025).

All in all, I think it is fair to say that chapter 13 provides one of the most comprehensive and thorough ethical analyses of current pediatric gender medicine. It's strongest feature probably is its focus on the most fundamental issue, grounded in a convincing appeal to the principles of beneficence and non-maleficence and explaining the appropriate role of respect for patient autonomy: to which extent does PMT have a favorable risk/benefit profile? The cumulative case that the chapter presents for the answer that there is currently no such case is very strong and convincing, and I expect that it will turn out to be hard to rebut.

In this respect, it is helpful to quote again from Block's reporting: "Jonathan D Moreno, professor emeritus at the University of Pennsylvania, who was a senior adviser to Barack Obama's bioethics commission, told *The BMJ* that the ethics section of the HHS report cited reputable bioethics texts and presented a "plausible" analysis. However, even if the risk-benefit ratio was unfavourable, he said, the question of "how this report will be used" was one of political philosophy "about the proper role of government in the practice of medicine," adding, "Typically, in this country, we have been restrained in that respect."" (Block, 2025). Indeed, it is very much preferable that professional bodies such as the AAP and WPATH would adhere to established norms and practices for guideline development. If they would have done so in the past, there would have been no need for the current HHS report on GD. Their refusal to change course at some point may understandably lead to governmental action (Gorin et al., 2025), but again, it is strongly preferable that the medical community itself ensures to provide clinically and ethically appropriate care for adolescents with GD.

Finally a few comments about the report as a whole. It is clearly not a neutral report in the sense of merely providing the relevant considerations for and against the current gender affirmative treatment model in the US. It decidedly argues against early medical intervention for GD in minors. However, it does so transparently on the basis of established principles of evidence based medicine, responsible clinical practice, and medical ethics, while covering the relevant literatures and dealing with all extant considerations presented in favor of the gender affirmative care model. Accordingly the most productive way to respond to the HHS report, especially after its authors might be revealed, is to engage directly with the report itself. I would say that if one thinks its authors are highly biased, it should be possible to point out where the reports engages in motivated reasoning, fails to do justice to the extant literature, or shows other problems. I myself do think the report in fact has a few problems. For example, chapter 11, while still providing valuable insights, is far more accusative than fitting for the type of report the HHS analysis aims to be, accusing even clinicians who have just become the target of legal procedures. Here the fundamental principle that one is innocent until proven guilty would have been better applied. Another issue that deserves much more careful reflection than now given in the report is the question as to the nature of transgender identities. While chapter 2 on the relevant language is important and provides essential insights, its skepticism regarding the *term* gender identity may easily be taken for a wholesale skepticism regarding the *experience* of gender incongruence and may come across as dismissive to the importance that gendered feelings have for trans persons.

Despite issues such as these, in my view, the report as a whole provides a comprehensive interdisciplinary and well-argued analysis of pediatric gender medicine. It is unique in its sort by providing and integrating so many relevant perspectives. Producing this report in such a short time span is a tour de force and a valuable service to all stakeholders, for which the authors are to be commended. Especially the ethics chapter reviewed here is a very welcome addition to existing systematic reviews, but also to, for example, the Cass review. I look forward to my fellow bioethicists engaging in a respectful and productive discussion of the report in the bioethics literature and elsewhere.

Cited Literature

Allen, L. R., Adams, N., Ashley, F., Dodd, C., Ehrensaft, D., Fraser, L., Garcia, M., Giordano, S., Green, J., Johnson, T., Penny, J., Rachlin, K., & Veale, J. (n.d.). Principlism and contemporary ethical considerations for providers of transgender health care. *International Journal of Transgender Health*, 0(0), 1–19. https://doi.org/10.1080/26895269.2024.2303462

Baron, T., & Dierckxsens, G. (2022). Two dilemmas for medical ethics in the treatment of gender dysphoria in youth. *Journal of Medical Ethics*, 48(9), 603–607. https://doi.org/10.1136/medethics-2021-107260

Beauchamp, T. L., & Childress, J. F. (2019). *Principles of Biomedical Ethics* (8th edition). Oxford University Press.

Block, J. (2025). US transgender care: Evidence for interventions is "very low," says review ordered by Trump. *BMJ*, 389, r884. https://doi.org/10.1136/bmj.r884

Brik, T., Vrouenraets, L. J. J. J., de Vries, M. C., & Hannema, S. E. (2020). Trajectories of Adolescents Treated with Gonadotropin-Releasing Hormone Analogues for Gender Dysphoria. *Archives of Sexual Behavior*, 49(7), 2611–2618. https://doi.org/10.1007/s10508-020-01660-8

Byrne, A. (2024). Another Myth of Persistence? *Archives of Sexual Behavior*, 53(10), 3705–3709. https://doi.org/10.1007/s10508-024-03005-1

Carmichael, P., Butler, G., Masic, U., Cole, T. J., Stavola, B. L. D., Davidson, S., Skageberg, E. M., Khadr, S., & Viner, R. M. (2021). Short-term outcomes of pubertal suppression in a selected cohort of 12 to 15 year old young people with persistent gender dysphoria in the UK. *PLOS ONE*, 16(2), e0243894. https://doi.org/10.1371/journal.pone.0243894

Cass, H. (2022). *Interim report – Cass Review*. https://cass.independent- review.uk/publications/interim-report/

Gorin, M., Smids, J., & Lantos, J. (2025). Toward Evidence-Based and Ethical Pediatric Gender Medicine. *JAMA*. https://doi.org/10.1001/jama.2024.28203

Institute of Medicine (US) Committee on Quality of Health Care in, I. of M. (US) C. on Q. of H. C. in. (2001). COMMITTEE ON QUALITY OF HEALTH CARE IN AMERICA. In *Crossing the Quality Chasm:*

A New Health System for the 21st Century. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK222278/

Kaltiala-Heino, R., Bergman, H., Työläjärvi, M., & Frisén, L. (2018). Gender dysphoria in adolescence: Current perspectives. *Adolescent Health, Medicine and Therapeutics*, 9, 31–41. https://doi.org/10.2147/AHMT.S135432

Katz, A. L., Webb, S. A., COMMITTEE ON BIOETHICS, Macauley, R. C., Mercurio, M. R., Moon, M. R., Okun, A. L., Opel, D. J., & Statter, M. B. (2016). Informed Consent in Decision-Making in Pediatric Practice. *Pediatrics*, 138(2), e20161485. https://doi.org/10.1542/peds.2016-1485

Kingdon, C., Stingelin-Giles, N., & Cass, H. (2025). The Cass Review; Distinguishing Fact from Fiction. The American Journal of Bioethics, 25(6), 5–10. https://doi.org/10.1080/15265161.2025.2504397

Oosthoek, E. D., Stanwich, S., Gerritse, K., Doyle, D. M., & de Vries, A. L. C. (2024). Gender-affirming medical treatment for adolescents: A critical reflection on "effective" treatment outcomes. *BMC Medical Ethics*, 25(1), 154. https://doi.org/10.1186/s12910-024-01143-8

Smids, J. (2025). Looking at Gender Affirming Care Through the Lens of Justice. The American Journal of Bioethics, 25(6), 84–87. https://doi.org/10.1080/15265161.2025.2497997

Steensma, T. D., Biemond, R., de Boer, F., & Cohen-Kettenis, P. T. (2011). Desisting and persisting gender dysphoria after childhood: A qualitative follow-up study. *Clinical Child Psychology and Psychiatry*, 16(4), 499–516. https://doi.org/10.1177/1359104510378303

Stolk, T. H. R., Asseler, J. D., Huirne, J. A. F., van den Boogaard, E., & van Mello, N. M. (2023). Desire for children and fertility preservation in transgender and gender-diverse people: A systematic review. *Best Practice & Research Clinical Obstetrics & Gynaecology*, 87, 102312. https://doi.org/10.1016/j.bpobgyn.2023.102312

Van Breukelen, G. J. P. (n.d.). How to improve research methodology in gender care: A non-binary choice. *European Journal of Developmental Psychology*, 0(0), 1–21. https://doi.org/10.1080/17405629.2025.2485221

Dr. Lane Strathearn

Treatment for Pediatric Gender Dysphoria: Review for Evidence and Best Practice

Department of Health and Human Services, May 1, 2025.

Peer Review by Lane Strathearn, MBBS, PhD
Professor of Pediatrics, Psychiatry, Neuroscience and Pharmacology, Psychological and Brain Sciences, University of Iowa.

Thank you for this opportunity to review “Treatment for Pediatric Gender Dysphoria: Review of Evidence and Best Practice”. I am a tenured Professor of Pediatrics, Psychiatry, Psychological and Brain Sciences, and Neuroscience and Pharmacology, at the University of Iowa. I am also the Director of the Division of Developmental and Behavioral Pediatrics, and Physician Director of the Center for Disabilities and Development (CDD). My NIH-funded research includes longitudinal studies of parents and infants, focusing on the effects of early experience and maltreatment on child development, as well as the neurobiology of mother-infant attachment. As co-director of an NIH-funded P50 Center, the Hawkeye Intellectual and Developmental Disabilities Research Center (Hawk-IDDRC), I also have a broad interest in the care of children with intellectual, developmental and behavioral conditions, including those with gender dysphoria.

In August 2024, as an editorial board member for the Journal of Pediatrics, I received approval from the editor to prepare a Commentary on the assessment and management of pediatric gender dysphoria, summarizing 8 linked systematic reviews commissioned for the U.K. Cass Review. At that time, there appeared to be strong support, both locally and across the U.S., for “gender affirming care”, but little if any acknowledgement of the limited evidence base. The Commentary aimed to highlight this discrepancy and was titled “What We Know and What We Don’t: Evaluating the Evidence for Gender-Affirming Care in Pediatrics”, raising many of the same concerns highlighted in this Review. Unfortunately, but not unexpectedly, the Commentary was not received favorably by peer reviewers, who used many of the arguments effectively countered in this publication. Despite submitting a comprehensive rebuttal of these arguments (attached) and resubmitting to the “Journal of Pediatrics: Clinical Practice”, as recommended by the Journal, the commentary was never published.

Overall, the current Review provides a comprehensive summary of the evidence base for many treatment practices in pediatric gender medicine, including social transition, puberty blockers, cross-sex hormones, surgery, and psychotherapy. It also provides a compelling historical context for the current U.S. medical care environment, including the impact of international guidelines, U.S. medical association responses, and information garnered from legal proceedings. The Review provides a strong focus on evidence- based medicine, outlining both the strengths and limitations, supplemented by indirect evidence from basic science and physiology to better understand mechanisms and the likely risk/benefit ratio of treatment. I

believe that this Review provides a valuable and much needed contribution to this important field of practice.

Below are specific minor comments referenced to sections in the text:

FOREWORD

P. 10, para 3: I think it is important to acknowledge that there is also insufficient evidence to clearly understand the "risk of potential harm" for some of these treatments. For example, the long-term outcomes (both risks and benefits) are uncertain for all treatment modalities, including psychosocial support, social transitioning, pubertal suppression, and/or masculinizing/feminizing hormone interventions. In one systematic review, the evidence is described as "inconsistent" for whether hormone treatments result in permanent adverse effects, such as infertility, height/growth restriction, or reductions in bone density (Taylor et al. *Arch Dis Child*. 2024). Nevertheless**,** the responsibility for medical practitioners to "first do no harm" means that the primary burden of evidence should be for the likelihood of benefit, especially when there is even a potential for harm. This is discussed in more detail in Chapter 13: Ethical Considerations.

EXECUTIVE SUMMARY

Part I: Background

P. 13, para 1: It should be acknowledged that many of these international recommendations include treatment which is limited to established research protocols.

PART I: BACKGROUND

2.1 Terminology in pediatric gender medicine

P. 36, Footnote 37: Typographic error "medicalization that…" should be "…medicalization than…"

CHAPTER 3 - HISTORY AND EVOLUTION OF ADULT AND PEDIATRIC GENDER MEDICINE

These chapters are somewhat based on conjecture and hearsay and may be vulnerable to bias.

CHAPTER 4 - INTERNATIONAL RETREAT FROM THE "GENDER-AFFIRMING" MODEL

4.1 The rise of the affirmative care model

P. 58: Fig 4.2 should include error bars to assess the variability of the mean scores. Are the scores normally distributed, to justify a mean BDI-II score?

PART II: EVIDENCE REVIEW

CHAPTER 5 - OVERVIEW OF SYSTEMATIC REVIEWS

5.7.5 Conclusion

P. 94, para 3: This section highlighted the lack of evidence for *all* GD treatment modalities, including psychotherapy.

CHAPTER 6 - LIMITATIONS OF SYSTEMATIC REVIEWS

P. 95-96: This sentence needs additional clarification: “It is well-established in adults that for the same drug, off-label uses are associated with considerably higher rates of adverse effects, especially when strong scientific evidence is lacking”. Why would the lack of “strong scientific evidence” increase the rate of adverse effects?

6.2.3 Chen et al., 2023

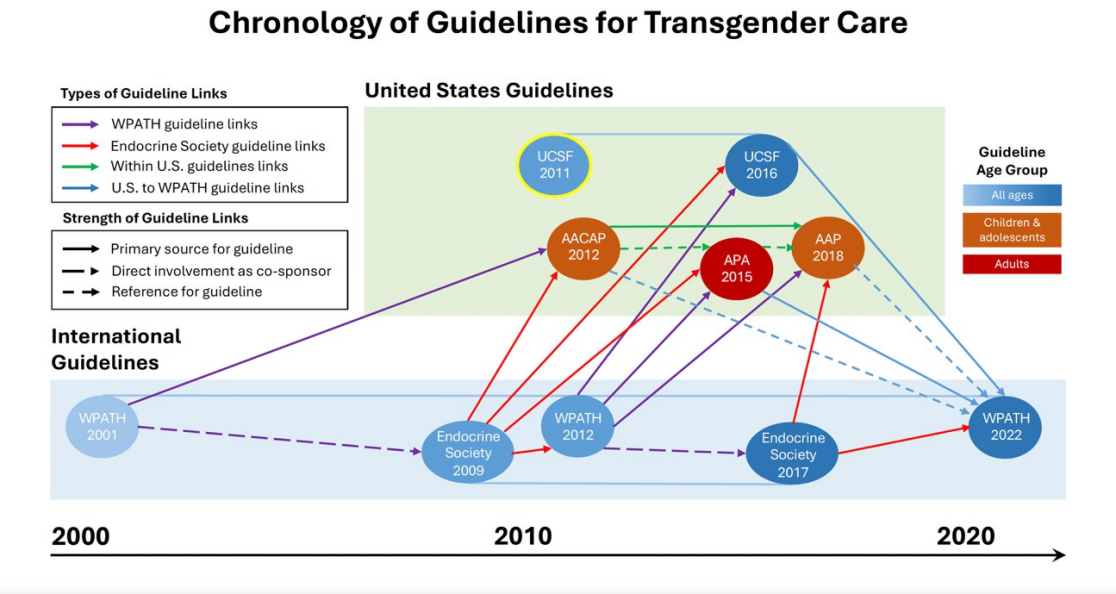
P. 102, para 3: It is unfair to compare the rate of suicide in those on CSH and GD with otherwise unaffected adolescents of a similar age. Although other studies are referenced in the footnotes, these rates should be compared in the actual text.

PART III: CLINICAL REALITIES

CHAPTER 9 - REVIEW OF INTERNATIONAL GUIDELINES

9.2.2 Interdependence of the existing guidelines and guidance documents

P. 136-7: Figures 9.2 and 9.3 are not referenced in the text. Figure 9.3 could be simplified to focus on guidelines within the US (see example below).



CHAPTER 11 - COLLAPSE OF MEDICAL SAFEGUARDING

11.3.1 Ambiguity in SOC-8

P. 162, para 1, ref 68: the embedded webpage link is not functional. The correct link is: https://www.washingtonpost.com/outlook/2021/11/24/trans-kids-therapy-psychologist/

11.3.2 SOC-8 guardrails abandoned

P. 164-171: Figures 10.1-4 are not referenced in the text. The “Boe v. Marshall” references do not come with information to access this information.

PART IV: ETHICS REVIEW

CHAPTER 13 - ETHICAL CONSIDERATIONS

13.2.3 Risk/benefit in pediatric medical transition

P. 219, para. 2: “Regarding the potential harms of psychotherapy for adolescents with GD, a systematic review of the evidence found no evidence of negative or adverse effects in any of the studies examined.” However, to be fair, and as noted when discussing potential harms of other medical treatment of GD, *no evidence* for harm does not equate with “no potential harm”. The studies reviewed were probably not specifically looking for or measuring potential harm.

BIBLIOGRAPHY

This link is not functional: Edwards-Leeper, L., & Anderson, E. (2021, November 24). The mental health establishment is failing trans kids. Washington Post. https://www.washingtonpost.com/outlook/2021/11/24/trans-kids-therapypsychologist/

Dr. Trudy Bekkering & Professor Patrik Vankrunkelsven

Patrik Vankrunkelsven, Director, Belgian Centre for Evidence-Based Medicine (Cebam)

Trudy Bekkering, Methodologist, Belgian Centre for Evidence-Based Medicine (Cebam)

Leuven, October 3, 2025

I. Introduction

The HHS Office of the Assistant Secretary for Health (OASH) invited us to peer review HHS's *Treatment for Pediatric Gender Dysphoria: Review of Evidence and Best Practices*.

Given our expertise in evidence-based medicine (EBM), we focused the review on the core of the report, namely the umbrella systematic review (SR) about the various treatments, chapter 5 (pp. 75 to 94).

Our report consists of three parts:

(II) Conclusions of peer review
(III) A summary of the content of chapter 5
(IV) Our evaluation using criteria from the Preferred Reporting Items for Overviews of Reviews (PRIOR) reporting checklist for umbrella reviews.

References:

(1) Pollock, M., Fernandes, R. M., Pieper, D., Tricco, A. C., Gates, M., Gates, A., & Hartling, L. (2019). Preferred reporting Items for overviews of reviews (PRIOR): A protocol for development of a reporting guideline for overviews of reviews of healthcare interventions. *Systematic Reviews*, 8(1), 335. https://doi.org/10.1186/s13643-019-1252-9

(2) Reporting guideline for overviews of reviews of healthcare interventions: development of the PRIOR statement. *BMJ* 2022; 378, e070849. https://doi.org/10.1136/bmj-2022-070849

II. Conclusions of Peer Review

1. Methods:

The use of an umbrella review is justified by the fact there are many SRs, most using the same studies. The review used robust methods:

- Followed Cochrane methods
- Searched Medline, Embase and PsychINFO from 2015 to 2025, complemented by ACCESSSS and Epistemonikos, also grey literature and Google Scholar. Full search string Medline Box 1.1 (appendix- separate file) constructs: gender dysphoria, youth, SR
- Screening title/abstract (tiab) and full text by 2 reviewers
- ROBIS to assess Risk of Bias (RoB)
- GRADE assessment
- Outcomes: gender dysphoria (GD), mental health and well-being, physiologic effects (e.g.,

suppression of sex hormones for puberty blockers (PBs)), need for or progression to further treatment, safety outcomes including side effects and adverse outcomes, and regret
- Table 2.1 (Appendix) - scope of SR, shows RoB of 4 domains and overall risk of bias in review for each review, and the interventions reviewed
- Section 2.2 (Appendix) - excluded SRs

2. General conclusions

We have no major remarks on the study design, nor on the conclusions. Minor remarks:

- The lack of rigorous reporting of conflict of interest (COI) by authors is the most important issue here, given the topic.
- A definition of an SR (to be included in the umbrella SR) would have been useful, but we found no issues on inclusion or exclusion of SRs.
- The registration of the protocol would have increased transparency, as would more details about how the results were summarized. However, the final results are described transparently and are easy to follow. There are also many tables with necessary and relevant information.
- No information was available on support, author information, and availability of data and other information (PRIOR items 24 to 27).

III. Summary of Chapter 5

Subject of the umbrella SR: What are the effects of social transition, PBs, CSHs, surgeries, and psychotherapy for youth with GD up to 26 years of age?

1. Results:

- 17 SR were included: 10 had low risk of bias overall, 7 had high risk of bias overall
- NB 2 NICE SR excluded because updates (by University of York) were published
- SR Baker 2021 excluded as this was in mature adults. However, as this study was cited by WPATH for a statement, it was checked for risk of bias separately (section 2.3 of Appendix 4) - high risk of bias due to limitations in ROBIS domains of "data collection and study appraisal" and "synthesis and findings."

(1) 5.2 Outcomes of social transition

- 2 SRs, both low risk of bias
- The results suggest that the **impact of social transition** on long-term GD, psychological outcomes and well-being, and future treatment decisions such as hormones or surgeries **remains poorly understood**. Evidence on regret associated with social transition is extremely limited. The certainty of evidence for these outcomes is very low.
- Most studies are cross sectional (not prospective, no comparison groups). It is unclear whether observed effects are causal.

(2) 5.3 Outcomes of puberty blockers (PB)

- 9 SRs, among which are 4 English reviews with low risk of bias
- The certainty of evidence is **very low** regarding the effect of PBs on GD (or gender incongruence),

improvement in mental health, and safety. There is **high certainty evidence** that PBs **exert physiological effects** (such as sex hormone suppression) and often cause infertility when followed by CSHs, depending on the patient's pubertal stage and sex. **Low certainty evidence** suggests that PBs may compromise bone health. A high proportion of youth proceed to CSHs.
- After PBs, though the certainty of evidence regarding any causal role PBs play in this progression is very low.

(3) 5.4 Outcomes of cross-sex hormones
- 8 SRs, among which are 4 English reviews with low risk of bias
- The certainty of evidence is **very low** regarding the effect on GD or incongruence, improvement in mental health, and safety metrics including fertility and bone health. There is **high certainty evidence that CSH exert physiological effects.**

(4) 5.5 Outcomes of surgery
- 3 SRs, of which 2 are with low risk of bias
- Most studies considered mastectomy only.
- There is **high certainty evidence** that mastectomy is associated with predictable surgical complications such as necrosis and scarring. The certainty of evidence is **very low** regarding the effect of surgery on GD or incongruence, improvement in mental health including suicidality and depression, and long-term outcomes such as sexual function, quality of life, and regret.
- Most studies are case series or small observational studies without comparison.

(5) 5.6 Outcomes of psychotherapy
- 5 SRs, of which 2 are with low risk of bias. The evidence on the effects of psychotherapy is limited. For mental health outcomes, the certainty of evidence was **very low**. However, no harms were reported.

2. Discussion
Certainty of evidence is very low. Not just because there are no randomized controlled trials (RCTs), as well designed observational studies would also be very helpful. There are no new or ongoing studies that would have an important impact. New studies are needed. New SRs are unlikely to yield novel insights.

IV. Checklist following PRIOR reporting

Section topic	Item No	Item	Achieved/location where item is reported
Title			
Title	1	Identify the report as an overview of reviews.	OK
Abstract			
Abstract	2	Provide a comprehensive and accurate summary of the purpose, methods, and results of the overview of reviews.	OK Executive summary report provides a good summary

Introduction			
Rationale	3	Describe the rationale for conducting the overview of reviews in the context of existing knowledge.	OK. In the present case, an overview of SRs was prepared because the field is already saturated with SRs, many of which evaluate the same studies. By assessing the quality of these SRs, an overview allows for a clearer understanding of the overall strength, consistency, and gaps in the evidence base.
Objectives	4	Provide an explicit statement of the objective(s) or question(s) addressed by the overview of reviews.	OK. Reviews the best available information regarding the risks, benefits, and uncertainties of interventions commonly used to
			address gender dysphoria (GD) in youth - interventions and outcomes reported
Methods			
Eligibility criteria	5a	Specify the inclusion and exclusion criteria for the overview of reviews. If supplemental primary studies were included, this should be stated, with a rationale.	OK. Appendix. Included if SR, reporting on youth below 26 yrs, assess the selected interventions (social transition, psychotherapy, puberty blockers (PBs), cross-sex hormones (CSH), or surgery.) - in addition, discussed results of review of Baker separately - review was excluded because it was on adults. Was discussed in umbrella review because this SR was cited to support statement 2.1 in WPATH SOC (see p 81 of report)
	5b	Specify the definition of “systematic review” as used in the inclusion criteria for the overview of reviews.	No definition of an SR (to be included in the umbrella SR) was found

Information sources	6	Specify all databases, registers, websites, organisations, reference lists, and other sources searched or consulted to identify systematic reviews and supplemental primary studies (if included). Specify the date when each source was last searched or consulted.	OK. searched Medline, Embase and PsychINFO from 2015 to 2025, complemented by ACCESSSS and Epistemonikos, also grey literature and Google Scholar. Full search string Medline Box 1.1 (appendix-separate file) constructs: gender dysphoria, youth, SR
Search strategy	7	Present the full search strategies for all databases, registers and websites, such that they could be reproduced. Describe any search filters and limits applied.	OK Box 1.1
Selection process	8a	Describe the methods used to decide whether a systematic review or supplemental primary study (if included) met the inclusion criteria of the overview of reviews.	OK. Two reviewers reviewed titles and abstracts and independently determined study eligibility. Once potentially eligible records were identified, a thorough review of full-text articles with a standardized and piloted screening form was performed. Reviewers resolved disagreement by discussion.
	8b	Describe how overlap in the populations, interventions,	OK.
		comparators, and/or outcomes of systematic reviews was identified and managed during study selection.	No description in methods was found, but the way the results are described is easy to follow. Evidence synthesis was based on outcomes from SRs published in English and assessed as having low risk of bias. Synthesis was organized based on intervention and outcomes For each outcome, this overview summarized the effect estimates and the certainty of evidence (confidence in the effect estimates, the quality of evidence). GRADE was used properly.

Data collection process	9a	Describe the methods used to collect data from reports.	OK. Extraction was done by 1 reviewer and checked by another.
	9b	If applicable, describe the methods used to identify and manage primary study overlap at the level of the comparison and outcome during data collection. For each outcome, specify the method used to illustrate and/or quantify the degree of primary study overlap across systematic reviews.	OK. Table 2.1 (appendix) provides an overview of included SRs, which interventions they are examining, and the RoB domains of ROBIS. Then, per intervention, an overview of included SRs.
	9c	If applicable, specify the methods used to manage discrepant data across systematic reviews during data collection.	NA
Data items	10	List and define all variables and outcomes for which data were sought. Describe any assumptions made and/or measures taken to identify and clarify missing or unclear information.	OK. Data extracted included review authors, research team, and research question answered; number and characteristics of included studies; study population; treatment; outcomes of interest; analysis and synthesis strategy; risk of bias assessment used for included studies.
Risk of bias assessment	11a	Describe the methods used to assess risk of bias or methodological quality of the included systematic reviews.	OK ROBIS was used.
	11b	Describe the methods used to collect data on (from the systematic reviews) and/or assess the risk of	OK
		bias of the primary studies included in the systematic reviews. Provide a justification for instances where flawed, incomplete, or missing assessments are identified but not reassessed.	Did not assess Risk of Bias (RoB) of primary studies in SR, but risk of bias was part of GRADE appraisal
	11c	Describe the methods used to assess the risk of bias of supplemental primary studies (if included).	NA

Synthesis methods	12a	Describe the methods used to summarise or synthesise results and provide a rationale for the choice(s).	OK. These methods are not clearly described, but reporting of the results is very structured: Per intervention -List the SRs, then how many RCTs were cited in the SRs. - List the SRs in English with a low risk of bias, then describe the number of included studies and study designs. - Then, list and summarize conclusions separately for each outcome For example: '5.3 Outcome 1. Gender dysphoria A total of four low risk of bias SRs assessed the impact of PBs on GD. Dopp 2024 included the most studies, though these included case reports and a qualitative study, which have limited value in estimating treatment effects. This systematic review narratively described that PBs lead to improved GD, without details on its methods for evidence synthesis. In contrast, the other three SRs reported no change in GD associated with PBs. The interpretation of these results requires caution regarding the certainty of evidence. All three SRs using the GRADE methodology to assess certainty of evidence concluded the certainty of evidence was very low. Taylor 2024a did not formally assess the certainty of evidence but found that "no high-quality studies using an appropriate design were

			identified, … no conclusions can be drawn," which is equivalent to very low certainty evidence. In summary, this overview concludes that the certainty of evidence is very low, and no conclusion could be drawn on the impact of PBs on GD.'
	12b	Describe any methods used to explore possible causes of heterogeneity among results.	NA
	12c	Describe any sensitivity analyses conducted to assess the robustness of the synthesised results.	NA
Reporting bias assessment	13	Describe the methods used to collect data on (from the systematic reviews) and/or assess the risk of bias due to missing results in a summary or synthesis (arising from reporting biases at the levels of the systematic reviews, primary studies, and supplemental primary studies, if included).	NA

Certainty assessment	14	Describe the methods used to collect data on (from the systematic reviews) and/or assess certainty (or confidence) in the body of evidence for an outcome.	OK. GRADE was used -this was copied from the SR or judged de novo in some cases (see below). This overview summarizes the GRADE ratings from the original SRs for the respective outcome wherever it is available. Nevertheless, two modifications were made: '1. Where a formal GRADE appraisal had not been performed by the SR, but expressions such as "we are very uncertain" or "no conclusions could be drawn" were used in the SR's conclusions, these were considered equivalent to a "very low quality" GRADE assessment. 2. Where SRs disagreed on GRADE assessment for the same outcome, this overview resolved the disagreement with de novo assessment following the GRADE methodology and reported the rationale.'
Results			
Systematic review and supplemental primary study selection	15a	Describe the results of the search and selection process, including the number of records screened, assessed for eligibility, and included in the overview of reviews, ideally with a flow diagram.	OK Fig 5.1 (report)
	15b	Provide a list of studies that might appear to meet the inclusion criteria, but were excluded, with the main reason for exclusion.	OK. Appendix 2.2
Characteristics of systematic reviews and supplemental primary studies	16	Cite each included systematic review and supplemental primary study (if included) and present its characteristics.	OK Table 2.1 (Appendix)

Primary study overlap	17	Describe the extent of primary study overlap across the included systematic reviews.	OK. Appendix - tables with evidence mapping per intervention - overview of primary studies and SRs, (e.g. Table 6.1 Evidence mapping of the SRs on CSHs and the primary studies that these SRs included)
Risk of bias in systematic reviews, primary studies, and supplemental primary studies	18a	Present assessments of risk of bias or methodological quality for each included systematic review.	OK. Table 2.1 - for each SR, reported per domain. Also, a separate description per SR (Chapter 4 Appendix) - summary of all included SRs
	18b	Present assessments (collected from systematic reviews or assessed anew) of the risk of bias of the primary studies included in the systematic reviews.	NA
	18c	Present assessments of the risk of bias of supplemental primary studies (if included).	NA
Summary or synthesis of results	19a	For all outcomes, summarise the evidence from the systematic reviews and supplemental primary studies (if included). If meta-analyses were done, present for each the summary estimate and its precision and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	OK. No meta-analyses were performed, but evidence tables were created that list the SRs for each intervention and outcome, including the number of studies, number of participants, review conclusions, and the certainty of
			evidence (plus reasons for downgrading the evidence). The last line contains their summary of the evidence plus certainty. For example, Table 5.2 Appendix Summary of evidence on gender dysphoria after PBs (p. 56)
	19b	If meta-analyses were done, present results of all investigations of possible causes of heterogeneity.	NA
	19c	If meta-analyses were done, present results of all sensitivity analyses conducted to assess the robustness of synthesised results.	NA

Reporting biases	20	Present assessments (collected from systematic reviews and/or assessed anew) of the risk of bias due to missing primary studies, analyses, or results in a summary or synthesis (arising from reporting biases at the levels of the systematic reviews, primary studies, and supplemental primary studies, if included) for each summary or synthesis assessed.	NA
Certainty of evidence	21	Present assessments (collected or assessed anew) of certainty (or confidence) in the body of evidence for each outcome.	OK. last row in the evidence tables
Discussion			
Discussion	22a	Summarise the main findings, including any discrepancies in findings across the included systematic reviews and supplemental primary studies (if included).	OK
	22b	Provide a general interpretation of the results in the context of other evidence.	OK
	22c	Discuss any limitations of the evidence from systematic reviews, their primary studies, and supplemental primary studies (if included) included in the overview of reviews. Discuss any limitations of the overview of reviews methods used.	OK (NB - SRs are at risk of reporting bias (that certain studies are not published) - Research in this field suffers from publication and reporting bias, but the extent of this problem is unclear
			- SRs cannot generate hypotheses, so important adverse events may remain undiscovered)
	22d	Discuss implications for practice, policy, and future research (both systematic reviews and primary research). Consider the relevance of the findings to the end users of the overview of reviews, e.g., healthcare providers, policymakers, patients, among others.	OK

Other information			
Registration and protocol	23a	Provide registration information for the overview of reviews, including register name and registration number, or state that the overview of reviews was not registered.	No information was found
	23b	Indicate where the overview of reviews protocol can be accessed, or state that a protocol was not prepared.	No information was found
	23c	Describe and explain any amendments to information provided at registration or in the protocol. Indicate the stage of the overview of reviews at which amendments were made.	No information was found
Support	24	Describe sources of financial or non-financial support for the overview of reviews, and the role of the funders or sponsors in the overview of reviews.	No information was found
Competing interests	25	Declare any competing interests of the overview of reviews' authors.	No information was found
Author information	26a	Provide contact information for the corresponding author.	No information was found
	26b	Describe the contributions of individual authors and identify the guarantor of the overview of reviews.	No information was found
Availability of data and other materials	27	Report which of the following are available, where they can be found, and under which conditions they may be accessed: template data collection forms; data collected from	No information was found
		included systematic reviews and supplemental primary studies; analytic code; any other materials used in the overview of reviews.	

- NA: Not applicable

Dr. Nadia Dowshen et al.

Dowshen, N., Baker, K., Garofalo, R., Chen, D., Inwards-Breland, D. J., Sequeira, G., ... & McNamara, M. (2025). A critical scientific appraisal of the Health and Human Services report on pediatric gender dysphoria. *Journal of Adolescent Health*, *77*(3), 342–345.

https://www.jahonline.org/article/S1054-139X(25)00246-0/fulltext (open access)

Professor G. Nic Rider et al.

Rider, G. N., Weideman, B. C., Ehrensaft, D., Choudhary, K., Connor, J. J., Feldman, J., ... & Berg, D. (2025). Scientific integrity and pediatric gender healthcare: Disputing the HHS Review. *Sexuality Research and Social Policy*, 1–6. https//:10.1007/s13178-025-01221-5

https://link.springer.com/article/10.1007/s13178-025-01221-5 (open access)

Replies

Reply to the American Psychiatric Association

We thank the American Psychiatric Association (APA) for its engagement with the Review. The APA has notified HHS that its peer review was authored by Dr. William M. Byne, M.D. and Dr. Jack Drescher, M.D., both distinguished figures in the field of gender medicine. We are grateful to Drs. Byne and Drescher for their criticisms and remarks.

The APA makes several substantive comments regarding the Review's (1) methodological rigor and study inclusion; (2) analysis of benefits and harms of pediatric medical transition (PMT); (3) engagement with the findings of the U.K.'s Cass Review; and (4) authorship and stakeholder involvement. Each will be addressed in turn below, followed by a summary.

1. ***Methodological Rigor***. The APA states that it cannot assess the Review's methodological rigor because of a lack of "methodological clarity" and "transparency," asserting this prevents verification or independent replication of the Review's findings. In particular, the APA claims that the Review "does not provide its search strategy," "fails to articulate how the studies are selected," does not explain "what criteria governed their inclusion or exclusion," provides no information on "how their quality was assessed," provides no information on the "analytical frameworks" used for the Review, and "did not … list the reviewed studies with full citations or digital object identifiers." The APA concludes that the Review's "claims fall short of the standard of methodological rigor that should be considered a prerequisite for policy guidance in clinical care."

   The APA ends its review with a list of 16 studies, presumably assuming that these studies had been overlooked and that they would potentially modify the Review's conclusions.

The Review's overview of systematic reviews (SRs) was peer-reviewed by two methodologists, Dr. Trudy Bekkering (Belgian Centre for Evidence-Based Medicine) and Professor Patrik Vankrunkelsven (Director, Belgian Centre for Evidence-Based Medicine). Bekkering and Vankrunkelsven used the PRIOR (Preferred Reporting Items

for Overviews of Reviews) checklist to assess the overview, and commended its robust methodology. They identified no major issues related to its design or conclusions, noting that “the final results are described transparently and are easy to follow” and that “there are also many tables with necessary and relevant information.”

Bekkering and Vankrunkelsven’s favorable peer review recognizes the methodological rigor of the Review’s approach. Contrary to the APA’s assertion, Appendix 4 provides a clear, transparent explanation of the Review’s search strategy/literature selection criteria (Section 1), exclusion criteria (Section 2.2), the key findings in the studies on which the Review relies (Sections 4–9) and does indeed list the reviewed studies with full citations and digital object identifiers (Section 11). This contradicts the APA’s peer review to such an extent that it suggests the reviewers failed to notice the references in the Review (including in the table of contents) to the 174-page Appendix 4. For example, Section 1.2.1 of the Review notes that “an overview of SRs was conducted, and its findings are presented in Chapter 5 and in Appendix 4.” In any case, the Review’s search strategy/literature selection criteria and exclusion criteria are also supplied in Chapter 5 of the Review.

The methodological framework used by the Review was the “Overview of Reviews” (also known as an “umbrella review”). This methodology is outlined in the *Cochrane Handbook for Systematic Reviews of Interventions* (Pollock et al., 2024). As Appendix 4 details, a total of 3,484 articles were screened according to prespecified criteria. Of these, 17 were identified as systematic reviews (SRs) that focused on the appropriate population and interventions. These SRs were assessed using the Risk of Bias in Systematic Reviews (ROBIS) assessment tool, and those at low risk of bias were included in the evidence synthesis and quality of evidence assessment, which followed the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) methodology.

The bibliography of Appendix 4 (Section 11) cites all the reviewed studies in the APA-7 format, including their DOIs.[3] The Review itself adheres to a modified APA-7 format, giving full citations and listing DOIs for publications not yet assigned to an issue.

Turning now to the APA's 16 "additional studies and reports for review and consideration," Table 1 below presents a study-by-study response:

Table 1: Review of studies supplied by APA

Study	Notes
1. Allen, L. R., Watson, L. B., Egan, A. M., & Moser, C. N. (2019). Well-being and suicidality among transgender youth after gender-affirming hormones. *Clinical Practice in Pediatric Psychology*, *7*(3), 302–311.	**Included in Appendix 4**. This study was identified by one or more of the low-risk-of-bias SRs included in the umbrella review. **Discussed in the Review.** This study was also discussed the main body of the Review (see footnote to Section 4.3.4).
2. Chen, D., Berona, J., Chan, Y.-M., Ehrensaft, D., Garofalo, R., Hidalgo, M. A., Rosenthal, S. M., Tishelman, A. C., & Olson-Kennedy, J. (2023). Psychosocial functioning in transgender youth after 2 years of hormones. *New England Journal of Medicine*, *388*(3), 240–250.	**Included in Appendix 4.** This study was identified by one or more of the low-risk-of-bias SRs included in the umbrella review. **Discussed in the Review.** This study is also discussed in the main body of the Review in great detail (see Section 6.2.3).
3. de Vries, A. L. C., Steensma, T. D., Doreleijers, T. A. H., & Cohen-Kettenis, P. T. (2011). Puberty suppression in adolescents with gender identity disorder: A prospective follow-up study. *Journal of Sexual Medicine*, *8*(8), 2276–2283.	**Included in Appendix 4.** This study was identified by one or more of the low-risk-of-bias SRs included in the umbrella review. **Discussed in the Review.** This study is also discussed in the main body of the Review in great detail (see Section 6.2.1).

[3] Several DOIs in the entire bibliography in Section 11 were inadvertently omitted and have been inserted.

4. de Vries, A. L. C., McGuire, J. K., Steensma, T. D., Wagenaar, E. C. F., Doreleijers, T. A. H., & Cohen-Kettenis, P. T. (2014). Young adult psychological outcome after puberty suppression and gender reassignment. *Pediatrics*, *134*(4), 696–704.	**Included in Appendix 4.** This study was identified by one or more of the low-risk-of-bias SRs included in the umbrella review. **Discussed in the Review.** This study is also discussed in the main body of the Review in great detail (see Section 6.2.1).
5. Green, A. E., DeChants, J. P., Price, M. N., & Davis, C. K. (2022). Association of gender-affirming hormone therapy with depression, thoughts of suicide, and attempted suicide among transgender and nonbinary youth. *Journal of Adolescent Health*, *70*(4), 643–649.	**Included in Appendix 4.** This study was identified by one or more of the low-risk-of-bias SRs included in the umbrella review. **Discussed in the Review.** This study is also referred to in a footnote in the main body of the Review (see Section 3.6).
6. Hughto, J. M. W., Gunn, H. A., Rood, B. A., & Pantalone, D. W. (2020). Social and medical gender affirmation experiences are inversely associated with mental health problems in a US nonprobability sample of transgender adults. *Archives of Sexual Behavior*, *49*(7), 2635–2647.	**Not included in Appendix 4.** Excluded from the analysis as the study population comprised mature adults, representing a different population than the Review's target population of youth.
7. LaFleur J., Heath L., Gonzalez V., et al. Gender-affirming medical treatments for pediatric patients with gender dysphoria. A report of the University of Utah College of Pharmacy Drug Regimen Review Center (DRRC). Salt Lake City, UT: University of Utah. 2024; https://le.utah.gov/AgencyRP/reportingDetail.jsp?rid=636	**Not included in Appendix 4.** This study postdates the publication of the Review. It is a scoping review that would not have met the umbrella review's inclusion criteria, as outlined in Appendix 4. Although presented as a systematic review, the study did not perform a formal synthesis of its evidence and so could not then appraise the certainty of that evidence, a necessary criterion for a systematic review.
8. Murad, M. H., Elamin, M. B., Garcia, M. Z., Mullan, R. J., Murad, A., Erwin, P. J., & Montori, V. M. (2010). Hormonal	**Not included in Appendix 4.** The study population comprised mature adults, representing a different

therapy and sex reassignment: A systematic review and meta-analysis of quality of life and psychosocial outcomes. *Clinical Endocrinology*, *72*(2), 214–231.	population than the Review's target population of youth.
9. Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.	**Not included in Appendix 4.** This publication is not about youth gender dysphoria (GD) or PMT. The Review does discuss how major medical and mental health associations (MMHAs) may be subject to confirmation bias (Section 12.1).
10. Olsavsky, A. L., Grannis, C., Bricker, J., Chelvakumar, G., Indyk, J. A., Leibowitz, S. F., Mattson, W. I., Nelson, E. E., Stanek, C. J., & Nahata, L. (2023). Associations among gender-affirming hormonal interventions, social support, and transgender adolescents' mental health. *Journal of Adolescent Health*, *72*(6), 860–868.	**Included in Appendix 4.** This study was identified by one or more of the low-risk-of-bias SRs included in the umbrella review.
11. Rosenthal, S. M. (2021). Challenges in the care of transgender and gender-diverse youth: An endocrinologist's view. *Nature Reviews Endocrinology*, *17*(10), 581–591.	**Not included in Appendix 4.** The publication did not report any original research. **Discussed in the Review.** This article is discussed in the main body of the Review (Section 4.3.2).
12. Taylor, J., Mitchell, A., Hall, R., Heathcote, C., Langton, T., Fraser, L., & Hewitt, C. E. (2024). Interventions to suppress puberty in adolescents experiencing gender dysphoria or incongruence: A systematic review. *Archives of Disease in Childhood*.	**Included in Appendix 4.** This SR's findings were deemed to be at low risk of bias and contributed to the evidence synthesis. **Discussed in the Review.** This study is also discussed the main body of the Review (see Sections 4.2.1, 5.1, 5.3).

13. Tordoff, D. M., Wanta, J. W., Collin, A., Stepney, C., & Inwards-Breland, D. J. (2022). Mental health outcomes in transgender and nonbinary youths receiving gender-affirming care. *JAMA Network Open*, *5*(2), e220978.	**Included in Appendix 4.** This study was identified by one or more of the low-risk-of-bias SRs included in the umbrella review. **Discussed in the Review.** This study is also discussed in the main body of the Review in great detail (see Section 6.2.2).
14. Turban, J. L., King, D., Carswell, J. M., & Keuroghlian, A. S. (2020). Pubertal suppression for transgender youth and risk of suicidal ideation. *Pediatrics*, *145*(2), e20191725.	**Included in Appendix 4.** This study was identified by one or more of the low-risk-of-bias SRs included in the umbrella review. **Discussed in the Review.** This study is also discussed in a footnote in the main body of the Review (see Section 4.3.4).
15. Turban, J. L., King, D., Kobe, J., Reisner, S. L., & Keuroghlian, A. S. (2022). Access to gender-affirming hormones during adolescence and mental health outcomes among transgender adults. *PLOS ONE*, *17*(1), e0261039.	**Included in Appendix 4.** This study was identified by one or more of the low-risk-of-bias SRs included in the umbrella review. Although the study was not discussed in the main body of the Review, the significant limitations of the study's data source (a 2015 online survey) are discussed (see Section 4.3.4).
16. van der Miesen, A. I. R., Steensma, T. D., de Vries, A. L. C., Bos, H., & Popma, A. (2020). Psychological functioning in transgender adolescents before and after gender-affirmative care compared with cisgender general population peers. *Journal of Adolescent Health*, *66*(6), 699–704.	**Included in Appendix 4.** This study was identified by one or more of the low-risk-of-bias SRs included in the umbrella review.

In brief:

- 12 of the 16 studies are in fact discussed in the Review and/or Appendix 4.
- Three of the remaining four studies do not pertain to youth outcomes or do not pertain to youth gender medicine at all.

- Only one study is new and potentially relevant: LaFleur et al. (2024), cited here as University of Utah College of Pharmacy, Drug Regimen Review Center (2025).[4] This report, commissioned by lawmakers in Utah (henceforth the "Utah Review"), was not publicly released until end of May 2025 (after the HHS Review's publication). Having assessed the study using ROBIS (Whiting et al., 2016), we find that the Utah Review would not have met the criteria for inclusion in the evidence synthesis. It is not a systematic review, because it failed to meet two key requirements of a systematic review (a formal evidence synthesis and an assessment of evidence certainty). See Appendix 4, Section 2.4.

It is not advisable to assess evidence simply by looking at the conclusions of individual studies, because not all studies are equally reliable. The cornerstone of evidence-based medicine is a *systematic review* of the evidence, which involves a search for studies using prespecified criteria, an assessment of the individual studies for risk of bias, and a determination of the quality (certainty) of the entire body of evidence for each key outcome (Guyatt et al., 2015). The Review's overview of systematic reviews adheres to this methodology.

The APA's challenges to the Review's methodological rigor are accordingly unfounded.

2. ***Analysis of benefits and harms of pediatric medical transition (PMT)***. The APA recognizes the Review's clarity concerning "the potential harms of intervening medically" but criticizes the Review for not applying "any kind of rational scrutiny to potential harms that have been associated with withholding intervention, including higher rates of depression, anxiety, suicidality, and social withdrawal."

The Review's analysis of the potential benefits and harms of PMT consists of (1) an overview of systematic reviews (Chapter 5 and Appendix 4); and (2) evidence from basic science and physiology (Chapter 7).

The overview of systematic reviews of interventions considered all relevant published literature regarding PMT, including studies that compared the outcomes for populations that received PMT with those that did not. The evidence synthesis found there was no

[4] This is also the citation given in the revised HHS Review.

credible evidence of benefits of PMT compared with no PMT in the outcomes referenced by the APA (depression, anxiety, suicidality)—and, by extension, found no credible evidence of harms from not providing PMT.

The basic science and physiology analysis assumed that endogenous puberty is not pathological, but a normal process of sexual development through which a child matures into an adult. Disrupting this process has the potential to result in physical harms. Therefore, the basic science and physiology analysis could only yield an assessment of the harms of interrupting a normal physiological process.

Contrary to the APA's assertions, then, the Review does engage in "rational scrutiny" of the benefits and harms of providing or withholding PMT.

3. ***Engagement with the findings of the U.K.'s Cass Review****.* The APA faults the Review for "draw[ing] heavily from the Cass Review which itself has been criticized by experts for its methodological flaws and biases." The APA also criticizes the Review for its failure to "take into consideration conclusions of the Cass Review that do not support the [Review's] outcome."

The APA cites two sources as "expert criticism" of the Cass Review. One is a non-peer-reviewed online essay whose authorship is commonly but erroneously attributed to Yale University (McNamara et al., 2024). The other is a peer-reviewed article (Noone et al., 2025) that primarily critiques the University of York systematic reviews (one of the main sources of evidence commissioned for the Cass Review) and also comments on the Cass Review itself. At least three papers to date have contested the central claims made by McNamara et al. (2024) (Cheung et al., 2025; Kingdon et al., 2025; McDeavitt et al., 2025), with the first and third of these papers also having commented on Noone et al. (2025).[5]

Like all scientific publications, the Cass Review has limitations. Further, disagreement is common in science, and debate should be welcomed. However, current debates surrounding the Cass Review are based largely on demonstrable mischaracterizations

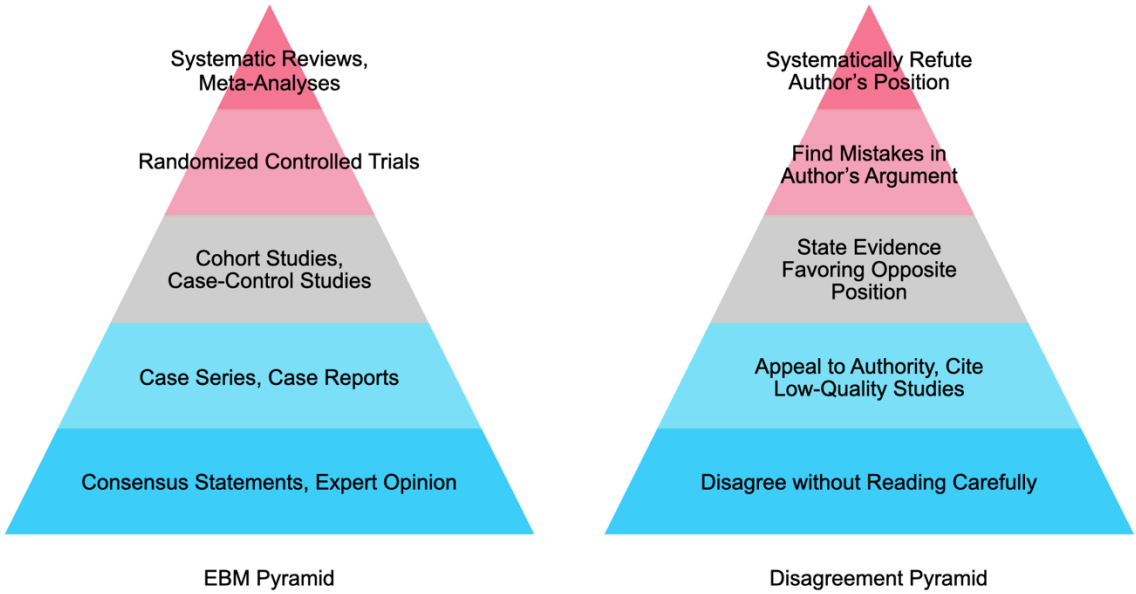
[5] The criticisms were specifically of a preprint, but the published paper does not significantly differ.

and simple errors of fact and appear to be part of a scientific misinformation campaign (Kingdon et al., 2025).

Cheung et al. (2025) criticize McNamara et al. (2024) for mistaking the Cass Review for a clinical practice guideline (CPG). Independent reviews are a U.K.-specific process deployed when an area of medicine begins to operate in a way that jeopardizes patient safety or compromises care quality. Independent reviews adhere to the "terms of reference" set out by the commissioning body rather than the standards for CPG development.

We agree with Baxendale (2025) that debates about the efficacy of medical interventions should be settled by parties using evidence from studies at the top of the evidence-based medicine (EBM) pyramid (see Appendix 3 of the Review), and addressing their opponents' arguments in good faith—rather than relying on authority, citing irrelevant studies, or reading perfunctorily. See Figure 1 below.

Figure 1: EBM and disagreement pyramids[6]



[6] Figure inspired by Baxendale (2025).

We now turn to the APA's second allegation, that the Review selectively picks from the Cass Review, omitting the "conclusion" that "for some, the best outcome will be transition." The APA asserts that this "conclusion" is inconsistent with the findings of the HHS Review.

The APA's claim that this was a "conclusion" of the Cass Review is another example of the ongoing misinformation campaign against it.[7] When properly understood in context, and given other critical observations made in the Cass Review, the quote is best understood as one consideration in a more nuanced line of clinical reasoning. Crucially, the quote does not support the APA's insinuation that the Cass Review is supportive of PMT in clinical settings.

For reference, the full quotation alluded to by the APA is:

> For some, the best outcome will be transition, whereas others may resolve their distress in other ways. Some may transition and then de/retransition and/or experience regret. The NHS needs to care for all those seeking support. (Cass, 2025, p. 21)

The Cass Review found the evidence for benefit to be "weak," that "clinicians ... are unable to determine with any certainty which children and young people will go on to have an enduring trans identity,"[8] and that "a diagnosis of gender dysphoria … is not reliably predictive of whether that young person will have longstanding gender incongruence in the future, or whether medical intervention will be the best option for them."[9]

Similarly, the HHS Review finds that the evidence underpinning the alleged benefits of medical interventions is very uncertain; that clinicians are unable to distinguish between minors for whom the (alleged) benefits would outweigh the harms; and that the field in its current form uses a model in which the child's wishes determine the course of

[7] See Kingdon et al. (2025), cited above, as well as Cheung et al. (2025), which describes how some criticism of the Cass Review—for instance, in the McNamara et al. (2024) online essay—appeared to have been written for the purpose of influencing court cases, as opposed to advancing scientific understanding. Baxendale (2025) also critiques the style of argument used in McNamara et al. (2024).
[8] Cass (2025, p. 22).
[9] Cass (2025, p. 34).

treatment. In addition to these overlapping areas of inquiry with Cass, the HHS Review conducted a comprehensive analysis of harms, concluding that some harms are physiologically certain and others plausible, as well as an ethical analysis that incorporates these findings along with well-established principles in medical ethics.

Any reasonable interpretation of the Cass Review's statement that "for some, the best outcome will be transition," must grapple with its findings about lack of evidence for benefit and deep uncertainties about diagnosis. Unfortunately, the APA fails to do so. Even granting for the sake of argument that "for some the best outcome will be transition" it would not follow that prescribing PMT interventions in clinical settings is ethically permissible because, as the Cass Review acknowledges, there is no way for clinicians to distinguish between patients whose gender dysphoria will persist into adulthood and those who will come to terms with their bodies. As we say in our response to Dr. Jilles Smids, "As with studies of any clinical intervention, the fact that studies to date do not find strong evidence that PMT improves health outcomes does not, in principle, rule out the possibility that some subpopulation of subjects benefits from the interventions while others are harmed by them. However, to date no subpopulation has been shown to benefit." Given the unfavorable risk/benefit profile and the inherent difficulties in diagnosis, ethical considerations support prioritizing less aggressive therapeutic alternatives.

Moreover, the APA's reading of this sentence from the Cass Review is not supported by subsequent administrative decisions in the U.K., where PMT interventions are now restricted to research settings.[10] The U.K.'s National Health Service (NHS) has accepted all of the Cass Review's recommendations for implementation. Puberty blockers for pediatric GD have been permanently banned in the U.K., while prescribing of cross-sex hormones for youth under 18 has been sharply curtailed. According to media reports, no new cases of cross-sex hormones for youth under 18 have been initiated through the NHS (Spencer, 2025).

[10] As of this publication, no such research has been approved.

Contrary to the APA's assertions, then, the Review appropriately engaged with the Cass Review's overall findings.

4. ***Authorship and stakeholder involvement***. With respect to the initial non-disclosure of the contributors to the Review, the APA says that this prevents others from being able to "assess the expertise of contributors, evaluate their qualifications in relevant fields, and identify potential conflicts of interest or ideological commitments." This, according to the APA, threatens "the integrity of scientific and policy analysis" of the Review. The APA also asserts that the perspectives of "key stakeholders—namely, transgender individuals [and] their families" may not have been adequately considered, to the detriment of the Review's conclusions.

We agree that it is vital to preserve the "integrity of scientific and policy analysis" of the Review. All scientific publications are at risk of bias or the perception of bias; publications in the contentious space of youth gender medicine are especially susceptible to both, due to the highly politicized nature of the field. To minimize bias, HHS took the following steps:

- First, with full recognition of the highly politicized climate that surrounds the care for gender-dysphoric youth, HHS deliberately sought expert contributors from a wide range of political positions, including those not politically aligned with the administration commissioning the Review.

- Second, the Review's evidence syntheses followed a well-established, rigorous, and reproducible methodology (Pollock et al, 2024). This ensures that if the same project of overviewing systematic reviews was conducted again under the same conditions, an independent team would arrive at comparable results and draw similar conclusions of very low certainty evidence for the benefits of PMT.

- Third, HHS conducted an external peer review of the findings, seeking input from organizations and individual experts with a diverse set of perspectives and positions on PMT. In addition to the APA, HHS sought input from the American Academy of Pediatrics (AAP) and the Endocrine Society (ES), as well as from

> individuals who are recognized as experts in this field (including those who would be expected to be critical of the Review or some important aspects of it).

Importantly, the Review does not make specific policy recommendations but instead attempts to provide the best available information to guide decision makers. We agree with the APA that stakeholder involvement should be a part of CPG development. The HHS Review is not a CPG. It does, however, draw on the Cass Review, which commissioned qualitative research to characterize experiences of patients, parents, and clinicians, and conducted interviews with over 1,000 individuals and organizations. Also, according to a 2024 systematic guideline appraisal, CPGs which recommend standard-of-care psychotherapy (not PMT) received high scores for stakeholder involvement, compared with the World Professional Association for Transgender Health (WPATH) and ES CPGs, which did not (Taylor et al., 2024a, 2024b).

Given the highly polarized nature of the topic, contributors' names were withheld during the peer-review process so that reviewers could focus on the content of the review, rather than on the individual contributors themselves.[11] This is an established practice in scientific review, designed to reduce reviewer bias and ensure impartial focus on substance. The scientific integrity of any document, including the Review itself, is best assessed through its content.

Another peer reviewer (Dr. Jilles Smids) has noted that allegations of bias should be accompanied by examples of where the Review "engages in motivated reasoning, fails to do justice to the extant literature, or shows other problems." The APA has provided no such examples.

Summary

The APA's central criticism concerns the Review's methodology:

> Our conclusions are that while the HHS Report purports to be a thorough, evidence-based assessment of gender-affirming care for transgender youth, its underlying methodology lacks sufficient transparency and clarity for its findings to

[11] As Dr. Jilles Smids notes in his peer review, he provided feedback on earlier versions of parts of the Review; accordingly, he had knowledge of some contributor identities prior to the May 1 publication.

> be taken at face value. Key elements including literature selection criteria, analytical frameworks, and justification for excluding other studies, and key findings in studies on which the Report relies, are either underexplained or absent. As a result, the Report's claims fall short of the standard of methodological rigor that should be considered a prerequisite for policy guidance in clinical care.

This unfounded criticism may have resulted from a failure to read core parts of the Review (principally Chapter 5, which summarizes the umbrella review's methodology). The APA's other criticisms are similarly unfounded. However, we have added a discussion of the Utah Review (one of the sources cited by the APA, whose publication postdates that of the HHS Review) in Chapter 5 (Section 5.7.3) and Appendix 4 (Section 2.4).

Finally, we appreciate that the APA, a leading mental health organization, did not mischaracterize the HHS Review's extensive discussion of psychotherapy for youth with GD (Chapter 14) as promoting "conversion therapy." We are encouraged that the organization chose not to participate in the denigration of psychotherapy in this context. Psychotherapy is demonstrably evidence-based for other types of psychological distress in children and adolescents, and has increasingly been recommended as an ethical, comparatively non-invasive treatment option that does not carry the significant risks associated with PMT.

Reply to Bester

We thank Dr. Johan Bester for his thorough and helpful review. Bester states that "the main findings and conclusions of the review are correct," and makes several suggestions for "minor improvements." We respond to some of these here.

1. Bester notes it should not be assumed that psychotherapy is a beneficial treatment for pediatric gender dysphoria (GD) merely because it is an effective treatment for some other mental health conditions, and he cautions against strongly endorsing psychotherapy given the weak evidence base. He recommends the Review "make a stronger suggestion for further studies of psychotherapy" as an intervention for pediatric GD.

The Review reports that the evidence for benefit of psychotherapeutic approaches for mental health conditions that often accompany GD (for instance, depression) is stronger than the evidence for their effects on GD itself and, therefore, that psychotherapy is a promising treatment for the former conditions in patients presenting with GD. The incidence of co-occurring mental health conditions is very high in this population and there is no good evidence that pediatric medical transition (PMT) is a safe or effective intervention for these indications, just as there is no good evidence that PMT is safe or effective in treating pediatric GD itself. However, crucially, the Review also points out that psychotherapy carries lower risks than PMT. Research indicating that psychotherapy is an effective treatment for a wide range of psychosocial problems, combined with its carrying a lower risk than PMT, suggests that the risk/benefit profile of psychotherapy for treatment of pediatric GD is favorable when compared to more medically aggressive alternatives such as puberty blockers, cross-sex hormones, and surgeries.

The Review's overview of systematic reviews (Appendix 4) concludes that the evidence for benefit of psychotherapeutic interventions for the treatment of pediatric GD is uncertain. This likely is due at least in part to the dearth of primary studies examining the effects of psychotherapy and the emphasis on puberty blockers, hormones, and surgery, which some U.S. clinicians and researchers incorrectly regard as comprising the standard of care. We agree with Bester that robust research on psychotherapeutic

approaches is needed. We encourage researchers to conduct such research and to incorporate their findings when developing trustworthy, evidence-based clinical practice guidelines for the management of pediatric GD.

2. While Bester agrees with the Review's insistence that a positive risk/benefit profile is a necessary condition for ethical prescribing, he recommends the Review include a more detailed discussion of informed consent and of "how medical decisions are usually made for minors, the ethical reasons why this is so, and then how decision-making and consent procedures differ in the case of gender transition." Bester's view is that "minors cannot be asked to consent for these treatments, nor lead the decision-making around them."

Certainly, there is more that could be said about informed consent. The Review focuses on the clinically and ethically prior question of whether it is permissible to offer PMT to patients in the first place. Issues of autonomy and consent become pressing only after it has first been established that it is clinically and ethically justified to offer some intervention to patients (see Chapter 13). Because a favorable risk/benefit profile is a necessary condition of any pediatric intervention being ethically justified, a robust discussion of the question of whether minors (or their legal guardians) can consent to PMT would be premature. While it may be valuable to explore the issue of patient autonomy and consent within a hypothetical context in which PMT were known to have a favorable risk/benefit profile, the scope of the Review was limited to an assessment of best practices within the context of existing evidence of risks and benefits.

3. "It has struck me for a while now that the pressure to allow minors to lead decision-making in this area departs markedly from how medical decision-making for minors is usually done. Usually, parents are decision-makers for minors, and together with clinicians make decisions that serve the best interests of the minor."

We agree with Bester that the child-led "affirming" approach that has come to dominate some U.S. gender clinics departs markedly from how pediatric medicine is generally practiced, where medical decision-making is grounded in the health-related best interests of the patient. As quoted in Section 13.2.2 of the Review, the AAP's Committee on Bioethics emphasizes that "parental authority regarding medical decision-making for

their minor child or young adult who lacks the capacity for medical decision-making is constrained compared with the more robust autonomy in medical decision-making enjoyed by competent adults" and, moreover, that clinicians' fiduciary duties "to protect and promote the health-related interests of the child and adolescent … may conflict with the parent's or patient's wishes …" (Committee on Bioethics et al., 2016, pp. e5, e2).The centrality of the best interest standard to ethical clinical practice in pediatrics is not controversial. The apparent rejection of this standard among some practitioners of PMT, whether implicit or explicit, is a further indication that the field of pediatric gender medicine in the U.S. has become exceptionalized.

Reply to Gribble

We thank Professor Karleen Gribble for her thoughtful review. She raises some important points, to which we respond below.

1. Gribble recommends (a) mentioning "the risks of using terminology suggesting that people can change their sex"; (b) reconsidering the terminology of "male-to-female," "female-to-male," and "sex reassignment surgery"; and (c) noting that "not everyone applies the concept of gender identity to themselves."

Regarding (a), we have added the following underlined text to footnote 16 in Section 2.1 ("Terminology in pediatric gender medicine"):

> American Psychological Association (2024a). <u>The APA has the "problematic implication" backwards: terminology that suggests a person's sex is a *mutable* characteristic is misleading to patients and should be avoided</u>.

Regarding (b), there is a tradeoff between coining new terminology which may tax the reader and using familiar terminology that is less-than-ideal. "Male-to-female" and "female-to-male" are very familiar and readily interpretable as indicating the aspirational direction of travel rather than a literal change of sex. "Sex reassignment surgery" has the disadvantage that it suggests a prior "assignment" but is less problematic than the older "sex change" or the current "gender confirmation/affirmation surgery." Rather than multiplying terminology, we think it best to keep to our original usage.

Regarding (c), we do quote from Sullivan (2025) in footnote 41 (Section 2.2, "Terminology in this Review"): "Questions on gender identity should recognize that the concept of gender identity as such will be unfamiliar, unclear or irrelevant to some respondents, and that many respondents may not perceive themselves as having a gender identity. Questions should not assume that respondents will agree that they have a gender identity." We have altered the start of that footnote to bring out Gribble's point more clearly:

> It should be emphasized that not everyone accepts that they have a gender identity. As the RSG puts it …

2. “[N]either the body of the Review nor the overview of the systematic reviews includes harm in terms of inability to breastfeed in the analysis of findings. I would suggest that this be addressed … It may be helpful to provide a citation explaining the nature of chest masculinisation surgery to make it clear why breastfeeding is prevented.”

In the footnote quoted by Gribble we do say that “loss of breastfeeding function” is one of the outcomes that “the Review considers harms” (footnote 30, Section 13.2.3). However, Gribble correctly observes that this is not mentioned in the text.

We have added a sentence after “Further, surgeries to remove healthy and functioning organs introduce a unique set of iatrogenic harms not encountered in other areas of medicine” at the beginning of Section 7.5:

> An example is mastectomy performed as part of PMT, which results in an inability to breastfeed and potential loss of nipple sensation.

We have also added a footnote to the above sentence:

> A mastectomy removes the mammary glands together with the ducts that transfer milk from them to the nipple. Loss of nipple sensation is invariably mentioned by surgeons performing “top surgery” as a potential side effect, although there is very low certainty evidence about the magnitude of the risk in adolescents and young adults (Miroshnychenko et al., 2025).

3. “The Review does not discuss breast binding … I would encourage the authors to consider reconceptualising breast binding and genital tucking in the Review as a physical intervention rather than as part of social transition.”

The first paragraph in Section 5.2 (“Outcomes of social transition”) ends: “As noted in the Cass Review, even though social transition is undertaken outside healthcare settings, ‘it is important to view [social transition] as an active intervention because it may have significant effects on the child or young person in terms of their psychological functioning and longer-term outcomes.’”

We have added text to the accompanying footnote (26), which cites Cass (2024, p. 158):

> Social transition may also involve breast binding for females or "tucking" (moving the testes into the inguinal canals and positioning the penis and scrotum in the perineal region) for males. This is a (non-medical) physical intervention with potentially adverse health effects, unlike haircuts or clothing changes. As a recent review puts it, "For chest binding, a significant number of negative health implications have been reported, with rates as high as 97.2%" (Bumphenkiatikul, 2025, p. 5). This review did not attempt to synthesize the quality of the evidence for either harms or benefits, however.

4. "Regret associated with chest masculinisation surgery is not mentioned but should be added."

We have added the underlined text to the first paragraph at the start of Section 7.6.2 ("Detransition and regret"):

> Patients of any age may experience regret regarding the permanent physical and physiologic effects of CSH, regardless of how they identify. For example, stably transgender-identified patients may regret loss of fertility. Developing baldness, or chafing/discomfort caused by clitoromegaly, may lead a patient who identifies as a transgender man to regret taking testosterone. <u>Such a patient, or (especially) a detransitioned female, may regret having had a mastectomy with the consequent loss of the ability to breastfeed.</u>

Reply to Santen

We thank Dr. Richard Santen for his thoughtful and substantive comments on the HHS Review. He finds the overview of systematic reviews ("umbrella review") "particularly helpful as it covers an extensive volume of data and provides an assessment of the level of validity of each review" and concludes that "the overall assessment of these studies was scientifically sound." Santen further judges Chapter 7, which supplements the evidence from systematic reviews with evidence from basic science and physiology, to "contain scientifically valid information." Santen's general assessment is that the Review's "summary of data and detailed discussions reasonably reflect an overview of the information currently available and its interpretation."

Santen has two major comments, the first about whether the practice of pediatric medical transition (PMT) should be designated "experimental," and the second about "panel stacking." We will address these in reverse order, before turning to Santen's more minor comments.

1. Santen recommends that the Review be more explicit about "the concept of 'stacking', its definition, and its role in guideline development."

"Panel stacking" refers to the practice of populating clinical practice guideline development groups with individuals who share a similar position regarding the treatment under consideration, often due to having financial or non-financial conflicts of interest. Because of its tendency to perpetuate groupthink, panel stacking represents a threat to the trustworthiness of any clinical practice guideline (CPG), especially in the absence of systematic reviews of evidence.[12] Managing conflicts of interest is essential in CPG development because CPGs make recommendations factoring in not only the evidence but a range of other considerations, such as "values and preferences."

The Review does discuss panel stacking in the development of World Professional Association for Transgender Health (WPATH) guidelines (Section 10.3.1, "Conflicts of

[12] Kepp et al. (2024).

interest management"). However, we agree it does not sufficiently address the same issue with respect to the Endocrine Society (ES) guidelines, as Santen helpfully notes.

It is important to emphasize that an "interest" is not the same as a "conflict of interest." A conflict occurs "when a past, current, or expected interest creates a significant risk of inappropriately influencing an individuals' judgement, decision or action when carrying out a specific duty."[13] Financial interests, such as deriving income from an area under review, are widely understood to compromise a CPG's trustworthiness. However, non-financial interests (such as personal beliefs, political positions, or personal histories) can in some cases have similar effects as financial ones, as strongly held beliefs may create cognitive distortions, preventing a person from adjusting his or her position when the evidence requires it.

The ES guidelines for this area of medicine have been heavily influenced by the Dutch clinician-researcher team that pioneered the practice of PMT. Three of the eight authors of the 2009 ES guidelines, which first introduced PMT into clinical practice, were the founders of the Dutch Protocol: Peggy Cohen-Kettenis, Henriette A. Delemarre-van de Waal, and Louis J. Gooren.[14] The group also included Norman Spack, who co-founded the first U.S. pediatric gender clinic. Most of the authors were also prominent WPATH members and leaders. As discussed in the Review (Chapter 9), the 2017 update to the ES guidelines continued to maintain a strong link with the Dutch clinical team and further cemented the relationship with WPATH through common authorship.

The intellectual commitments of ES guideline panel members find expression in the fact that ES recommended PMT despite not having conducted systematic reviews (SRs) of evidence for benefits and risks for the relevant population—a strong departure from norms governing how CPGs should be drafted. In addition, the 2017 ES guideline's "values and preferences" place a "high value" on a satisfying cosmetic outcome and a "lower value on avoiding potential harm from early pubertal suppression"[15]—a surprising

[13] Akl et al. (2022).
[14] Hembree et al. (2009).
[15] Hembree et al. (2017, p. 3881).

inversion of basic principles of medical ethics in the context of endocrinological interventions for minors with no physical pathology (see Chapter 13).

It is important to recognize that COIs or perceptions of COIs are hard to avoid whenever subject-area experts are involved in CPG development.[16] However, COI management is essential. It includes transparent disclosures and careful COI management (e.g., recruiting to the panel individuals with a diversity of positions and ensuring that recommendations are based on impartial appraisal of evidence, conducted by methodologists).

We believe the Review adequately describes COI problems in the development of WPATH guidelines, but we have added a three-paragraph summary of the points above in Section 9.2.3.

> "Panel stacking" refers to the practice of populating clinical practice guideline development groups with individuals who share a similar position regarding the treatment under consideration, often due to having financial or non-financial conflicts of interest. Because of its tendency to perpetuate groupthink, panel stacking represents a threat to the trustworthiness of any CPG, especially in the absence of systematic reviews of evidence.[39] Managing conflicts of interest is essential in CPG development because CPGs make recommendations factoring in not only the evidence, but a range of other considerations, such as "values and preferences" (see Section 10.3.1).
>
> ES guidelines for this area of medicine have been heavily influenced by the Dutch clinician-researchers who pioneered the practice of PMT. Three of the eight authors of the 2009 ES guidelines, which first introduced PMT into clinical

[16] Even though the Review is not a CPG, we recognize that the same allegation of "intellectual COIs" could be leveled at the Review team itself. Ultimately, intellectual COIs in this highly contentious area of medicine are unavoidable because all knowledgeable individuals have considered opinions. As stated in our reply to the American Psychiatric Association, the Department of Health and Human Services took considerable effort to minimize the influence of intellectual COIs by involving experts from a diversity of political backgrounds and ensuring a robust evidence review from an expert methodologist. However, the Review's analysis should be assessed on its merits, and if there is evidence that interests might have led to an inappropriate interpretation, decision, or action, that evidence should first be identified. No such evidence has been provided by any of the peer reviews.

practice, were the founders of the Dutch Protocol: Peggy Cohen-Kettenis, Henriette A. Delemarre-van de Waal, and Louis J. Gooren.[40] The group also included Norman Spack, who co-founded the first U.S. pediatric gender clinic. Most of the authors were also prominent WPATH members and leaders. The 2017 update to the ES guidelines continued to maintain a strong link with the Dutch clinical team and further cemented the relationship with WPATH through common authorship.

The intellectual commitments of ES guideline panel members find expression in the fact ES recommended PMT despite not having conducted SRs of evidence for benefits and most risks—a strong departure from norms governing how CPGs should be drafted—and the unusual "values and preference" statements mentioned above.

(Associated footnotes: [39] Kepp et al. (2024); [40] Hembree et al. (2009).)

2. Santen asserts that it is "essential" to state "whether gender-affirming hormone therapy (i.e., puberty blockers and cross-sex hormone therapy) is experimental or accepted practice." He points out that health authorities in several countries (e.g., Sweden, Finland, and the U.K.) deemed some or all aspects of the endocrine protocol "experimental" and restricted it to research. Santen recommends including a separate discussion of "experimental" status in the HHS Review.

While the Review reflects on this theme in several places, the following additional observations may be helpful.

(a) PMT entered clinical practice without proper testing. The use of PMT in the Netherlands was initially rolled out under what can be best described as the "innovative practice" framework. The framework allows for certain promising treatments to be attempted on a small scale, provided the drug has already been approved for another indication, the affected population is expected to be small, and no alternatives exist. Ethical considerations require that once such treatments gain momentum, they must be placed into high-quality research settings as soon as possible to prevent "runaway

diffusion."[17] PMT never entered the clinical trials phase; despite its widespread use, it is best understood as pre-clinical.

(b) Existing NIH research skipped critical steps. In 2014, a group of leading American gender clinicians applied for NIH funding for a proposed observational study, "The Impact of Early Medical Treatment in Transgender Youth." In their grant proposal, the researchers wrote that their study "will be the first in the U.S. to evaluate longitudinal outcomes of medical treatment for transgender youth and will provide essential evidence-based data on the physiological and psychosocial effects and safety of treatments currently used for transgender youth."[18] Despite receiving significant funding, and hundreds of children being subjected to the risks of PMT, the research omitted Phase II/III testing, which is aimed at evaluating efficacy for a new indication. Instead, the research examined PMT as if it had already been established as standard practice, resembling Phase IV (postmarketing review) research. This is a highly unusual practice.

(c) NIH applications disclose how little is known about PMT. Following the original application, in 2019 the same group of researchers wrote that extant guidelines were based on "very limited data" and "minimal data examining the long-term physiologic and metabolic consequences of gender-affirming hormone treatment in youth."[19] As recently as 2024, the researchers' request for reauthorization of funding continued to describe the evidence base for PMT as "scant."[20]

(d) The term "experimental" has multiple meanings. Currently, the interventions that comprise PMT (notably, puberty blockers, estrogen and testosterone blockers for males, testosterone for females) are used "off-label," which means the drugs are FDA-approved for other indications. Off-label treatment can reflect established as well as experimental practice.

The term "experimental" has technical meanings in the context of U.S. law and policy. Experimental therapies are typically excluded from coverage under state definitions of

[17] Abbruzzese et al. (2023).
[18] Regenstreif (2023).
[19] Olson-Kennedy et al. (2019).
[20] National Institutes of Health (2024).

"medical necessity." A widely recognized criterion for an intervention to be considered experimental is that its safety and efficacy profile is inadequately known. However, different states have different thresholds for "medically necessary" vs. "experimental." A few examples:

- Massachusetts defines "experimental" services as "any service for which there is insufficient authoritative evidence that such service is reasonably calculated to have the effect described in [Massachusetts' statutory definition of 'medical necessity']."[21]
- New Jersey defines "medical necessity" and "experimental" in a way that gives more weight to "expert … opinion" and "community acceptance."[22]
- Tennessee considers a therapy experimental "if there is inadequate empirically based objective clinical scientific evidence of its safety and effectiveness for the particular use in question. This standard is not satisfied by a provider's subjective clinical judgment on the safety and effectiveness of a medical item or service or by a reasonable medical or clinical hypothesis based on an extrapolation from use in another setting or from use in diagnosing or treating another condition."[23]

Santen is right that a decision to label PMT "experimental" would have significant consequences. From the payor perspective, it likely would justify denial of coverage. From a research perspective, it would require submitting PMT to proper, IRB-approved clinical trials, likely following animal studies for drug safety. The latter is particularly important as it is now widely recognized that puberty blockers are not a standalone intervention but nearly always followed by cross-sex hormones. Given the complexities involved, we believe discussion of this issue is beyond the scope of the HHS Review.

[21] Commonwealth of Massachusetts (2017). Medical necessity is defined in terms of two conditions: "(1) [the service] is reasonably calculated to prevent, diagnose, prevent the worsening of, alleviate, correct, or cure conditions in the member that endanger life, cause suffering or pain, cause physical deformity or malfunction, threaten to cause or to aggravate a handicap, or result in illness or infirmity; and (2) there is no other medical service or site of service, comparable in effect, available, and suitable for the member requesting the service, that is more conservative or less costly to the MassHealth agency …"
[22] New Jersey (n.d.).
[23] Tennessee (2024).

However, healthcare decision-makers, including payors and regulators, should examine this issue carefully and adjust their policies and actions accordingly.

3. Santen takes issue with the HHS Review's claim that "the natural history of pediatric gender dysphoria is poorly understood, though existing data suggests it will remit without intervention in most cases." The Review, he argues, fails to distinguish persistence in childhood versus persistence in adolescent gender dysphoria (GD). Although GD "is known to commonly resolve" in children, Santen explains, the evidence that it commonly resolves in adolescents without a prepubertal history of GD is "not considered scientifically sound." (Another reviewer—Strathearn—made a similar comment in a prepublication review.) Santen recommends deleting the sentence from the Review.

The Review addresses the differences between childhood and adolescent persistence in Section 4.3.2. As the Review notes, the claim that adolescent GD (unlike childhood GD) is stable has been asserted without evidence and is a central justification for PMT. The Review emphasizes that there is a dearth of research on this question ("the natural history … is poorly understood"). We agree that the Review should note the tentative nature of emerging research on low diagnostic stability.

"Tentatively" was added to Section 4.3.2.1 ("New evidence about the natural history of gender dysphoria"):

> Although the natural history of GD—i.e., its course absent medical interventions—is currently impossible to measure given the wide availability of interventions, newer evidence tentatively suggests that GD has a low diagnostic stability.

4. Santen suggests that the ranges specified in the Review (Sections 7.4.3 and 7.4.4) for testosterone in females and estradiol in males are too high and should be 10 to 35 ng/100 mL for testosterone (cf. 2-45 ng/dL in the Review) and 10 to 40 pg/mL for estradiol (cf. 60–190 pg/mL in the Review).

The reference ranges cited in the Review reflect ones commonly used in laboratory settings.[24] "[A] standard reference range for estradiol" in Section 7.4.4 refers to total estrogen, not estradiol; accordingly, "estradiol" has been replaced by "total estrogen." The Review has also been updated to reflect more recent ranges, which are substantially similar to the older ones. The new reference ranges are 10 to 55 ng/dL for testosterone in females and 56 to 213 pg/mL for total estrogen in males. Santen's proposed ranges are also reasonable. Were we to rely on them, the result would be an even larger discrepancy between the reference range for normal female testosterone and the range recommended by PMT guidelines. We have added a footnote in Section 7.4.3:

> Labcorp (2025b). There is variability in laboratory reference ranges for testosterone (as well as for estrogen; see Section 7.4.4 below).

While experts may disagree about the specific reference ranges, the critical point is that whatever reference range is used, the hormonal regimens recommended by WPATH and the Endocrine Society for purposes of medical transition far exceed the normal ranges of estrogen/estradiol in males and testosterone in females.

5. Santen disagrees with the HHS Review's comment (Section 7.6.1, "Adverse psychiatric effects") about anabolic steroid abuse as "the amounts of anabolic steroid that cause the symptoms described [cardiovascular and psychiatric adverse reactions] are very much higher than the amounts used as cross-sex hormone therapy."

The Review discusses the use of testosterone in females, where the normal reference range of testosterone is much lower and narrower than in males. By extension, the relative increase in testosterone above the reference range is very large, creating, in our view, risk for harm. Santen suggests that risk for harm is related to absolute, not relative, values. The disagreement seems to hinge, in part, on whether one agrees with the Review's citation of Gomez-Lumbreras & Villa-Zapata report of the FDA's Event

[24] See, e.g., the reference ranges used by Labcorp for total estrogen, estradiol, and testosterone (Labcorp, 2024, 2025a, 2025b).

Reporting System (FAERS) data and the Laidlaw & Jorgensen comment about the data (Section 7.6.1). We appreciate Santen's perspective and agree that more evidence—specifically, on whether it is the absolute level of testosterone or the level relative to the normal female range that increases risk for psychiatric problems—would allow for a more confident assessment of the phenomenon.

We have made some changes to Section 7.6.1. "In men" was added to a sentence in the second paragraph:

> One study assessing medium (300–1000 mg/week) and high (>1000 mg/week) anabolic steroid use in men found that 23% of users ...

The following has been added to the end of the third paragraph:

> It is unknown whether these patients had testosterone levels between 320 to 1000 ng/dL (the range recommended by the Endocrine Society for females undergoing medical transition), or levels outside of this range. What is known is that the patients were female, were categorized as having a "transgender" related treatment indication, presented with psychiatric problems, and were on testosterone.[99] Although it is not possible to determine causation from FAERS data, this underscores the importance of considering adverse psychiatric events as a potential risk in female patients initiating testosterone for medical transition.

(Associated footnote: [99] See Gomez-Lumbreras & Villa-Zapata (2024), Table 2.)

6. Santen notes the existence of new published research associated with the NIH-funded Olson-Kennedy et al. initiative and encourages the contributors to address this research ("with the caveat that it is not peer-reviewed").

Two studies associated with the Olson-Kennedy et al. initiative have been published in 2025: "Mental and emotional health of youth after 24 months of gender-affirming care initiated with pubertal suppression" (Olson-Kennedy, Durazo-Arvizu et al. 2025) and "Emotional health of transgender youth 24 months after initiating gender-affirming hormone therapy" (Olson-Kennedy, Wang et al. 2025). The first was published as a

preprint (not peer-reviewed), and only after the HHS Review was published on May 1. For a critical analysis, see Society for Evidence-Based Gender Medicine (2025).

The Review cites the second study on cross-sex hormones. As explained in Section 5.7.3 ("Robustness of this overview's conclusion"), "rather than extending beyond what the evidence can support, this overview is confined to summarizing the conclusions of SRs [systematic reviews]. As a result, it may not include some of the most recently published studies due to the timing of the SRs' literature searches. However, a targeted search [the footnote cites Olson-Kennedy, Wang et al. (2025)] of recently published studies did not reveal any published or ongoing studies that would significantly change the conclusions, especially those pertaining to benefits. This is due to ongoing problems such as an absence of comparison groups, inadequate sample sizes, and limited follow-up."

7. Santen recommends that the Review "highlight the differences in results between birth assigned males and females [in Chen et al. (2023)] as an adjunct to the discussion of the Olson-Kennedy manuscript."

The Review does highlight those differences in a section devoted to Chen et al. (2023) (6.2.3): "The only statistically significant improvement in both sexes was in 'appearance congruence' as measured by the 'transgender congruence scale,' which has not been validated in minors. The authors also reported that there were statistically significant improvements in depression, anxiety and life satisfaction. However, these improvements were small and of questionable clinical significance. The statistically significant improvements were observed only in females, whereas males experienced no significant improvement in these measures."[25]

[25] Footnotes omitted.

Reply to Smids

We are grateful to Dr. Jilles Smids for his incisive comments on the HHS Review. Smids concludes that the Review "as a whole provides a comprehensive interdisciplinary and well-argued analysis of pediatric gender medicine." In particular, Smids states that Chapter 13 provides "one of the most comprehensive and thorough ethical analyses of current pediatric gender medicine."

While Smids's review is positive overall, he raises some important concerns. We respond to his main critical points below.

1. Smids acknowledges that Chapter 2's treatment of relevant terminology is "important and provides essential insights" but worries that its "skepticism regarding the *term* gender identity may easily be taken for a wholesale skepticism regarding the *experience* of gender incongruence and may come across as dismissive to the importance that gendered feelings have for trans persons."

Chapter 2 of the Review explains that while advocates of pediatric medical transition (PMT) use and rely on the term "gender identity," the term's meaning has shifted since it was first introduced by clinician researchers in the mid-20th century. At present there is no scientifically useful or indeed even coherent definition of the term in the field's authoritative clinical practice guidelines and policy statements, such as those published by the World Professional Association for Transgender Health (WPATH) and the American Academy of Pediatrics (AAP). Because "gender identity" plays a central role in decisions about medical interventions, the terminological issues discussed in the Review represent a serious problem for the field. Chapter 2 also describes other examples of scientifically ungrounded, misleading, or euphemistic terminology and argues that clinical solutions arrived at by deploying such language violate clinicians' "professional duty to apprise their patients of their conditions and the treatment options in language that is accurate, ethically neutral, and in no way misleading."

The Review recognizes that some children and adolescents experience discomfort or distress regarding their sexed bodies or associated social roles and expectations. It aims to describe and assess current best practices for the treatment of children and

adolescents facing precisely this challenge. The Review notes that leading U.S. professional medical societies and clinicians working in gender clinics have adopted terminology that is unhelpful at best for describing their patients' experiences or problems, but it does not deny or minimize these experiences or problems. On the contrary, the Foreword states that when patients seek professional help, "they and their families should receive compassionate, evidence-based care tailored to their specific needs."

Discerning what care is tailored to patients' specific needs requires clear language and scientifically accurate terms. Toward that end, we emphasize again that "the understandable desire to avoid exclusionary or pathologizing language—combined with beliefs firmly embedded in the field—has led to a vocabulary and a mode of communicating that is scientifically ungrounded, that presupposes answers to ethical controversies, and that is in other ways misleading" (Section 2.1).

2. While Smids acknowledges that the Review's research ethics analysis appropriately relies on "established research ethical principles" requiring a reasonable anticipation of a positive balance of benefits over risks, he is not wholly convinced by the analysis. This is because he is not certain that in the research context "we are in a position to claim that predictions of possible overall benefit are unreasonable even for a small subset of GD adolescents currently undergoing PMT."

As with studies of any clinical intervention, the fact that studies to date do not find strong evidence that PMT improves health outcomes does not, in principle, rule out the possibility that some subpopulation of subjects benefits from the interventions while others are harmed by them. However, to date no subpopulation has been shown to benefit. Moreover, clinicians are unable to predict which patients will experience persistent GD into adulthood and which will experience a resolution of symptoms. Nor do those clinicians who follow the American "gender-affirming" model try to make such predictions (see, e.g., Chapter 11).

We agree with Smids on the need for further research, but for reasons set out in Section 13.5 we find that "administering PMT to adolescents, even in a research context, is in tension with well-established ethical norms for human subjects research."

Delineating specific areas of future research in pediatric gender medicine is an urgent challenge that is outside the scope of the Review. However, it is important that the etiology and natural history of pediatric GD continue to be studied and that the short- and long-term health effects of hormonal interventions be characterized. We encourage researchers to explore alternative ways to study this vulnerable population, for example by analyzing existing data, recruiting research participants from the adult population who received a diagnosis of GD as children or adolescents (whether they medicalized or not), and conducting trials using less invasive and risky psychosocial interventions. We also emphasize that clinical research typically proceeds with a reasonably clear account of health and disease in the relevant population and with a good understanding of the clinical aims of the interventions. Here we note once again that in the field of pediatric gender medicine the rationales for medical intervention are much contested, a problem revealed in the fundamentally different nosological approaches adopted by the DSM-5 and the ICD-11, as discussed in Section 13.3.

3. According to Smids, Chapter 11 of the Review is "far more accusative than fitting for the type of report the HHS analysis aims to be, accusing even clinicians who have just become the target of legal procedures." While he credits Chapter 11 with "providing valuable insights," he claims the "fundamental principle" that ought to have guided the chapter is the principle that "one is innocent until proven guilty."

Chapter 11 concerns safeguarding failures in pediatric gender medicine. The chapter describes how leading clinicians and clinics have strayed from ethical standards of pediatrics in ways that put young and vulnerable patients at risk of serious harm. The evidence set forth in the chapter includes direct quotes from leading clinicians, while additional testimonial evidence is provided by whistleblowers. It would be irresponsible for a comprehensive assessment of pediatric medical transition in the U.S. to ignore these clinical realities. While we agree with Smids that "one is innocent until proven guilty," the chapter does cite sufficient evidence for its conclusions. Of course, the chapter makes no claims regarding the law, as doing so would exceed its scope as well as the professional expertise of its contributors.

Reply to Strathearn

We thank Dr. Lane Strathearn for the time and effort spent in compiling helpful peer review comments. Strathearn praises the HHS Review as a "comprehensive summary of the evidence base for many treatment practices in pediatric gender medicine" and "a valuable and much needed contribution to this important field of practice." He also notes its "strong focus on evidence-based medicine, outlining both the strengths and limitations, supplemented by indirect evidence from basic science and physiology to better understand mechanisms and the likely risk/benefit ratio of treatment."

We appreciate Strathearn's first-hand example of probable publication bias. Articles articulating a more neutral or critical account of the problems and uncertainties in pediatric gender medicine are too often rejected by leading academic journals (sometimes accompanied by dismissive peer-reviews that make politicized arguments rather than focusing on science and evidence). In contrast, studies that claim positive effects of pediatric medical transition (PMT) appear to pass peer review easily, even when the conclusions are not supported by the data presented. A recent article (Cohn, 2025) describes examples of this sort in gender medicine research. The Review recognizes these problems and discusses them in various parts, especially Section 6.3.

Strathearn points out some minor errors (a dead link, figures not referenced in the text, etc.). These have been fixed. He also suggests that Figure 9.3 be simplified to focus on guidelines used in the U.S. Since the original figure was reproduced from a published study, we opted to keep the original version.

We now turn to Strathearn's more significant comments.

1. Strathearn suggests that in the Foreword, "it is important to acknowledge that there is also insufficient evidence to clearly understand the 'risk of potential harm' for some of these treatments. For example, the long-term outcomes (both risks and benefits) are uncertain for all treatment modalities … Nevertheless, the responsibility for medical practitioners to 'first do no harm' means that the primary burden of evidence should be for the likelihood of benefit, especially when there is even a potential for harm."

The following underlined text was added to the Foreword (paragraphs four and seven):

> Having recognized the experimental nature of these medical interventions and their potential for harm (which has been inadequately studied, especially with respect to long-term outcomes), health authorities in a number of countries have imposed restrictions.
>
> Nevertheless, the "gender-affirming" model of care includes irreversible endocrine and surgical interventions on minors with no physical pathology. These interventions carry risk of significant harms including infertility/sterility, sexual dysfunction, impaired bone density accrual, adverse cognitive impacts, cardiovascular disease and metabolic disorders, psychiatric disorders, surgical complications, and regret, and there has been inadequate research into the frequency and severity of these harms. Meanwhile, systematic reviews of the evidence have revealed deep uncertainty about the purported benefits of these interventions.

2. Strathearn suggests that Part 1 of the Executive Summary should mention that some countries have restricted puberty blockers, cross-sex hormones and surgeries to research settings.

The following underlined text was added to Part 1 of the Executive Summary:

> … health authorities in an increasing number of countries have restricted access to puberty blockers and cross-sex hormones, and, in the rare cases where they were offered, surgeries for minors. These authorities now recommend psychosocial approaches, rather than hormonal or surgical interventions, as the primary treatment, and in some cases have restricted the latter to nationally-overseen research protocols.

3. Strathearn found Chapter 3 to be "somewhat based on conjecture and hearsay" and noted it could be vulnerable to bias.

Chapter 3 provides a brief history of adult and pediatric gender medicine. It follows well-established scholarly conventions, supporting its claims with peer-reviewed and primary

source literature, which readers can consult for verification. Strathearn does not identify specific examples of errors or mischaracterizations, and neither, for that matter, do the proponents of PMT (the APA, Dowshen et al., and Rider et al.), to whom we have replied here.

4. Strathearn notes that in Section 4.1, Figure 4.2 should include error bars to assess the variability of the mean scores. He also raises a question about the distribution of the scores.

Figure 4.2 was updated to include 95% confidence intervals, with “95% confidence intervals added” placed in a footnote. The data reported in the original Dutch research are insufficient to answer Strathearn’s reasonable question about distribution.

5. Strathearn suggests that the uncertain evidence for psychotherapy outcomes should also be mentioned in Section 5.7.5 (“Conclusion”).

We have added a sentence at the end of this section:

> This overview synthesizes the best available clinical evidence from population-level data, highlighting a consistent pattern across interventions for children and adolescents with GD. The benefits and harms of social transition remain unknown; PBs, CSH, and surgeries consistently produce certain physical and physiological effects; and there is considerable uncertainty regarding their psychological and long-term health outcomes. Likewise, there is uncertainty regarding the effects of psychotherapy for GD.

6. Strathearn requests clarification regarding the following statement in the introduction to Chapter 6: “It is well-established in adults that for the same drug, off-label uses are associated with considerably higher rates of adverse effects, especially when strong scientific evidence is lacking.”

A citation in the introduction to Chapter 5 (Eguale et al., 2016) reports:

> We found that off-label use of drugs was associated with ADEs after adjusting for important patient and drug characteristics. Moreover, we noted a risk gradient with higher rates of ADEs for off-label uses lacking strong scientific evidence.

> Although the intrinsic nature of the drug to cause ADEs is the same for on-label and off-label uses, it may be modified by a number of factors, including the off-label disease condition. In addition, the lack of approval from a regulatory body implies a lack of safe dose ranges and inadequate information on contraindications, which in aggregate make ADEs more likely. We found that 4 in 5 off-label prescriptions lacked strong scientific evidence, and this group had higher rates of ADEs.

7. Strathearn suggests that the suicide rate comparisons in Section 6.2.3 be improved and more thoroughly cited.

The following underlined text was added to the discussion of suicide:

> However, two of the study subjects died by suicide within one year of initiating hormones, representing <u>an annualized suicide rate of 317 per 100,000 patients. The suicide rate in Chen et al. was higher than rates that have been reported by PGM clinics in the U.K. and Finland (13 per 100,000 and 51 per 100,000).[40] One Belgian study[41] has also reported a comparatively high annual suicide rate (1,126 per 100,000); like in Chen et al., all patient suicides in that study were among patients taking CSH.[42]</u>

(Associated footnotes: [40] Society for Evidence-Based Gender Medicine (2023a); [41] Van Cauwenberg et al. (2021); [42] Society for Evidence-Based Gender Medicine (2024a). See also Section 4.3.4.)

Additionally, the following sentence was added after the end of the second paragraph in Section 4.3.4, after which Table 4.1 was added:

> Table 4.1 reports suicide rates from four PGM clinics and notes whether the suicides were in patients who had received hormonal interventions.

Table 4.1. Suicide mortality in youth referred to pediatric gender medicine clinics (with estimated per 1,000 patient-years rates)

Study; Country	Age range	Years	Referred youths	Suicides	%	Per 1K patient-years	Were patients who died by suicide taking PBs and/or CSH?
Van Cauwenberg et al. (2021); Belgium	12–18	2007–2016	148	5	3.38	11.26	All suicides among patients taking CSH
Chen et al. (2023); U.S.	12–20	2016–2021	315	2	0.63	3.17	All suicides among patients taking CSH
Ruuska et al. (2024); Finland	<23	1996–2019	2,083	7	0.34	0.51	Unknown (38% of cohort treated with PBs and/or CSH)
Biggs (2022); U.K.	<18	2010–2020	15,032	4	0.03	0.13	Unknown (59% of cohort treated with PBs and/or CSH)

Table adapted from Society for Evidence-Based Gender Medicine (2024a).

8. Referring to Chapter 13's discussion of psychotherapy, Strathearn correctly notes that, as in the case of medical interventions, "*no evidence* for harm does not equate with 'no potential harm.'"

The following underlined text was added to paragraph three of Section 13.2.3:

> Regarding the potential harms of psychotherapy for adolescents with GD, a systematic review of the evidence found no evidence of negative or adverse effects in any of the studies examined <u>(although absence of evidence for harm does not imply evidence of no harm, psychotherapy does not carry the medical or surgical risks associated with PMT).</u>

Reply to Bekkering & Vankrunkelsven

We thank methodologists Dr. Trudy Bekkering and Professor Patrik Vankrunkelsven for their meticulous peer review comments.

Bekkering and Vankrunkelsven focus on Chapter 5, “Overview of Systematic Reviews,” (“umbrella review”) and Appendix 4, which includes the full methodological details of our overview of systematic reviews (SRs). They commend the Review’s robust methodology and agree with the rationale for an umbrella review, which is justified “by the fact there are many SRs [in this field], most using the same studies.” They note the umbrella review’s adherence to Cochrane standards, the comprehensive literature search across multiple databases, and appropriate use of the Risk of Bias in Systematic Reviews (ROBIS) tool and the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework for assessing risk of bias and certainty of evidence.

Bekkering and Vankrunkelsven find no major issues with the Review’s conclusions. As they put it:

> Certainty of evidence is very low. Not just because there are no randomized controlled trials (RCTs), as well designed observational studies would also be very helpful. There are no new or ongoing studies that would have an important impact. New studies are needed. New SRs are unlikely to yield novel insights.

Bekkering and Vankrunkelsven have some “minor remarks,” which we address below.

1. “The lack of rigorous reporting of conflicts of interest (COI) by authors is the most important issue here, given the topic.”

As we explain in our reply to the American Psychiatric Association, “Given the highly polarized nature of the topic, contributors’ names were withheld during the peer-review process so that reviewers could focus on the content of the review, rather than on the individual contributors themselves. This is an established practice in scientific review, designed to reduce reviewer bias and ensure impartial focus on substance. The scientific integrity of any document, including the Review itself, is best assessed through its content.” Conflict of interest disclosures are reported in the revised Review.

2. “A definition of an SR (to be included in the umbrella [review]) would have been useful, but we found no issues on inclusion or exclusion of SRs.”

Section 5.1 of the Review quotes this definition of a systematic review from the *Cochrane Handbook for Systematic Reviews of Interventions*:[26]

> [A systematic review] attempts to collate all empirical evidence that meet[s] prespecified eligibility criteria in order to answer a specific research question. It uses explicit, systematic methods that are selected with a view to minimizing bias, thus providing more reliable findings from which conclusions can be drawn and decisions made.

We have also added a footnote to Section 1 of Appendix 4:

> An SR needs to have: 1) a defined research question according to PICO elements: Population, Intervention, Control/Comparator, Outcome; 2) pre-defined eligibility criteria for studies; 3) adequate systematic search methods that identify all studies that would meet the eligibility criteria; 4) an assessment of the validity of the findings of the included studies, for example through a risk-of-bias assessment; and 5) a systematic presentation and synthesis of the characteristics and findings of the included studies, which may include a meta-analysis. Scoping reviews, overviews of systematic reviews (umbrella reviews), and narrative reviews, are not SRs.

3. “The registration of the protocol would have increased transparency, as would more details about how the results were summarized. However, the final results are described transparently and are easy to follow. There are also many tables with necessary and relevant information.”

Bekkering and Vankrunkelsven correctly note that the protocol was not pre-registered. PROSPERO protocol registration bolsters openness and accountability. Unfortunately, given the time constraints, preregistration was not possible in this case. Preregistration also would have revealed contributor names. Given the polarization of this issue, it was

[26] Higgins et al. (2019, p. xxiii).

important for the peer-review process that names not be disclosed until that process was completed.[27] In short, competing considerations had to be balanced, and trade-offs had to be made. We think the methodological information in Appendix 4 is sufficiently detailed.

4. “No information was available on support, author information and availability of data and other information.”

With respect to support, the Review was commissioned from the contributors by the HHS contractor for this project. There were no other sources of support.

Regarding contributor information, see reply to point 1 (above).

Regarding data availability, all relevant details regarding the methodology and results are in Chapter 5 and Appendix 4. We appreciate that Bekkering and Vankrunkelsven found the results to be “described transparently and ... easy to follow." Their use of the Preferred Reporting Items for Overviews of Reviews (PRIOR) checklist[28] provides an independent check of our overview’s transparency, accuracy, and completeness.

[27] See footnote 11 in our reply to the APA.
[28] Pollock et al. (2019).

Reply to Dowshen et al.

In August 2025, the *Journal of Adolescent Health* published a commentary titled "A critical scientific appraisal of the Health and Human Services Report on pediatric gender dysphoria" (Dowshen et al., 2025).

The authors conclude that the Review engages in "numerous violations of scientific norms, misrepresentation of scientific evidence, and mischaracterizations of both gender identity in youth and the standard of care." The Review, they suggest, "is a dangerous example of government incursion into the provision of evidence-based medical care."

The commentary's allegations are serious; moreover, its authors are leaders in the field of pediatric gender medicine.[29] We have therefore decided to treat the commentary as an unsolicited peer review. We are grateful for the opportunity to address the collated feedback and concerns of gender clinicians and researchers who believe pediatric medical transition (PMT) is beneficial to patients and in line with existing standards in pediatric medicine.

We have organized our responses under five headings, A–E, corresponding to the themes of Dowshen et al.'s objections and comments.

[29] The authors are: Dr. Nadia Dowshen, gender clinician and co-director of the gender clinic at Children's Hospital of Philadelphia (CHOP); Dr. Kellan Baker, health policy expert and lead author of a systematic review on hormonal interventions and mental health outcomes commissioned by WPATH (which was analyzed by the HHS Review); Dr. Robert Garofalo, gender clinician and division head of Adolescent Medicine at Lurie Children's Hospital of Chicago; Dr. Diane Chen, pediatric gender medicine researcher and lead author of the NIH-funded research paper Chen et al. (2023) (which was analyzed by the HHS Review); Dr. David J. Inwards-Breland, gender clinician who has founded two pediatric gender clinics (Seattle Children's, UC San Diego) and currently serves as medical director for the gender clinic at Lurie Children's; Dr. Gina Sequeira, gender clinician who has served as co-director of the gender clinic at Seattle Children's Hospital; Dr. Jamie E. Mehringer, gender clinician and Assistant Professor of Pediatrics at the University of Rochester who founded the first pediatric gender clinic in the state of Vermont; and Dr. Meredithe McNamara, Assistant Professor of Pediatrics at Yale University and co-founder of the "Integrity Project," which publishes essays characterizing criticism of PMT as "misinformation."

A. Violations of scientific norms

1. Dowshen et al. suggest the Review lacks independence and that "the findings … were predetermined by the EO [executive order] that predated the writing of the report itself …"

We agree that scientific independence is critical to the Review's credibility. We note that no members of the contributor team are employed by the commissioning administration, that the empirical conclusions of the Review were arrived at via a transparent, reproducible methodology, and that the Review followed standard, scholarly norms of citation and argumentation.

The Review's central conclusions are based on an overview of systematic reviews (SRs) or "umbrella review" (see Chapter 5 and Appendix 4) and an ethics analysis (see Chapter 13). Each of these followed well-established principles in the relevant fields: evidence-based medicine and biomedical ethics, respectively. Each underwent independent peer-review by subject matter experts, and in both cases these experts concluded that the analyses are robust and consistent with high professional standards.[30] In our reply to the American Psychiatric Association (APA) we quote the observation of one reviewer (Dr. Jilles Smids), that if the Review's findings reflect the authors' bias, "it should be possible to point out where the reports engages in motivated reasoning, fails to do justice to the extant literature, or shows other problems." As we demonstrate below, Dowshen et al. have done none of these things.

2. Dowshen et al. are concerned that the Review "declines to name its authors, making assessment of their financial, intellectual, or other conflicts of interest impossible."

As HHS has stated, the decision to withhold names until completion of peer-review was intended to help maintain the integrity of the review process. This is standard practice in scientific publishing and promotes impartial engagement with the document's content rather than its provenance. Please also see our reply to the APA's peer review.

[30] See the peer reviews by Bekkering & Vankrunkelsven (evidence-based medicine) and Bester and Smids (biomedical ethics).

3. Dowshen et al. claim that the Review lacks credibility because “over 20%” of its references are not from the “peer-reviewed scientific literature.”

Focusing on the ratio of peer-reviewed to non-peer-reviewed sources is misguided. The important question is whether all relevant evidence is appropriately represented. The Review was tasked with evaluating both evidence and best practices. In addition to engaging with the relevant peer-reviewed scientific literature, the Review discusses non-peer-reviewed publications where appropriate. For example, a non-peer-reviewed essay, produced by the “Integrity Project” and posted on the Yale University Law School website in 2024, is relevant to the evidence for PMT and is cited in the Review.[31] The Integrity Project was co-founded by Dr. Meredithe McNamara, lead author of that essay and senior author of Dowshen et al.

The Review’s appraisal of best practices examines clinical realities in the U.S., which are frequently documented in court filings and media outlets such as *The New York Times*, *The Washington Post*, *The Boston Globe*, *Reuters*, and *The Economist*. Chapter 10 cites internal World Professional Association for Transgender Health (WPATH) documents which were obtained via the discovery process in a lawsuit. The email exchanges between senior WPATH members reveal that WPATH’s guideline development process flouted well-recognized standards.

Dowshen et al. do not dispute the veracity of claims made in Chapter 10 of the Review, which describes WPATH’s suppression of SRs and its revision of clinical recommendations in response to political pressures.

B. Misrepresentations of scientific evidence

4. Dowshen et al. claim the Review misused the very low GRADE designation of evidence quality for PMT as justification for “rejecting the standard of care for TGD [transgender and gender diverse] youth.”

Dowshen et al. mischaracterize the Review. First, as we explicitly note, the Review is not a clinical practice guideline and does not make policy recommendations. Second,

[31] McNamara et al. (2024).

contrary to Dowshen et al.'s suggestion, our conclusion regarding the risk/benefit profile of PMT does not rely exclusively on the fact (now demonstrated by many SRs and confirmed by the Review's umbrella review) that the evidence for benefit is of very low certainty. Rather, as explained in Chapter 8 and Chapter 13, this conclusion is supported by a standard risk/benefit analysis that incorporates both the purported benefits and the known risks and harms of the relevant interventions, as compared to the risk/benefit profile of the alternatives.

Decision-makers, including patients, their families, medical providers, and policymakers, must consider, among other factors, the strength of the evidence and ethical considerations, both of which were within the scope of the Review.

5. Dowshen et al. claim the Review "misrepresents … studies, often ignoring their primary conclusions." They give Chen et al. (2023) as an example. Dowshen et al. describe the findings of this study very positively: "appearance congruence, positive effect [sic], life satisfaction, and depression and anxiety symptoms all improved significantly following 2 years of hormone therapy." Dowshen et al. allege that the Review "ignores" these findings and "focuses solely on the two deaths by suicide among the study's 315 participants."

It is not true that the Review "ignored" this study's findings. The findings of Chen et al. (and other commonly cited studies) are extensively discussed in Chapter 6 of the Review (see Section 6.2.3 for discussion of Chen et al.). This study is also part of the umbrella review's analysis and contributed to the findings of very low certainty of evidence.

Although self-report scores for appearance congruence, depression, anxiety, and life satisfaction improved at 24 months compared to baseline,[32] these findings should not be described without serious qualifications. For example, males showed no improvement in any outcome except for "appearance congruence," which was measured, as pointed out in the Review, on a scale that has never been validated in

[32] The improvement in another measure, positive affect, from baseline compared to 24 months was not statistically significant. See the Review, footnote 46, p. 109.

minors. Originally, the Review reported that appearance congruence improvement was "the only statistically significant finding" but omitted the qualifying phrase "in males." The Review acknowledged the statistically significant improvements in females.

A *statistically* significant improvement, however, should not be confused with improvement that is *clinically* significant or meaningful.[33] In Chen et al., the mean Beck Depression Inventory score improved over 24 months from 16.01 to 13.85 (63-point scale); the mean Revised Children's Manifest Anxiety Scale improved from 59.84 to 57.32 (T-score, where 50 is the population average and 10 is one standard deviation); and the mean life satisfaction score improved from 40.03 to 44.68 (T-score) on a subscale of the NIH Toolbox Emotion Battery. These are small improvements of questionable significance to clinicians and patients. For the Beck Depression Inventory, for example, researchers have suggested a 17.5% decrease from baseline score may represent a "minimal clinically important difference."[34] The mean decrease of 2.16 points on this (63-point) scale in Chen et al. (2023) does not, according to this criterion, meet the minimal threshold of clinical importance.[35] Given the known and plausible harms of these interventions, even if such minor benefits were established via well-conducted studies (e.g., randomized controlled trials), the risk/benefit profile of hormonal interventions would remain unfavorable.

There are many other problems with Chen et al., including the fact that follow-up data for mental health outcome measures were unavailable for 31–34% of participants (introducing selection bias), the shifting hypotheses between the preregistered protocol and the publication, and the failure to report many preregistered outcomes such as suicidality, self-harm, and gender dysphoria. Most crucially, the uncontrolled observational design precludes any conclusion about whether cross-sex hormone (CSH) treatment *caused* any improvement. Nonetheless, that did not prevent Chen et

[33] A statistically significant improvement at the conventional $p<.05$ level simply means that the probability of an improvement as large as the one found, assuming the treatment has no effect (i.e. assuming that the "null hypothesis" is true), is less than 5%. This is compatible with the magnitude of the improvement being very small and of questionable clinical importance.
[34] Button et al., 2015.
[35] McDeavitt, 2024. It is unclear whether improvements for females meet this threshold (see Figure S3 in the supplementary appendix to Chen et al.).

al. from erroneously describing the results on the first page in explicitly causal language: CSH “improved appearance congruence and psychosocial functioning.” Contrary to Dowshen et al.’s assertion that the Review “focuses solely on the two deaths by suicide,” all of these problems were noted in the Review. Given points raised by a peer reviewer (Dr. Lane Strathearn) regarding the study’s elevated suicide rate, along with the fact that it continues to be uncritically cited as evidence for mental health benefits of CSH (see Part C of the reply to Rider et al.), Section 6.2.3 has been further clarified.

As Dowshen et al. do not engage with the substance of the critiques described above, it is impossible to know where or how they may disagree. Of note, two of the authors of Dowshen et al., Chen and Garofalo, are authors of Chen et al.

6. Dowshen et al. claim that, because Chen et al. (2023) purportedly had very positive findings and because “other cohort studies [report] improvements in psychosocial functioning after treatment …” the Review selectively misuses research.

Here, Dowshen et al. cite five additional papers in support of their claim about improved psychological functioning.[36] Pre-post studies in this field have shown inconsistent results with respect to psychological improvements (see Chapter 4). Their chosen references include Achille et al. (2020), in which (after regression analysis) depression improved only in males (the opposite finding from Chen et. al.); Chelliah et al., (2024), in which the reported improvement was a small decrease in the mean Quick Inventory of Depressive Symptomatology (QIDS) score from 10.7 to 8.2 (on a 42-point scale); and a study in which there was a small improvement in self-reported depression but not in the clinician-reported depression outcome measure (Kuper et al., 2020).

Achille et al. (2020), Kuper et al. (2020) and Dopp et al. (2024) (another of the five citations), were all included in the overview of SRs and contributed to the conclusion of very low certainty evidence. Chelliah et al. (2024) is also cited in the Review but appeared after the publication of the most recent SR included in the umbrella review;

[36] Achille et al. (2020); Dopp et al. (2024); Chelliah et al. (2024); Kuper et al. (2024); Olson-Kennedy, Wang et al. (2025).

we accordingly conducted an additional ROBINS-I V2[37] analysis of this study, which found it to be at critical risk of bias.[38]

The last of the five citations (Olson-Kennedy, Wang et al., 2025) is a paper reporting data derived from the same patient cohort as Chen et al. (2023). It is misleading to imply that this is one of the "other" cohort studies, when in fact it is the same cohort as that reported in Chen et al. (It is unclear why Trans Youth Care researchers are spreading the outcome data from this research project over multiple publications. Future systematic reviewers should be vigilant for "salami slicing."[39])

The problems in these studies illustrate why it is inappropriate to rely on low-quality observational studies when rigorous SRs are available. Dowshen et al.'s critique is an illuminating example of how the field of pediatric gender medicine often relies on an inverted hierarchy of evidence (see our reply to the APA's peer review, Figure 1). This inversion, not the Review's analysis, is what constitutes a "misrepresentation of the scientific evidence."

7. Dowshen et al. criticize the Review for omitting a reference to Nunes-Moreno et al. (2025).

Dowshen et al. are correct that this study was not included in the Review. Like Chelliah et al. (2025), Nunes-Moreno et al. appeared after the publication of the most recent SR included in the Review's umbrella review. The key question is whether this study merits reconsideration of the umbrella review's conclusion.

The study investigated the association between puberty blockers (PBs), cross-sex hormones (CSH), and suicidality among youth with gender dysphoria (GD), using the

[37] Cochrane (2024).
[38] See Appendix: ROBINS analyses.
[39] "Salami slicing" refers to the practice of "splitting data from the same research into small units, each of which is submitted—and in many cases published—separately." The practice may be "driven by an author's desire or need to achieve a larger number of publications, in order to gain recognition, move up on the academic career ladder, attract research funds by increasing the institution's visibility and/or obtain financial gain" (Karlsson & Beaufils, 2013). Salami slicing is problematic because it misleads readers (who may think each study represents a new data set) and creates a perception that the body of original research is larger than it really is.

PEDSnet electronic health record network. In Cox regression models,[40] CSH were associated with a significant reduction in suicidality risk (HR = 0.564, 95% CI 0.36–0.89), while PBs showed a non-significant trend (HR = 0.79, 95% CI 0.47–1.31).

Risk of bias assessment with the ROBINS-I V2 tool highlighted critical concerns, chiefly uncontrolled confounding from baseline mental health, family support, and cointerventions. Intervention classification, participant selection, and deviations from intended interventions were assessed as low risk of bias, but missing data, outcome measurement, and selective reporting were judged to be at serious risk. Taken together, the overall risk of bias for the CSH and PB results was assessed as critical, reflecting unresolved confounding and multiple serious risks across domains. (See Appendix, p. 177.)

Although Nunes-Moreno et al. leverages a large multicenter dataset, it has similar limitations as previously reported observational studies. Consideration of this study would not have changed the conclusion of prior SRs on PBs and CSH, nor the conclusion of Appendix 4's umbrella review. In evidence-based medicine, strength of evidence is determined by quality, not quantity, of studies.

8. Dowshen et al. claim the Review "provides no evidence for its assertion that puberty-pausing medications and hormone therapy are harmful to TGD youth," and that the Review "even states that evidence of harms is 'sparse'."

On the Review's alleged statement that evidence of harms is "sparse," Dowshen et al. have selectively quoted. The full quotation is:

> Evidence for harms associated with pediatric medical transition *in systematic reviews* is also sparse, but this finding should be interpreted with caution. (Executive Summary, emphasis added)

Chapter 6 of the Review explains why consideration of harms associated with PMT needs to go beyond evidence from SRs. As Guyatt and colleagues remark, "Many, if not

[40] A Cox regression model is a type of statistical model used to estimate how different factors affect the risk of some event occurring (in this case, an emergency department or inpatient visit for suicidality).

most, systematic reviews fail to address some key outcomes, particularly harms, associated with an intervention."[41]

For example, the fact that studies have not reported infertility data (and therefore SRs have been unable to capture it) does not mean infertility can be ignored in a comprehensive evidence appraisal. Far from "[providing] no evidence" of harms, Chapter 7 of the Review presents detailed indirect evidence derived from basic science, endocrinology, and developmental physiology, demonstrating plausible and biologically expected harms. Dr. Richard Santen, former president of the Endocrine Society and one of the Review's peer reviewers, found it to be "scientifically valid" and to "reasonably reflect an overview of the information currently available and its interpretation." Because Dowshen et al. do not engage with the substance of the findings in this chapter, it is impossible to know how or why they may disagree.

9. Dowshen et al. assert that the Review's comments on the lack of long-term outcome data are "misleading." They mention two Dutch studies providing "over 20 years of follow-up data," and an American study providing "up to 10 years" of follow-up data.

The Dutch studies are both cited in the Review, specifically with respect to rates of continuation from PBs to CSH and to bone mineralization outcomes.

The first study, van der Loos, Klink et al. (2023), evaluated treatment trajectories. It is misleading to describe this study as a supplying "20 years of follow-up data." Here, 20 years refers to the *intake period* (1997 to 2018). With respect to follow-up after hormone initiation, the study's median follow-up was 4.6 years.

The second Dutch study, van der Loos, Vlot et al. (2023), evaluated bone density of 25 males and 50 females treated with PBs followed by CSH. As discussed in Chapter 7 of the Review, this study found that Z-scores returned to pre-treatment baseline by median age 28 in females, but that in males, Z-scores at the lumbar spine remained below pre-

[41] Guyatt et al. (2011, p. 397).

treatment baseline at follow-up.[42] A major limitation of this study is the 40% non-participation rate.

The American study, Olson et al. (2024), assessed satisfaction and regret after initiation of hormones. It is misleading to describe this as a 10-year follow-up, as the mean follow-up was 4.86 years after starting PBs and 3.4 years after starting CSH (median follow-up was not reported and may have been considerably lower). Notably, the study may not be representative, as participants in this cohort were completely socially transitioned before puberty (some were as young as age two at the time of social transition; average age at social transition was 6.49[43]). Further, the absence of physiological or psychiatric outcome data in this study is a critical limitation.

As the Review explains (Section 13.4), satisfaction and regret, though important data points, are not valid proxies for evaluating the justification for PMT. A Letter to the Editor responding to Olson et al. (2024) points out that "Patient satisfaction is generally considered a complementary measure of health care quality and is typically assessed after the safety and effectiveness of the intervention are established."[44] Because Dowshen et al. never engage with our analysis on this point, it is impossible to know where or how they disagree with our conclusions.

C. Safety of PMT

10. Dowshen et al. suggest that PBs are safe for "TGD youth" because "they have been safely and effectively used for decades to treat cisgender youth with medical conditions such as precocious puberty."

Chapter 7 of the Review contrasts the use of PBs in treating central precocious puberty (CPP) with their use for GD. There are three key differences.

[42] A bone mineral density Z-score is a measure that compares individual bone density to age- and sex-matched population norms.
[43] deMayo et al. (2025, p. 39).
[44] Sinai et al. (2025).

First, *purpose*: CPP is a physical pathology and PBs are used to stop abnormally timed puberty. GD is not a physical pathology and PBs in this case are used to stop normally timed puberty.

Second, *diagnosis*: CPP is diagnosed using objective tests such as blood work, and the natural history of the condition is well-understood, whereas the diagnosis of GD relies on subjective criteria and has poor predictive validity.

Third, *prognosis*: In CPP cases, PBs are stopped and puberty resumes. For pediatric GD cases, over 90% of youth treated with PBs proceed to CSH, and for these patients puberty (properly defined[45]) does not resume.

The senior author of Dowshen et al., McNamara, recently acknowledged that PBs should not be assessed as a standalone intervention, but rather as a component of a single treatment modality comprised of both PB and CSH.[46] By her own admission, then, it makes little sense to assert that PBs are safe for use in GD on the grounds that they are safe for use in CPP. The combined-use pathway (PBs followed by CSH) presents a fundamentally different risk profile. For example, infertility is not an expected treatment outcome when PBs are used for CPP, but it is an expected outcome when PBs are used prior to or alongside CSH for pediatric GD.

We also note that Dowshen et al.'s description of CPP patients as "cisgender youth" is inaccurate. Any child diagnosed with CPP will be a candidate for PBs, irrespective of how he or she identifies. CPP and GD are two distinct clinical scenarios and the implication that concerns about PB use in one scenario but not the other are due to identity-based discrimination is seriously misleading.

11. Dowshen et al. cite "a recent comprehensive review commissioned by the Utah state legislature and completed by experts at the University of Utah," which "concluded

[45] Critics may object that puberty *does* resume when patients proceed from PBs to CSH, but as we explain in the Review (Section 2.1, footnote 6), "it is misleading to suggest that the 'right puberty' induced by estrogen in males or testosterone in females amounts to a cross-sex version of puberty, because puberty centrally and definitionally involves maturation of the capacity for reproduction."
[46] U.S. v. Skrmetti (No. 23-477), Expert Researchers and Physicians, Amicus curiae brief (2024, pp. 16-17).

that puberty-pausing medications and hormone therapy can also be used safely in TGD youth."

The "Utah Review"[47] cited by Dowshen et al. was published after the HHS Review and therefore was not included in its analysis. We discuss the Utah Review in our reply to the APA (where we note that this review did not synthesize evidence or assess its quality, and therefore does not qualify as an SR). Because the Utah Review has been cited in peer-reviewed publications (e.g., Dowshen et al.) and popular media, we have included a formal methodological appraisal in the Review (see Section 5.7.3 and Appendix 4), which finds the Utah Review to be at high risk of bias in all domains.

D. Gender identity and the charge of "conversion therapy"

12. Dowshen et al. say that a "central premise" of the Review "is the unsupported claim that gender identity among adolescents is inherently unstable."

First, the Review is concerned with gender dysphoria, not "gender identity," and it did not adopt any claim about the stability of the latter as a "central premise." Second, as discussed in the Review, the assumed permanence of adolescent (in contrast to childhood) GD has served as the basis for the Dutch Protocol, but this assumption was never based on credible evidence. In the Review, we discussed other, more recent evidence suggesting that for a significant number of children and adolescents, gender dysphoria appears to be a transient phenomenon. Other reviewers (Santen and Dr. Lane Strathearn) have pointed out the importance of recognizing the limitations of the research in this area. We address this in our response to Santen.

The key point to consider here is that the burden of proof for PBs, CSH, and surgeries as treatments for adolescent GD—a mental health condition—rests on those advocating for these interventions. If there is no credible evidence for the permanence of adolescent GD or for the safety and efficacy of PMT, and tentative evidence that the diagnosis is unstable in many or most adolescents, the precautionary principle (Chapter 13) applies.

[47] University of Utah College of Pharmacy, Drug Regimen Review Center (2025).

13. Dowshen et al. object to the Review's broadly positive treatment of ("exploratory") psychotherapy, claiming that it is "an ill-defined practice that aims to change a young person's identity, which is akin to conversion therapy." "Decades of evidence," they say, "demonstrate that conversion practices are both ineffective and dangerous for the psychological health of transgender [youth]." Dowshen et al. allege that the Review includes a "recommendation" for "nonevidence-based conversion practices."

The claim that the Review recommends conversion practices is false. Unfortunately, Dowshen et al. fail to engage with Section 14.5.2.1, which anticipates and refutes this charge. Instead, they repeat an earlier accusation of their co-author, Dr. Kellan Baker, that the Review "pushes the dangerous and discredited practice of conversion therapy to try to force transgender people to change a fundamental, deeply rooted part of who they are."[48] Repetition of claims is no substitute for substantive engagement. Moreover, Dowshen et al.'s chosen citations are manifestly inadequate.[49]

Prominent PMT advocate Dr. Jack Turban, director of the Gender Psychiatry Program at UCSF, has said that "conversion efforts and exploratory psychotherapy are distinct, mutually exclusive practices."[50] And even WPATH—the leading organization supporting PMT—said, in a statement condemning the HHS Review, that "[they] unequivocally oppose" "[equating] conversion therapy with psychotherapy" for "youth who are exploring their gender identity."[51]

---

[48] Riedel (2025).

[49] Substance Abuse and Mental Health Services Administration (2023); see especially p. 27. Setting aside methodological problems with the few studies on "gender identity change efforts" cited therein, they have no bearing on the psychotherapeutic approaches described in Chapter 14. Psychotherapy, as described in this chapter, aims to provide support, mitigate psychological distress, facilitate self-understanding, and improve patients' quality of life and interpersonal relationships. All psychotherapy is "exploratory"; modalities described in Chapter 14 that may be helpful for this population include cognitive behavioral therapy, dialectical behavioral therapy, psychodynamic psychotherapy, and family therapy.

Dowshen et al. cite Ashley (2023) to support the incorrect claim that "exploratory therapy is an ill-defined practice that aims to change a young person's identity." That paper is cited in Chapter 14 of the Review as an example of the concerning trend of denigration/mischaracterization of psychotherapy, which itself is further described in Section 14.5.2.1.

[50] Chiles v. Salazar (No. 24-539), Dr. Jack L. Turban and Dr. Lisa R. Fortuna, Amicus curiae brief (2025, p. 15).

[51] World Professional Association for Transgender Health & United States Professional Association for Transgender Health (2025).

We also note that the APA did not raise any objections to the Review's psychotherapy chapter in their peer review.

E. Guidelines and clinical practice

14. Dowshen et al. criticize the Review for its discussion of deteriorating standards and the collapse of medical safeguarding in the U.S. They object to the inclusion of "unverified" whistleblower accounts, characterizing the whistleblowers as "individuals not directly involved in clinical decision-making for patients."

Contrary to Dowshen et al.'s characterization, most of the whistleblowers (Chapter 11) are practicing clinicians who have treated this patient population. Their accounts are highly relevant for understanding the clinical realities of pediatric gender medicine in the United States. Whistleblowers play a vital role in upholding U.S. healthcare safety and patient protection standards.

15. Dowshen et al. further suggest that the whistleblower accounts should be discounted because WPATH and ES guidelines require "an interdisciplinary team that performs a comprehensive biopsychosocial assessment" prior to initiation of PBs/CSH or referral for surgery.

The problem is that accounts of whistleblowers describe clinicians offering inappropriate treatments *within* the context of "multidisciplinary" (or "interdisciplinary") teams and "assessments" (see Chapter 11 and Section 14.3).

Likewise, "assessments" may be cursory or perfunctory,[52] and some advocates for PMT view the requirement for any mental health assessment, however brief, with skepticism. As one author of Dowshen et al. has put it, "If the medical provider thinks they have the answer [to whether medical interventions are appropriate] then they're the wrong medical provider. The answer lies within the young person and the family."[53] Rejecting the notion that minors should be required to undergo assessment, Dr. Johanna Olson-Kennedy, another prominent gender clinician, explained that "We don't actually have

[52] See Section 11.3.3 of the Review.
[53] Figliola (2025).

data on whether psychological assessments lower regret rates," and that "I don't send someone to a therapist when I'm going to start them on insulin."[54]

The involvement of multiple professionals therefore does not guarantee that the etiologies of patients' gender-related distress are explored, nor that the possibility that this distress may resolve with time or through less invasive means is adequately considered. It is noteworthy that Dowshen et al. do not engage with the Review's lengthy description and analysis of these issues.

16. Dowshen et al. assert that guidelines which recommend PMT as the standard of care (e.g., WPATH and Endocrine Society guidelines) are "informed by the best available evidence, which demonstrates improved outcomes in mental health, psychological well-being, and suicidality." Furthermore, Dowshen et al. claim the treatment approach recommended in WPATH's guideline is "evidence-based" and "based on more than 70 systematic reviews."

The best available evidence (i.e., from SRs) does not support Dowshen et al.'s assertion that psychological outcomes improve with PMT. Appendix 4's umbrella review reveals that the effects on psychological outcomes are unknown.

Chapters 9–11 of the Review identify serious problems with WPATH and Endocrine Society (ES) guidelines. With respect to the care of children and adolescents (Chapters 6 & 7), WPATH's guideline is consensus-based, not evidence-based, as it is not based on evidence from SRs. Indeed, *Standards of Care,* Version 8 (SOC-8) states—falsely—that an SR of hormonal interventions in minors is "not possible."

Section 10.3.2 details how WPATH suppressed the publication of some systematic reviews it had commissioned to inform SOC-8, including reviews covering treatment of minors. This raises serious concerns about the scientific integrity of WPATH as a self-described healthcare organization. Unfortunately, Dowshen et al. never engage with these revelations or their significance. (Of note, one of our reviewers, Dr. Richard Santen, a former president of the Endocrine Society, encouraged us to add a section

[54] Singal (2018).

about problems in the development of that Society's gender medicine guidelines, which we did.)

17. Dowshen et al. describe the WPATH/ES clinical practice guidelines (and other guidance that reference these) as "the existing standard of care" and imply that WPATH's guidelines should be considered trustworthy in part because they are "widely endorsed," "maintained since 1979," "currently in their eighth edition," "took almost a decade to develop," and represent "the consensus recommendations of more than 100 experts in transgender health."

We would like to clarify that there is no accepted "existing standard of care" for treating pediatric patients with GD, and that guidelines/policies from around the world recommend very different treatment approaches.

With respect to Dowshen et al.'s list of attributes, none of them is recognized as relevant to a guideline's trustworthiness. AGREE II, a widely used tool to assess trustworthiness of clinical practice guidelines, specifies six relevant domains: Scope and Purpose, Stakeholder Involvement, Rigor of Development, Clarity of Presentation, Applicability, and Editorial Independence. The methodological rigor of a guideline's development—specifically, whether it relied on SRs rather than expert consensus—is regarded as the most important of these domains.[55] Medicine should be evidence-based, not eminence-based. Dowshen et al. do not dispute any of the factual findings regarding WPATH SOC-8's development (described in Chapter 10).

[55] Hoffmann-Eßer et al. (2018).

Reply to Rider et al.

In October 2025, *Sexuality Research and Social Policy* published a commentary titled "Scientific integrity and pediatric gender healthcare: Disputing the HHS Review" (Rider et al., 2025), which asserts:

> Although the HHS Review has a different tone than the Executive Order that directs it, the HHS Review presents the White House's political agenda as objective science, relying on misleading evidence and data to advance its aims.

The commentary's fifteen authors include several prominent gender clinicians.[56]

As with Dowshen et al. (2025), we have decided to treat Rider et al. (2025) as unsolicited peer review.

Part A of this reply summarizes our responses to criticisms in Rider et al. that appear in other peer reviews and publications, namely Dowshen et al. and the peer review by the American Psychiatric Association (APA). Part B responds to the few points that do not appear in the other peer reviews. Part C discusses how Rider et al.'s commentary exemplifies problems common in this field: misrepresentation of research, inadequate citation practices, and poor understanding of evidence-based medicine (EBM) principles. The continued willingness of peer-reviewed journals to publish demonstrably false or misleading claims about evidence-based medicine and medical practices concerning child and adolescent health is deeply regrettable.

A. Critiques previously addressed in replies to Dowshen et al. and/or the APA

1. Rider et al. object that "the authors [of the Review] were unnamed."

As HHS initially explained, the identities of the contributors were temporarily withheld to "help maintain the integrity of this [peer review] process." Withholding names in peer-

[56] Most of Rider et al.'s authors are affiliated with the Eli Coleman Institute for Sexual and Gender Health at the University of Minnesota, a leading gender medicine clinic that is described on its website as "one of the largest clinical, teaching, and research institutions in the world specializing in human sexuality and gender" (University of Minnesota, 2025). The American psychologist Dr. Diane Ehrensaft, Director of Mental Health at UCSF's Child & Adolescent Gender Center, is also an author. Ehrensaft, who is heavily cited throughout the HHS Review, is one of the world's leading pediatric gender clinicians, and is responsible for pioneering the "child-led" approach to pediatric gender medicine in the United States. (See Section 11.1.1 of the Review.)

review is standard practice in academic publishing. Please see point 2 in our reply to Dowshen et al., as well as Section 4 of our reply to the APA.

2. Citing Dowshen et al. (2025), Rider et al. object that “more than a fifth” of the Review’s references “are from popular media articles, blogs, or social media.”

To repeat part of our reply to Dowshen et al. (point 4), focusing on the ratio of peer-reviewed to non-peer-reviewed sources is misguided. The important question is whether all relevant evidence is appropriately represented. The Review was tasked with evaluating both evidence and best practices. In addition to engaging with the relevant peer-reviewed scientific literature, the Review discusses non-peer-reviewed publications where appropriate. The Review’s appraisal of best practices examines clinical realities in the U.S., which are frequently documented in court filings and media outlets such as *The New York Times*, *The Washington Post*, *The Boston Globe*, *Reuters*, and *The Economist*.

3. Rider et al. criticize the Review’s provenance (i.e. the January 2025 Executive Order directing the Secretary of Health and Human Services to commission the Review), describing the Review as an “[entity] with a political agenda targeting [patients, families, and providers]” and claiming that the Review “presents the White House’s political agenda as objective science.”

The Executive Order directed HHS to “publish a review of the existing literature on best practices for promoting the health of children who assert gender dysphoria.”[57] If the findings of the Review were dictated by preexisting political agendas, it should be possible to identify errors within it. As explained in our response to Dowshen et al. and the APA, and as further demonstrated below, no such examples have been offered. Please see point 1 in our reply to Dowshen et al., as well as Section 4 of our reply to the APA.

4. Rider et al. characterize pediatric medical transition (PMT) (three times) as “medically necessary,” claiming that “scientific evidence demonstrat[es] its safety

[57] The White House (2025).

and effectiveness in improving short- and long-term health outcomes for TGNB [transgender and nonbinary] adolescents."

Rider et al. repeatedly exhibit a misunderstanding of basic EBM principles regarding quality (or certainty) of evidence. It is simply incorrect that scientific evidence "demonstrates" PMT's "safety and effectiveness." The Review's umbrella review (Appendix 4) shows this quite clearly, and Rider et al. say nothing that casts doubt on its findings. Please see point 6 in our reply to Dowshen et al., regarding the inversion of the evidence hierarchy, and Section 1 of our reply to the APA, regarding the list of individual studies provided in their peer review. We discuss other examples of Rider et al.'s misunderstanding of EBM in points 8 and 9 below.

5. Rider et al. criticize the Review's engagement with the Cass Review. They claim the Review did not address the Cass Review's alleged "omission of key findings from the broader literature," or the fact that it has been "negatively critiqued and challenged repeatedly by professional organizations and individual experts in the field of pediatric gender care." They also claim the HHS Review selectively quotes from the Cass Review's conclusions.

The Cass Review's findings have been accepted by both major political parties in the U.K. and its recommendations are being implemented by the U.K.'s National Health Service. It is not surprising that gender clinicians and the professional associations that represent them would disparage a review that upended their favored treatment model in the U.K.

We direct readers to comprehensive rebuttals to critiques of the Cass Review: see footnote 77 in Part C below and Section 3 of our reply to the APA, which also addresses Rider et al.'s claim that the Review selectively quotes from the Cass Review. In short, the critiques are rife with demonstrable falsehoods and some appear motivated by legal goals rather than scientific ones. It is also worth noting that one of the references Rider et al. provide (Horton, 2024) in support of their claim that "thorough scientific and legal scholarship, as well as the critiques from field experts ... directly rebut the evaluation of evidence..." in the Cass Review was published a month *before* the final Cass Review was published.

6. Rider et al. claim the Review "promotes a harmful practice known as 'exploratory therapy' ... which has been argued to be a form of conversion therapy encouraging a child or adolescent to accept the gender associated with their sex designated to them at birth."

Here, Rider et al. rely on the work of lawyer Florence Ashley, who opposes requirements for mental health assessments prior to PMT initiation and advocates for the wide availability of PMT because it facilitates a minor's "gender embodiment goals."[58] Ashley has repeatedly conflated psychotherapy for pediatric GD with conversion therapy.[59] In addition, Rider et al. attempt to bolster their case with misleading citations. One reference, the United States Joint Statement, explicitly states that "Exploration of issues pertaining to gender identity and sexual orientation in a way that does not favor or presume a particular identity or experience, would *not* be considered conversion therapy."[60] Another reference is a United Nations report which concludes that conversion therapy "may constitute torture."[61] However, that report's examples of conversion therapy include gay individuals being "blindfolded and pummeled with basketballs, bound with duct tape, rolled up into blankets and subjected to anti-gay slurs." These and similar practices are indeed abhorrent, but they have no bearing whatsoever on talk therapy for minors with gender dysphoria (GD).

Please see also point 13 in our reply to Dowshen et al.

7. Rider et al. compare hormonal interventions for pediatric GD to the use of hormonal interventions for other pediatric medical conditions—e.g., central precocious puberty (CPP)—in "prepubescent and pubescent cisgender youth," suggesting that "TGNB adolescents" are being unfairly singled out.

This framing is extremely misleading. To summarize point 10 in our reply to Dowshen et al., Chapter 7 of the Review contrasts the use of puberty blockers (PBs) in treating CPP with their use for GD. When used for CPP, PBs arrest abnormally timed puberty (as

[58] See the Review, Section 13.3.
[59] See the Review, Section 14.5.2.1.
[60] United States Joint Statement (2023).
[61] Madrigal-Borloz (2020).

opposed to normally timed puberty), and are not followed by administration of CSH, which may result in lifelong infertility and sexual dysfunction as well as other risks to health.[62] Also, like Dowshen et al., Rider et al. inaccurately describe CPP patients as "cisgender youth." Any child diagnosed with CPP will be a candidate for PBs, irrespective of how they identify. CPP and GD are two distinct clinical scenarios, and it is entirely wrong to suggest that concerns about PB use in one scenario but not the other are due to identity-based discrimination.

B. Novel points in Rider et al. (2025)

8. Rider et al. criticize the Review for "omitting" context related to the ambient "political climate and proposed or existing legislative bans on GAMC [gender-affirming medical care] for TGNB adolescents and their caregivers." Further, they claim this "distorts the application of evidence-based medicine."

This criticism seems intended to deflect from the content of the Review by putting the focus on the ambient "political climate." EBM is concerned with clinical decision-making based on the best available evidence for the safety and efficacy of treatments, not with the broader political climate or legislation. We agree that the political climate has made scientific debate very difficult, but we emphasize again that if errors appear in the Review, it should be possible to clearly identify them. Rider et al. identify no errors.

9. Rider et al. claim the Review "overlook[s] bias in the systematic reviews [SRs] it cites and deemphasiz[es] other layers in the 'hierarchy of evidence.'" According to Rider et al., "multifaceted data and studies across multiple levels in the 'hierarchy of evidence' comprise the robust body of evidence supporting [PMT]."

With respect to alleged quality problems in SRs of PMT that have found very low certainty evidence, Rider et al. reference an analysis critiquing the Cass Review, Noone et al. (2025), that claimed two specific SRs were biased/flawed.[63] Using the Risk of Bias Assessment Tool for Systematic Reviews (ROBIS), the Review came to a different conclusion than Noone et al., finding that these two SRs on puberty blockers and cross-

[62] As the Review notes, data suggest that the vast majority of those on PBs continue to CSH (Section 4.3.2.2).
[63] Taylor, Mitchell, Hall, Heathcote et al. (2024); Taylor, Mitchell, Hall, Langton et al. (2024).

sex hormones were generally at *low* risk of bias. Further, even if those two SRs were excluded, the conclusion that the quality of the evidence for benefit of PMT is very low certainty would be unaffected.[64] We refer Rider et al. to the peer review included in this Supplement by methodologists Dr. Trudy Bekkering and Professor Patrik Vankrunkelsven, which concluded that the Review's umbrella review was conducted appropriately. If Rider et al. disagree with the Review's analysis, it would have been helpful for them to explain why. Merely citing Noone et al. does not advance scientific understanding.

Rider et al.'s claim that the Review "deemphasiz[es] other layers in the 'hierarchy of evidence'" seems to be a suggestion that it should have ignored or minimized the findings of quality systematic reviews in favor of emphasizing conclusions reached by some individual studies ("other layers"). Doing so, however, would constitute an inversion of the hierarchy of evidence and a violation of a core principle of EBM. Low quality evidence of the kind favored by Rider et al. cannot be characterized as "robust."

10. Rider et al. say that the Review "fails to acknowledge … that most pediatric healthcare is guided by evidence of similar quality and strength as that supporting [PMT]."

It is not true that "most pediatric healthcare is guided by evidence of similar quality and strength." Dr. Hilary Cass, author of the Cass Review and a past president of the Royal College of Paediatrics and Child Health, observed that the evidence for the efficacy of PMT is very weak, even compared to other areas of pediatric medicine.[65] But we need not appeal to the authority of Cass: her judgment is supported by Rider et al.'s *own citation* (Matheny Antommaria et al., 2025). Matheny Antommaria et al. analyzed 14 current pediatric clinical practice guidelines, finding that 58% were based on Level A or Level B evidence: "Level A evidence includes well-designed and -conducted randomized controlled trials; Level B randomized controlled trials with minor limitations

[64] The two SRs represent only a fraction of the seven low risk of bias SRs that contributed to the umbrella review's evidence synthesis (Rider et al. mistakenly refer to it as a meta-analysis) with respect to PMT (PBs, CSH, and surgeries). Five other English-language PMT SRs were found to be at low risk of bias: Dopp et al. (2024); Ludvigsson et al. (2023); Miroshnychenko et al. (2024); Miroshnychenko, Ibrahim et al. (2025); Miroshnychenko, Roldan et al. (2025).
[65] Ghorayshi (2024).

or consistent evidence from multiple observational studies."[66] PMT is not supported by evidence at these levels: no randomized trials have been conducted, and the extant observational studies are generally low-quality.[67] Therefore, Rider et al.'s claim that "most" pediatric healthcare is supported by evidence of a similarly low quality as that supporting PMT is false.

There is a more important point. Quality of evidence, as assessed via a rigorous systematic review, can inform stakeholders regarding what is known about an intervention's effectiveness. But it is not the only consideration in clinical decision-making for PMT or for any other intervention. Harms must also be considered, as well as the natural history of the condition and the risk/benefit profiles of alternative treatment options. Please also see Section 2 of our reply to the APA.

Furthermore, Rider et al.'s discussion of the evidence verges on inconsistency. In one passage they refer to "substantial evidence of benefits" of PMT, implying there are studies furnishing high-quality evidence. But two paragraphs earlier they apparently concede that "most pediatric healthcare" and PMT both "[fall] short in the 'hierarchy of evidence'." Rider et al. cannot have it both ways.

11. The Review discusses "rapid onset gender dysphoria" (ROGD); according to Rider et al., this is a "largely discredited diagnosis."

The Review addresses the recent surge in adolescent females with GD and various attempts to explain the novel development of gender dysphoria in this clinical population. "ROGD" is simply a label for a new clinical phenomenon; contrary to Rider et al.'s assertion, ROGD was never presented as a "diagnosis." For a discussion, we refer Rider et al. to Section 4.3.1.4 of the Review. Rider et al. give a citation to support the claim of "discreditation"; this is discussed in Part C below.

12. Rider et al. claim that "the HHS review likens the field of GAC ["gender-affirming care"] to the Tuskegee syphilis study" and furthermore claim that this is not a legitimate comparison because parents provide consent for interventions that cause

[66] Matheny Antommaria et al. (2025, p. 2; see also Table 2).
[67] See Ludvigsson et al. (2023) for concrete suggestions on how the quality of observational research could be improved.

infertility, etc. in their assenting children, whereas the Tuskegee participants did not provide informed consent.

This is a misreading. Rider et al. cite Section 13.2.4 of the Review, which merely observes that the Belmont Report was "published in 1978 in the wake of the U.S. Public Health Service's Untreated Syphilis Study at Tuskegee."[68] No comparison between Tuskegee and "the field of GAC" is made or implied.

The Review does reference Tuskegee one other time but in a different chapter. Advocates of PMT sometimes object to scrutiny of the practice on the grounds that the number of minors undergoing these interventions is relatively small. Tuskegee is cited in Section 11.2 as an example of a medical experiment widely recognized as profoundly unethical despite involving a relatively small number of people (and much smaller than the number of youth receiving PMT).

We also note here that Dr. Steven Williams, past president of the American Society of Plastic Surgeons (ASPS),[69] invoked Tuskegee in a 2024 interview with Dr. Blair Peters, a plastic surgeon who performs gender surgeries. Peters claimed that if there were valid concerns about PMT, "physician groups providing it [would] be the first ones to raise the alarm and stop it." Williams disagreed: "'Assuming that doctors always do the right things—that's probably not the right assumption either.' Then, referring to his own racial identity, he said: 'In all honesty, again: *Black man*. So, you know …Tuskegee experiments, those types of things. Those were doctors. They were doing terrible stuff.'"[70]

C. Rider et al. (2025) as an instance of general problems

Rider et al.'s commentary—in its aim, tone, and content—exemplifies serious, pervasive and continuing problems in the field of pediatric gender medicine.

[68] The page number given by Rider et al. (226) is to the first version of the Review, published on May 1. The corresponding number in the May 15 version is p. 223.
[69] The ASPS does not endorse PMT.
[70] Ryan (2024).

First, the commentary employs inflammatory rhetoric. Rider et al. allege that the Review has "little regard" for the "civil rights" of vulnerable youth[71] and conflates psychotherapy for pediatric GD with conversion therapy, which it describes as akin to "torture."

It is inappropriate for a peer-reviewed journal to publish such extremely serious allegations, which impugn not only the Review but the moral character of its contributors, without evidence to support them. It is also unusual to see a peer-reviewed journal allow performative expressions such as "we condemn," which may be appropriate for political tracts but not scientific discussion.

Second, as discussed above, Rider et al. build their argument against the Review upon a variety of informal fallacies and suspect reasoning. Examples include genetic fallacies (e.g., judging the Review's content and conclusions based on the Executive Order that commissioned it), red herrings (e.g., purporting to critique the Review while instead focusing on political or legislative issues), and appeals to authority (e.g., claiming that PMT is beneficial because some medical organizations say so). Additionally, Rider et al. levy ad hominem attacks, seemingly implying that the Review's contributors—presumably unknown to Rider et al. at time of writing—are "unqualified individuals with no expertise in the field of pediatric gender care." (Assuming that the relevant "expertise" here is treating gender dysphoric adolescents in clinical settings, see Section 10.3.1 of the HHS Review and point 1 in our response to Dr. Richard Santen for discussions of conflicts of interest.) Such rhetorical tactics are depressingly commonplace in this field.[72]

---

[71] Advocates for PMT have long framed it as a matter of civil rights. See Section 12.2. of the Review, which describes how Dr. Diane Ehrensaft, an author of Rider et al., conceptualizes her work in pediatric gender medicine as having "finally created a civil rights movement."

[72] Another example is McNamara, Abdul-Latif et al. (2022), an essay which criticized an umbrella review of PMT commissioned by the state of Florida. McNamara et al. referred to (then) assistant professor Romina Brignardello-Petersen, who conducted the review, as someone whose "only clinical experience appears to be in dentistry," and compared this to asking "dermatologists to conduct a review of the scientific literature on neurosurgery" (p. 10). The clear suggestion was that Brignardello-Petersen was unqualified; however, her PhD in clinical epidemiology and health care research was not mentioned. (Clinical epidemiology is epidemiology used to inform clinical decision making and is the core of evidence-based medicine.)

McDeavitt et al. (2025) found that four papers critiquing the Cass Review exhibited similar problems, "making explicit and implicit claims about the professionalism of the Cass Review team and other researchers ... terms such as 'pseudoscience' and 'debunked' were used to describe contemporary peer-reviewed research and cogent hypotheses ... In some cases, an author and an author's professional

Rider et al.'s decision to substitute scholarly engagement with appeals to the authority of U.S.-based medical organizations is especially unfortunate given that the Review devotes entire sections to presenting evidence of how these organizations have misled their members, patients, and the public.[73] Rider et al. never dispute any of this evidence.

Another tactic used by Rider et al. is uncritical citation of a denunciatory statement as evidence that something—in this case, ROGD—is "discredited."[74] This is part of a widespread pattern of citation problems in the pediatric gender medicine literature. A statement that itself contains no bibliography or hyperlinks to relevant literature is cited as factual; this is an example of "dead-end referencing."[75] Rider et al. also employ selective citation, as when, in discussing critiques of the Cass Review, they omit any reference to the literature that has carefully rebutted the main allegations made in these critiques.[76,77]

In several places Rider et al. illustrate an observation made in Section 2.1 of the Review, that this field has "a mode of communicating that is scientifically ungrounded, that presupposes answers to ethical controversies, and that is in other ways misleading." Rider et al. take for granted, for example, that children "as young as five years old" may be "transgender or nonbinary," as if these categories could unproblematically be applied to an age group that has a rudimentary understanding of sex differences in terms of stereotypes.[78] Presumably Rider et al. are following the lead of one of their co-authors, Dr. Diane Ehrensaft, who has taught that toddler actions,

---

organization were emphasized more than or instead of the contents of the respective articles ... Additionally, negative characteristics [were imputed] to the [Cass] Review team."

[73] Chapters 9–12 of the Review scrutinize the origins of the purported medical consensus in the U.S (see Sections 12.1–12.2, especially).

[74] Coalition for the Advancement and Application of Psychological Science (2021).

[75] Coverdale et al. (2024).

[76] Omission of important references can occur when "authors make statements that close off an area of controversy by citing only one side ... authors may note that there are multiple studies on a topic yet cite only the one that supports their thesis, while more methodologically rigorous and less dated studies with contrary positions or findings are left out" (Coverdale et al., 2024).

[77] Rebuttals: Cheung et al. (2025); Kingdon et al. (2025); McDeavitt et al. (2025). These publications have found critiques of the Cass Review to be full of errors. Baxendale (2025, p. 10) also comments on one of these critiques (McNamara et al., 2024).

[78] Halim et al. (2017).

such as removing hair barrettes or unbuttoning onesies, can be "pre-verbal communication[s] about gender."[79]

The "Woozle effect" is the frequent citation of an inadequate source to support a particular claim; this can create the illusion that the source is authoritative.[80] The Woozle effect plagues gender medicine research and advocacy, and makes an appearance in Rider et al. For instance, Rider et al. cite Chen et al. (2023) as part of the "scientific evidence demonstrating … improvements in well-being and quality of life." Here Rider et al. reproduce the words of Budge et al. (2024), which also cites Chen et al. as part of "existing research [which] demonstrates the effectiveness of [PMT]," through "improvement in well-being and quality of life." Likewise, Dowshen et al. (2025) say that Chen et al. found improvements in "appearance congruence, positive effect [sic], life satisfaction, and depression and anxiety symptoms." None of these three papers mentions the elevated suicide rate in Chen et al.; the missing outcome measures; the improvements of questionable clinical importance in female patients; the fact that male patients did not improve in mental health, only in "appearance congruence" (on a scale that has not been validated in minors); or the alteration of the study's central hypotheses between the written protocol and the published paper.[81]

---

[79] The quotation is from a 2016 lecture on the "gender affirming" model (Ehrensaft, 2016, 2:07:55). Prior to pioneering the child-led "gender-affirming" treatment model that is now dominant in the U.S.—see Review, Sections 11.1, 11.3—Ehrensaft was an expert in issues of alleged preschool ritualistic satanic abuse, and uncritically accepted children's accounts of those events (Ehrensaft, 1992). In recent years psychologists' role in the "Satanic Panic" controversies of the 1980s and 90's has been heavily criticized (e.g., Yuhas, 2021).

[80] Woozle effect (2025). A classic example is a letter that appeared in the *New England Journal of Medicine* in 1980, claiming that the risk of addiction from narcotics (opioids) is very low. This "five-sentence letter … was heavily and uncritically cited as evidence that addiction was rare with long-term opioid therapy" and may have contributed to the opioid crisis (Leung et al., 2017).

[81] More examples of Chen et al. citations: it is one of the studies cited to support PMT's association with: "mental health benefits and decreased suicidality" (Borah et al., 2023); "significant improvements in depression, anxiety, positive affect, and life satisfaction" (Huit et al., 2024); "improvements in anxiety, depression, and body image" (Olson et al., 2024); "a range of positive outcomes and lower rates of negative outcomes such as suicidality" (Twenge et al., 2025). Restar (2023) is a particularly egregious example: after citing Chen et al. for "positive affect and life satisfaction, and decreases in depression and anxiety symptoms," Restar then says, "*Notably*, this study also reported a total of 3.5% suicidal ideation—a comparable rate to the U.S. general population rate …" (emphasis added). (One issue is that any "comparison" is compromised by the fact that Chen et al. do not explain how "suicidal ideation" in their patients was measured.) What Restar failed to note was the second part of the relevant sentence in Chen et al.: "The most common adverse event was suicidal ideation (in 11 participants [3.5%]); death by suicide occurred in 2 participants."

Please see also point 5 in our reply to Dowshen et al., and Sections 4.3.4 and 6.2.3 of the Review.

Rider et al. also twice cite Tordoff et al. (2022), another influential study exemplifying the Woozle effect. Tordoff et al. purport to show reductions in psychiatric morbidity following the provision of hormonal PMT interventions. According to Google Scholar, this study has been cited over 700 times—more than 200 times per year on average since it was published—despite the fact that an online supplementary table of the paper reveals no statistically significant improvement in patients receiving the interventions. The Review discusses Tordoff et al. in detail in Section 6.2.2.

As researchers who have carefully observed pediatric gender medicine for years, we fully expect the Woozle effect to apply to Rider et al. (2025) and Dowshen et al. (2025), which will almost certainly be uncritically cited in future peer-reviewed articles as proof that the Review has been “debunked” despite the serious problems in these papers. We strongly urge peer reviewers and journal editors to attend more carefully to the lax scholarly norms in this field and to work to strengthen them.

Appendix: ROBINS analyses

ROBINS-I V2

The Risk Of Bias In Non-randomized Studies – of Interventions, Version 2 (ROBINS-I V2) assessment tool (for follow-up studies)

November 2024

VERSION 2: LAUNCH VERSION, 22 November 2024

Outline of ROBINS-I V2

ROBINS-I aims to assess the risk of bias in a specific result from an individual non-randomized study that examines the effect of an intervention on an outcome. This document describes the ROBINS-I V2 tool for **follow-up (cohort) studies**. Assessments should relate to risk of **material bias** rather than risk of any bias. Material bias should be interpreted as bias sufficient to cause an important change to the magnitude of the estimated effect, compared with the true value.

Before undertaking a ROBINS-I assessment (or series of assessments, e.g., in the context of a systematic review), users of the tool should specify the important confounding factors that are likely to influence the association between the intervention and the outcome (see section “At planning stage”).

The start point for an assessment of a specific study is to specify the result from the study that is being assessed for risk of bias. A ‘screening’ section then facilitates identification of results that are at “Critical risk of bias”, allowing the user to avoid a detailed assessment.

A key feature of the ROBINS-I approach is the specification, for each study, of the causal effect estimated by the result under consideration through specification of a hypothetical ‘target trial’. This is essential for assessment of risk of bias, because the causal effect defines the result that would be seen (other than the impact of sampling variation) in the absence of bias.

If multiple assessors will implement ROBINS-I independently, the *Preliminary considerations to plan the assessment* should be agreed between all assessors before each assessor works individually through evaluation of the confounding factors and bias domains.

ROBINS I includes seven domains of bias:

- Domain 1: Risk of bias due to confounding
- Domain 2: Risk of bias in classification of interventions
- Domain 3: Risk of bias in selection of participants into the study (or into the analysis)
- Domain 4: Risk of bias due to deviations from intended interventions
- Domain 5: Risk of bias due to missing data
- Domain 6: Risk of bias arising from measurement of the outcome
- Domain 7: Risk of bias in selection of the reported result

Each bias domain in ROBINS-I is addressed using a series of **signalling questions** that aim to gather important information about the study and the analysis being assessed. Most signalling questions have response options ‘Yes’, ‘Probably yes’, ‘Probably no’, ‘No’ and ‘No information’, with ‘Yes’ and ‘Probably yes’ having the same implications for risk of bias and similarly for ‘No’ and ‘Probably no’. Some questions have additional response options (a ‘weak’ and a ‘strong’ version of ‘Yes’ or ‘No’) to help discriminate between higher and lower risk of bias. After the relevant signalling questions have been completed, an algorithm maps the answers to the signalling questions onto a proposed judgement about **risk of bias** in the result that arises from this domain. The judgements and their broad interpretations are as follows.

Judgement	Interpretation
*Low risk of bias**	There is little or no concern about bias with regard to this domain.
Moderate risk of bias	There is some concern about bias with regard to this domain, although it is not clear that there is an important risk of bias.
Serious risk of bias	The study has some important problems in this domain: characteristics of the study give rise to a serious risk of bias.
Critical risk of bias	The study is very problematic in this domain: characteristics of the study give rise to a critical risk of bias, such that and the result should generally be excluded from evidence syntheses.

*For Domain 1 (Risk of bias due to confounding), this is referred to as “Low risk of bias (except for concerns about uncontrolled confounding)”, in which confounding is very well addressed but cannot be eliminated as a possibility. This is because a risk of bias due to uncontrolled confounding cannot be excluded in an observational study.

ROBINS-I is intended to provide a framework for making informed and reasonable judgements about risk of material bias in studies of the effects of intervention on outcome. On occasion, answers to the signalling questions may not yield an appropriate risk of bias judgement based on the algorithm. Therefore, suggested risk of bias judgements produced by the algorithms can be overridden, in which case justification should be provided. We aim for transparency and reasonableness rather than mechanistic adherence to every word of the tool’s contents.

Optionally, a **predicted direction of bias** may be selected, balancing the various issues addressed within the domain. Response options for this depend on the type of bias being addressed.

After completing all seven bias domains, an **overall judgement** is made for the risk of bias (and optionally for the predicted direction of any bias). The risk-of-bias Judgement is derived from the domain-level judgements using an

algorithm. As for bias domain-level judgements, justification should be provided when the overall judgement suggested by the algorithm is overridden.

An online implementation of ROBINS-I V2 including automatic selection of relevant signalling questions and algorithm-derived risk-of-bias judgements is available via www.riskofbias.info.

Chelliah et al. (2024)

This was a prospective cohort study examining changes in gender dysphoria, minority stress, and mental health among adolescents with gender dysphoria following one year of hormone therapy. The study included 115 participants aged 12 to 18 years and assessed outcomes at baseline and at one-year follow-up, including body image scale, depression and anxiety inventories, and psychosocial quality of life measures. The authors reported significant reductions in body dissatisfaction, depression, anxiety, and victimization, along with improvements in psychosocial functioning, based on paired t-test analyses.

Risk of bias due to confounding: Critical

The greatest limitation of this study was the high potential for uncontrolled confounding. Changes in psychosocial and mental health outcomes could plausibly be explained by important co-interventions, such as psychotherapy, family support, or social transition, rather than hormone therapy alone. Additionally, natural progression or alleviation of dysphoria, depression, or anxiety over time cannot be ruled out. The analytic approach, which relied on paired t-tests, did not adequately account for baseline or time-varying confounders.

Bias in classification of interventions: Low

Because the study used a before–after design in which all participants received hormone therapy, classification of intervention status was straightforward and unlikely to be misclassified.

Bias in selection of participants: Moderate

Concerns arose regarding participant enrollment, particularly as the study period overlapped with a prior cohort, raising the possibility of arbitrary decisions in defining eligibility. One participant was excluded for missing baseline data. In sum, the criteria for defining the cohort were not fully transparent.

Bias due to deviations from intended interventions: Critical

All participants were treated within a multidisciplinary gender clinic; however, potential deviations from intended interventions, such as additional co-interventions, were not systematically captured or reported.

Bias due to missing data: Critical

Of 156 initially eligible participants, follow-up data were available for 115. The missing data may be related to outcomes.

Bias in measurement of outcomes: Serious

Validated self-report instruments were used to assess body dissatisfaction, depression, anxiety, and psychosocial functioning. While these measures are widely accepted, both healthcare providers and participants knew the type of interventions that they received, which introduced the possibility of measurement bias, particularly if participants anticipated improvement after beginning hormone therapy.

Bias in selection of reported results: Moderate

The study presented both paired t-test and regression analyses. Although sensitivity analyses excluding 14 participants previously included in another study were conducted, only results with these participants included were reported.

Overall ROBINS-I judgement

Considering all domains, the study was ultimately judged to be at critical risk of bias. This rating was primarily driven by the lack of adequate control for confounding, incomplete reporting of deviations from intended interventions, and substantial missing data. While the findings suggest improvements in psychosocial outcomes following hormone therapy, the validity of attributing these changes causally to the intervention is highly limited.

The ROBINS-I V2 tool: Chelliah et al.

At planning stage: list confounding factors

P1. List the important confounding factors relevant to all or most studies on this topic. Specify whether these are particular to specific intervention-outcome combinations.

Guidance notes

A confounding factor is a prognostic factor that predicts the interventions received. Important confounding factors are those that have the potential to introduce material bias into an estimated effect. Factors that are expected to have only very weak associations with the intervention or with the outcome, such that failure to account for them in the analysis will not have a material impact on the estimated effect of intervention on outcome, need not be considered here. Important confounding factors should be pre-specified at the planning stage, for example in the protocol of a systematic review that will include studies of the effects of interventions. The identification of potential confounding factors requires content knowledge and may usefully be informed by examination of relevant literature. Important confounding factors should be specified at the level of the broad research question (e.g. using a single list of confounding factors for a systematic review). This broad question may cover several specific interventions and/or outcomes. If confounding factors are specific to particular intervention-outcome combinations, then this should be stated.

Characteristics including natal sex, age of gender dysphoria diagnosis, starting age of intervention/duration of gender dysphoria diagnosis before treatment

Comorbidities such as anxiety, depression, baseline suicidality, ADHD, etc.

Co-interventions such as psychological support, family support, social transition, surgery

For each study result: preliminary considerations

Guidance notes

The following questions should be answered only for the specific result that is being evaluated for the current ROBINS-I assessment.

In case of multiple alternative analyses being presented, it is important to specify the numeric result (e.g. RR = 1.52 (95% CI 0.83 to 2.77) and/or a reference (e.g. to a table, figure or paragraph) that uniquely defines the result being assessed.

Some characteristics of a study or a result may lead directly to the result being at critical risk of bias, and so make detailed risk-of-bias assessments unnecessary. A series of preliminary questions in this section aim to identify such situations.

Two preliminary questions are used to examine whether there is a need to examine time-varying confounding in the first domain of the tool (Bias due to confounding). If participants could switch between intervention groups then associations between intervention and outcome may be biased by time-varying confounding. This occurs when prognostic factors influence switches between intended interventions. For example, in a cohort study of the effect of antiretroviral therapy (ART) on rates of AIDS and death in people with HIV, follow-up time for each participant was split according to receipt of ART. Because CD4 counts during follow-up influenced the decision to start ART, CD4 count was a time-varying confounder.

The **target randomized trial specific to the study** is a hypothetical randomized trial, which need not be ethical or feasible, that compares the health effects of the same interventions, conducted with the same eligibility criteria as the non-randomized study. In general, such target trials will not use blinding of participants or of health professionals administering interventions.

If multiple assessors will implement ROBINS-I independently, the questions in this section should be agreed between all assessors before each assessor works individually through the risk-of-bias assessment itself.

A. Specify the result being assessed for risk of bias

Guidance notes (specifying the numerical result)

A ROBINS-I assessment of risk of bias is specific to a particular study result. This is because different results from the same study may be at importantly different risks of bias (consider, for example, an unadjusted estimate of intervention effect compared with an estimate that is adjusted for numerous important confounding factors). Consequently, it may be necessary to undertake several ROBINS-I assessments of different results from the same study. If the study presents multiple alternative analyses, specify the numerical result (e.g. RR=1.52 (95% CI 0.83 to 2.77)) and/or a reference (e.g. to a table, figure or paragraph) that uniquely defines the result being assessed.

A1. Specify the numerical result being assessed

Change scores for body dissatisfaction, −18.0 (18.1), depression, −2.8 (5.6), anxiety, −6.3 (15.9), and psychosocial quality of life, 7.6 (16.4).
The risk of bias considerations were similar across these outcomes.

A2. Provide further details about this result (for example, location in the study report, reason it was chosen) [optional]

Table 1. “Significant reductions in body dissatisfaction (t (107) = 10.39, p < .001), parent gender-related nonaffirmation (t (98) = 3.15, p < .01), and victimization (t (98) = 3.06, p < .01) were found between baseline and year one. Reductions in anxiety (t (80) = 3.54, p < .01) and depression (t (108) = 5.16, p < .001) were also found along with improvements in quality of life (t (108) = -4.86, p < .001). However, changes were not significant for family social support, friend social support, and parent gender-related acceptance.” The rationale of choosing these outcomes: Dowshen et al. listed this study as one of the “cohort studies reporting improvements in psychosocial functioning after treatment”

B. Decide whether to proceed with a risk-of-bias assessment

Guidance notes (whether to proceed with a risk-of-bias assessment) Some characteristics of a study or a result may lead directly to the result being at critical risk of bias, and so make detailed risk-of-bias assessments unnecessary. The questions in this section aim to identify such situations.

B1 Did the authors make any attempt to control for confounding?	Confounding is a substantial problem in most non-randomized studies, and it is usually important to control for the important confounding factors.	N
B2 If N/PN to B1: Is there sufficient potential for confounding that an unadjusted result should not be considered further?	If there is sufficient potential for confounding that an unadjusted result should not be considered further, then the result is judged to be at ‘Critical risk of bias’.	Yes

B3 Was the method of measuring the outcome inappropriate?	This question aims to identify methods of outcome measurement (data collection) that are unsuitable for the outcome they are intended to evaluate. This enables a rapid assessment that a result should be regarded as at 'Critical risk of bias'. The question does not aim to assess whether the choice of outcome being evaluated was *sensible* (e.g. because it is a surrogate or proxy for the main outcome of interest). In most circumstances, for pre-specified outcomes, the answer to this question will be 'N' or 'PN'. Answer 'Y or 'PY' if the method of measuring the outcome is inappropriate, for example because: (1) important ranges of outcome values fall outside levels that are detectable using the measurement method; or (2) the measurement instrument has been demonstrated to have such poor reliability or validity that estimates of the relationship between intervention and the measured outcome are not useful. (3) The measurement method differed substantially between people in the intervention and comparator groups, so that differences between the groups are not interpretable.	PN

If the answer to either B2 or B3 is 'Yes' or 'Probably yes', the result should be considered to be at 'Critical risk of bias' and no further assessment is required.

We decided to continue the assessment to document limitations in this study with details.

C. Specify the analysis in the current study for which results are being assessed for risk of bias

Specify the outcome to which this result relates.

Change scores for body dissatisfaction, depression, anxiety, and psychosocial quality of life. The risk of bias considerations were similar across these outcomes.

C1. Specify the participant group on which this result was based.

Single group, before and after

C2 to C3. Determine whether there is a need to consider time-varying confounding.

C2. Was the analysis based on splitting participants' follow up time according to intervention received, or was follow-up censored when participants in one group switched to another group (e.g. when comparison group participants started the intervention)?

☐ Yes (it is a before-after study)	Use Variant A of Domain 1 Proceed to next question

C3. If **Y** to C2, were intervention discontinuations or switches likely to be related to factors that are predictive of the outcome?

Use Variant A of Domain 1

☐ Yes — Use Variant B of Domain 1

D. Specify a (hypothetical) target randomized trial specific to the study

Guidance notes

Evaluations of risk of bias are facilitated by considering the non-randomized study as an attempt to emulate a pragmatic randomized trial, which we refer to as the **target trial**. The first part of a ROBINS-I assessment for a particular study is to specify a target trial - the hypothetical randomized trial whose results should be the same as those from the non-randomized study under consideration, in the absence of bias. Its key characteristics are the types of participant (including exclusion/inclusion criteria) and descriptions of the intervention strategy and comparator strategy. These issues were considered in more detail by Hernán (2016). Differences between the target trial for the individual non-randomized study and the generic research question of the review relate to issues of heterogeneity and/or generalizability rather than risk of bias.

Because it is hypothetical, ethics and feasibility need not be considered when specifying the target trial. For example there would be no objection to a target trial that compared individuals who did and did not start smoking, even though such a trial would be neither ethical nor feasible in practice.

Selection of a patient group that is eligible for a target trial may require detailed consideration, and lead to exclusion of many patients. For example, Magid et al (2010) studied the comparative effectiveness of ACE inhibitors compared to beta-blockers as second-line treatments for hypertension. From an initial cohort of 1.6m patients, they restricted the analysis population to (1) persons with incident hypertension, (2) who were initially treated with a thiazide agent, and (3) who had one of the two drugs of

interest added as a second agent for uncontrolled hypertension, and (4) who did not have a contraindication to either drug. Their "comparative effectiveness" cohort included 15,540 individuals: less than 1% of the original cohort.

A note on terminology: Throughout ROBINS-I V2, we refer regularly to "intervention" and "comparator". The comparator may be an alternative active intervention, a control condition or no intervention at all.

We sometimes refer to the "intervention strategy" and "comparator strategy", because an intervention typically consists of a package of care or procedures, and may be implemented over a period of time rather than on a single occasion. Specification of the whole strategy of interest is particularly important when interest is in a 'per protocol' effect.

In non-randomized studies, assignment to the intervention or comparator is inferred from the recorded intervention for each participant. This is in contrast to randomized trials, in which participants are randomly assigned to the intervention or comparator. We refer to the participants assigned to each strategy as the "intervention group" and "comparator group".

Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology* 2016;183:758-64; doi:10.1093/aje/kwv254.

Magid DJ, Shetterly SM, Margolis KL, Tavel HM, O'Connor PJ, Selby JV, Ho PM. Comparative effectiveness of angiotensin-converting enzyme inhibitors versus beta-blocker as second-line therapy for hypertension. *Circulation: Cardiovascular Quality and Outcomes* 2010;3:453-458; doi:10.1161/CIRCOUTCOMES.110.940874.

D1. Specify the participants and eligibility criteria	Transgender youth aged 12 to 18 years old
D2. Specify the intervention strategy	Puberty blockers or cross sex hormones – this hypothetical target trial should specify which treatment it aims to evaluate, puberty blockers or cross sex hormones
D3. Specify the comparator strategy	Placebo for both

E. Decide on the effect of interest

E1. Is your aim for this study…?

☐ to assess the intention-to-treat effect (the effect of *assignment to* an intervention strategy or comparator strategy)

E2. **If the aim is to assess a per-protocol effect**, briefly define the changes to the intervention or comparator strategies that will be considered to be protocol deviations and, optionally, those changes that will not be considered. For example, the protocol deviations considered could be: “Starting intervention among comparator group participants, while acceptable changes could be “stopping intervention because of intervention-related toxicities occur or disease progression” or “changes to intervention after the trial baseline”.

F. Information sources

Guidance notes

Evaluation of a study should be based on the maximum possible amount of available information. In addition to published papers describing a study's methods and results, such information may be derived from the study protocol, unpublished reports or through correspondence with the study investigators.

Which of the following sources have you obtained to help you inform your risk of bias judgements (tick as many as apply)?

- ☐ Journal article(s)
- ☐ Study protocol
- ☐ Statistical analysis plan (SAP)
- ☐ Non-commercial registry record (e.g. ClinicalTrials.gov record)
- ☐ Company-owned registry record (e.g. GSK Clinical Study Register record)
- ☐ "Grey literature" (e.g. unpublished thesis)
- ☐ Conference abstract(s)
- ☐ Regulatory document (e.g. Clinical Study Report, Drug Approval Package)
- ☐ Individual participant data
- ☐ Research ethics application
- ☐ Grant database summary (e.g. NIH RePORTER, Research Councils UK Gateway to Research)
- ☐ Personal communication with investigator
- ☐ Personal communication with sponsor

Please specify any additional sources not listed above

Evaluation of confounding factors

Complete a row for each important confounding factor listed in advance (subsection (i) below); and either relevant to the setting of this particular study or identified by the study authors (subsection (ii)). **"Important" confounding factors are those for which, in the context of this study, adjustment is expected to lead to a meaningful change in the estimated effect of the intervention.**

Guidance notes

Confounding is of fundamental importance to the analysis and interpretation of non-randomized studies of the effect of interventions on outcomes. ROBINS-I addresses two types of confounding: baseline confounding and time-varying confounding.

Baseline confounding occurs when one or more prognostic factors, present before the start of the intervention, predict intervention received. Appropriate methods to control for confounders measured at baseline include stratification, regression, matching, standardization, and inverse probability weighting. The analysis may control for individual variables or for estimated propensity scores (inverse probability weighting is based on a function of the propensity score).

Time-varying confounding needs to be considered in studies that partition follow-up time for individual participants according to intervention received.

We use the term **confounding factor** for each broad source of potential confounding. It may not be possible to measure a factor well, and we distinguish between the confounding factor and the **variables** used to measure it. These variables may be used, for example, as covariates in a regression analysis.

In the context of a particular study, variables need not be included in the analysis: (a) if they are not associated with the outcome, conditional on intervention received (noting that lack of a statistically significant association is not evidence of a lack of association); (b) if they are not associated with intervention; (c) if adjustment makes no or minimal difference to the estimated effect of intervention on outcome; (d) because the confounder was addressed in the study design, for example by restricting to individuals with the same value of the confounder; (e) because a negative control demonstrates that there was unlikely to have

been confounding due to this variable or that uncontrolled confounding was likely to be minimal; or (f) because external evidence suggests that controlling for the variable is not necessary in the context of the study being assessed.

In some studies, researchers may include a very large set of potential confounding variables in an analysis without considering their associations with outcome and intervention. Users of ROBINS-I should focus on (i) the confounding factors they determined a priori to be important and (ii) other factors for which adjustment is expected to lead to an important change in the estimated effect of the intervention on the outcome in the context of the current study.

Users of ROBINS-I should evaluate the confounding factors that they prespecified as important for the intervention-outcome relationship under study. The tool also allows the user to evaluate a second list of any further confounding factors that are either relevant to the setting of this particular study or which the study authors identified as potentially important. It is likely that new ideas relating to confounding and other potential sources of bias will be identified after the drafting of the review protocol, and even after piloting data collection from studies selected for inclusion in the systematic review. For example, such issues may be identified because they are mentioned in the introduction and/or discussion of one or more papers. This could be addressed in practice by explicitly recording whether potential confounders or other sources of bias are mentioned in the paper.

In very rare situations it is possible that no confounding factors are present, either because interventions received are known to be unrelated to any prognostic factors for the outcome of interest, or because no such prognostic factors exist. In such situations, the risk of bias due to confounding may be assessed as low.

The purpose of this preliminary assessment of confounding factors is to review the extent to which the result being assessed was controlled for confounding, considering both the prespecified confounding factors and any further confounding factors identified as important in the context of the study being assessed. This enables users of ROBINS-I to answer the signalling questions for the Domain 1 assessment (Risk of bias due to confounding). "Important" confounding factors are those for which, in the context of this study, adjustment is expected to lead to an important change in the estimated effect of the intervention.

The preliminary assessment consists of the following steps for each confounding factor.

- determine which variables (if any) were measured for the factor;
- determine which of these variables were controlled for in the analysis;
- for variables that were not controlled for, look for evidence that controlling for the variable was not necessary in this particular study;
- determine whether the confounding factor was measured validly and reliably by the variables used to measure it (this is assessed at the level of the confounding factor rather than the level of the individual variables used to measure the factor);
- determine the likely direction of bias if the analysis fails to adjust for this variable (alone).

The direction of bias, if the analysis fails to adjust for a particular variable (alone), will be that the effect estimate is biased *upwards* or biased *downwards*. For example, if older age predicts that a particular intervention is more likely to be received and the outcome is mortality, then this confounding would bias the estimated effect downwards: unless we adjust for age the intervention will appear more positively associated with higher mortality than it should. In the presence of *positive confounding* (the confounder is positively associated with both intervention and outcome, or negatively associated with both intervention and outcome), the bias will be upwards. In the presence of *negative confounding* (the confounder is positively associated with intervention and negatively associated with outcome, or vice versa), the bias will be downwards.

(i) Important confounding factors listed in advance						
Confounding factor	Measured variable(s) for this factor, if any	Was this variable (or were these variables) controlled for in the analysis? (Y / N)	If this confounding factor was controlled for, was it measured validly and reliably by this variable (or these variables)?* (NA / Y / PY / PN / N / NI)	If this confounding factor was not controlled for, is there evidence that controlling for it was unnecessary?** (NA / Y / PY / PN / N)	OPTIONAL: Is failure to adjust for this confounding factor expected to bias the effect estimate upwards or downwards? (Upward bias (overestimate the intervention effect) / Downward bias (underestimate the intervention effect) / No information or unpredictable)	Comments
Natal sex	Electronic health record sex	Y	Y			Paired t test
Age of gender dysphoria diagnosis	Age at first diagnosis of gender dysphoria	Y	Y			Paired t test
Starting age of intervention/duration of gender dysphoria	Y (It seems that the starting age of	Y				They completed survey measures

	intervention is the same as age of diagnosis)					as part of an initial assessment when establishing care. After the assessment, participants were matched with a physician to initiate gender-affirming hormone therapy
Comorbidities	N	Y				Paired t test
Baseline anxiety	Y	Y				Paired t test
Baseline depression	Y	Y				Paired t test

Baseline suicidality	N	Y				Paired t test
Psychological support	Y (Friend support)	N				
Family support	Y	N				
Social transition	N	N				
Surgery	N	N				

(ii) Additional important confounding factors relevant to the setting of this particular study, or identified by the study authors						
Confounding factor	Measured variable(s) for this factor, if any	Was this variable (or were these variables) controlled for in the analysis? (Y / N)	If this confounding factor was controlled for, was it measured validly and reliably by this variable (or these variables)?* (NA / Y / PY / PN / N / NI)	If this confounding factor was not controlled for, is there evidence that controlling for it was unnecessary?** (NA / Y / PY / PN / N)	OPTIONAL: Is failure to adjust for this confounding factor expected to bias the effect estimate upwards or downwards? (Upward bias (overestimate the intervention effect) / Downward bias (underestimate the intervention effect) / No information or unpredictable)	Comments

Race	Race	Y	Y			Paired t test
Health insurance	Type of health insurance	Y	Y			Paired t test

* “Validity” refers to whether the confounding variable or variables accurately measure the confounding factor, while “reliability” refers to the precision of the measurement (more measurement error means less reliability).

** In the context of a particular study, variables need not be included in the analysis: (a) if they are measured validly and reliably and are not associated with the outcome, conditional on intervention (noting that lack of a statistically significant association is not evidence of a lack of association; (b) if they are measured validly and reliably and are not associated with intervention; (c) if they are measured validly and reliably and adjustment makes no or minimal difference to the estimated effect of the primary parameter; (d) because the confounder was addressed in the study design, for example by restricting to individuals with the same value of the confounder; (e) because a negative control demonstrates that there was unlikely to have been confounding due to this variable or that uncontrolled confounding was likely to be minimal; or (f) because external evidence suggests that controlling for the variable is not necessary in the context of the study being assessed”.

Risk of bias assessment

Responses underlined in green are potential markers for low risk of bias, and responses in red are potential markers for a risk of bias. Where questions relate only to sign posts to other questions, no formatting is used.

Guidance notes

The questions in this domain focus on the confounding factors that were identified as important in the preliminary evaluation in section E.

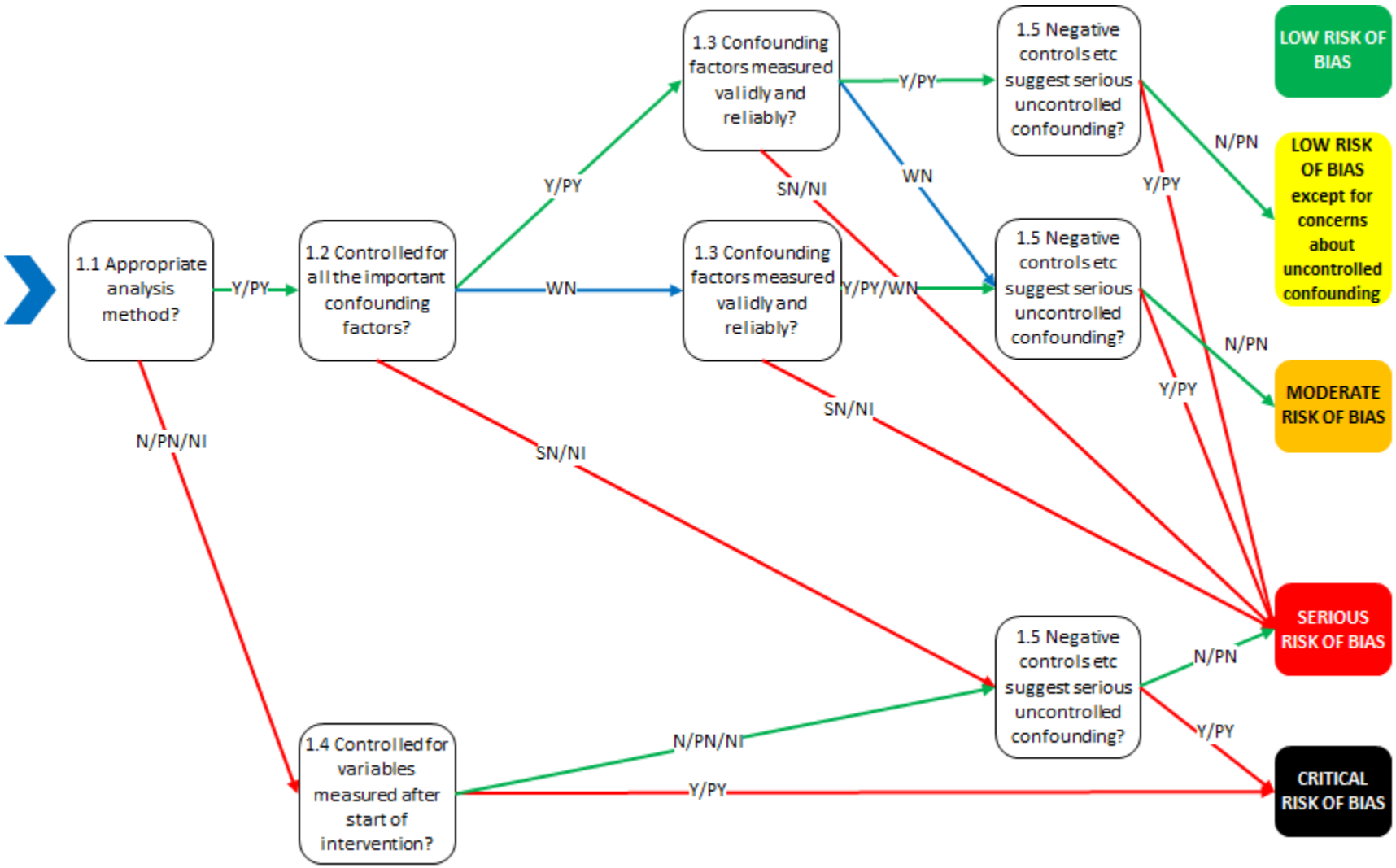
We use the term uncontrolled confounding to refer to confounding that was not controlled by the design or analysis of the study – and is therefore likely to bias the estimated effect of intervention. This may arise because (i) confounding factors were not (or could not) be measured; (ii) variables used to measure confounding factors were insufficient to characterize the confounding factor; or (iii) variables that characterize the confounding factor were measured but not included in the analysis.

Domain 1, Variant B (the analysis was based on splitting participants' follow up time according to intervention received, so both baseline and time-varying confounding need to be addressed – Y to C2 and Y to C3)

Signalling questions	Elaboration	Response
1.1 Did the authors use an analysis method that was appropriate to control for time-varying as well as baseline confounding?	The authors used paired t test, which is not sufficient.	N
1.2 If Y/PY to 1.1: Did the authors control for all the important baseline and time-varying confounding factors for which this was necessary?	Co-interventions, such as psychotherapy, surgery, or family and friend support, were not adjusted for. Natural progress of the condition, alleviation of dysphoria, anxiety, etc., or progression of these conditions cannot be controlled.	SN (no, and uncontrolled confounding was probably substantial)
1.3 If Y/PY/WN to 1.2: Were confounding factors that were controlled for (and for which control was necessary) measured validly and reliably by the variables available in this study?	As for variant A, question 1.2.	NA
1.4 If N/PN/NI to 1.1: Did the authors control for time-varying factors or other variables measured after the start of intervention?	No	N
1.5 Did the use of negative controls, or other considerations, suggest serious unmeasured confounding?	Paired t test results suggested improvement; however, the study design cannot properly assess the influence of other variables.	Y

Risk of bias judgement	As for variant A.	Critical risk of bias
Optional: What is the predicted direction of bias due to confounding?	As for variant A.	

Algorithm for reaching default risk of bias judgement:

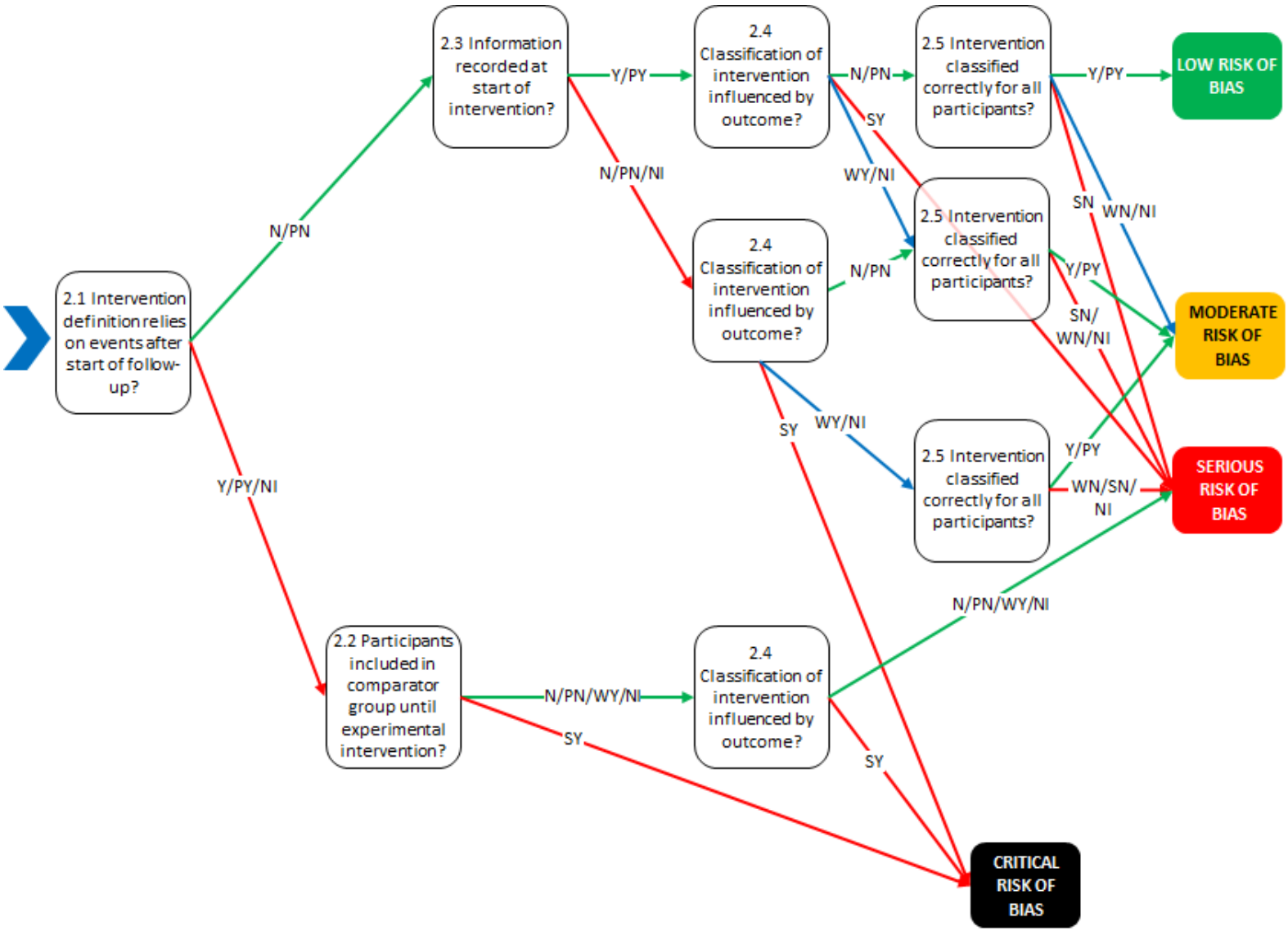


2. Bias in classification of interventions

Signalling questions	Elaboration	Response options
Questions about immortal time bias arising from definition of intervention groups		
2.1 Did assignment of participants to the intervention group or the comparator group rely on events or measurements that occurred after the start of follow up?	All participants in this before-after study	PN
2.2 If Y/PY to 2.1: Were participants included in the comparator group until they fulfilled the definition of the intervention (or vice versa)?	-	NA
Questions about differential misclassification		
2.3 If N/PN to 2.1: Was all information used to classify intervention and comparator groups recorded at or before the time the interventions started?	It was a before-after study, and the information to classify participants as receiving intervention was recorded probably at the start of the intervention	PY
2.4 Was classification of intervention status influenced by knowledge of the outcome or risk of the outcome?		PN
Question about non-differential misclassification		

2.5 If N/PN to 2.1 and WY/N/PN/NI 2.4: Was intervention status classified correctly for all, or nearly all, participants?	Probably nearly all	PY
Risk of bias judgement	See algorithm.	Low
Optional: What is the predicted direction of bias in classification of interventions?	-	Favours intervention / Favours comparator / Towards null /Away from null / Unpredictable

Algorithm for reaching default risk of bias judgement:



3. Bias in selection of participants into the study (or into the analysis)

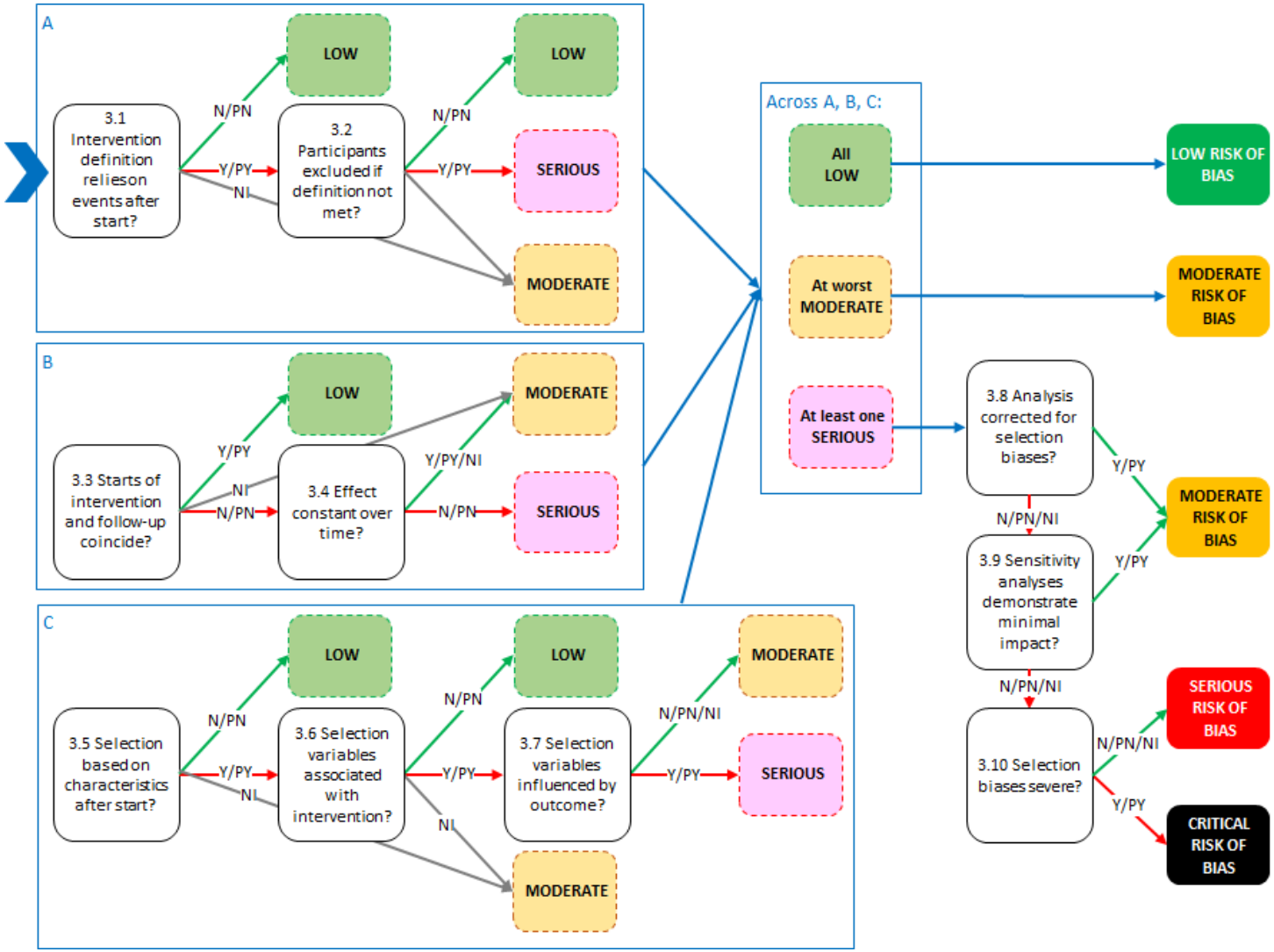
In the target trial, start of follow up is the time at which participants meet eligibility criteria and are assigned to interventions. In answering the signalling questions for this domain, consider what is the start of follow up in the study under consideration, for both the intervention and comparison groups.

Signalling questions	Elaboration	Response options
A. Questions about immortal time bias arising from definition of intervention groups		
3.1 (=2.1) Did assignment of participants to the intervention group or the comparator group rely on events or measurements that occurred after the start of follow up?	It was a before-after study, and the information to classify participants as receiving intervention was recorded **after the start of follow up**	Y
3.2 If Y/PY to 3.1: Were participants excluded after the start of follow-up because they did not meet the definition of either the intervention or the comparator?	One participant was excluded because "missing an initial assessment."	PY
B. Questions about prevalent user bias		
3.3 Were start of follow up and start of intervention the same for most participants?		PY
3.4 If N/PN to 3.3: Is the effect of intervention expected to be constant over the time period studied?	-	NA
C. Questions about other types of selection bias		

3.5 Was selection of participants into the study (or into the analysis) based on participant characteristics observed after the start of intervention (additional to the situations addressed in 3.1 and 3.3)?	The authors reported "156 eligible participants," however, it is unclear how these 156 eligible participants were enrolled and selected. The authors mentioned "the study period for the previous and current studies overlapped slightly resulting in 14 participants included in both samples," suggesting that there was arbitrary decision on the study period. It is unclear how this decision impacted participant selection, such as age, natal sex, cointervention of the study participants.	PY
3.6 If Y/PY to 3.5: Were the post-intervention variables that influenced selection likely to be associated with intervention?	It is unclear whether this study duration was associated with intervention or cointervention.	NI
3.7 If Y/PY to 3.6: Were the post-intervention variables that influenced selection likely to be influenced by the outcome or a cause of the outcome?		NA
D. Questions about analysis, sensitivity analyses and severity of the problem		
3.8 If Y/PY to 3.2, N/PN 3.4 or Y/PY to 3.7: Is it likely that the analysis corrected for all of the potential selection biases identified in 3.1-3.2, 3.3-3.4 or 3.5-3.7 above?	Only one participant excluded in 3.2 question.	PY
3.9 If N/PN to 3.8: Did sensitivity analyses demonstrate that the likely impact of the	-	NA

potential selection biases identified in 3.1-3.2, 3.3-3.4 or 3.5-3.7 above was minimal?		
3.10 If N/PN to 3.9: Were potential selection biases identified in 3.1-3.2, 3.3-3.4 or 3.5-3.7 above sufficiently severe that the result should not be included in a quantitative synthesis?	-	NA
Risk of bias judgement	See algorithm.	Moderate
Optional: What is the predicted direction of bias in selection of participants into the study?	-	Favours intervention / Favours comparator / Towards null /Away from null / Unpredictable

Algorithm for reaching default risk of bias judgement:



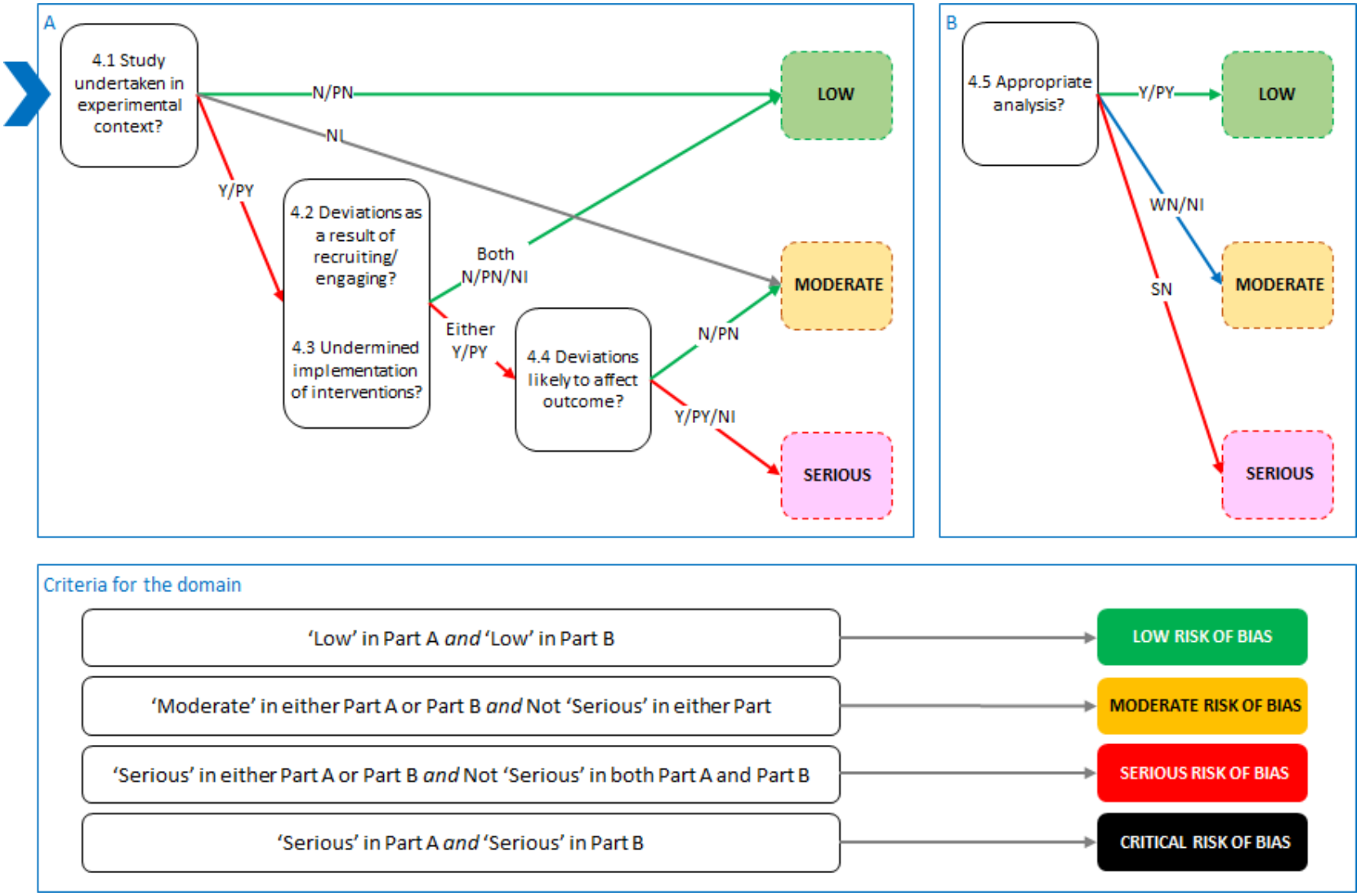
4. Bias due to deviations from intended interventions

Domain 4, Variant A: Effect of assignment to intervention

Signalling questions	Elaboration	Response options
4.1 Was the study undertaken in an experimental context?	No	N
4.2. If Y/PY to 4.1: Did participants deviate from the intended intervention as a result of the processes of recruiting and engaging them in the study?	-	NA
4.3. If Y/PY to 4.1: Did study personnel consciously or unconsciously undermine implementation of the intended interventions?	-	NA
4.4. If Y/PY/NI to 4.2 or 4.3: Were these deviations from intended intervention likely to have affected the outcome?	Cointervention, or non-compliant may have happened	PY
4.5. Was an appropriate analysis used to estimate the effect of assignment to intervention?	Only paired t test	N

Risk of bias judgement	See algorithm.	Critical
Optional: What is the predicted direction of bias in classification of interventions?	If the likely direction of bias can be predicted, it is helpful to state this. The direction might be characterized as being in favour of the intervention, as being in favour of the comparator, or as towards (or away from) the null.	Favours intervention / Favours comparator / Towards null /Away from null / Unpredictable

Algorithm for reaching default risk of bias judgement (effect of assignment to intervention):



5. Bias due to missing data

Guidance notes

Missing outcome data may arise, among other reasons, through attrition (loss to follow up), missed appointments and incomplete data collection. Additionally, in non-randomized studies data may be missing for characteristics including interventions received and confounders.

A general rule for consideration of bias due to missing data is that we should consider biases introduced by the missing data, compared with the effect estimate from an analysis in which all the data we intended to collect were available. Unfortunately, a single threshold for an acceptable proportion of missing data cannot meaningfully be defined. For example, a result based on 95% complete outcome data might be biased if the outcome was rare and if reasons for missing outcome data were strongly related to intervention group. Therefore, the potential for bias due to missing data should be assessed unless complete data on intervention status, the outcome and confounding variables were available for all, or nearly all, participants.

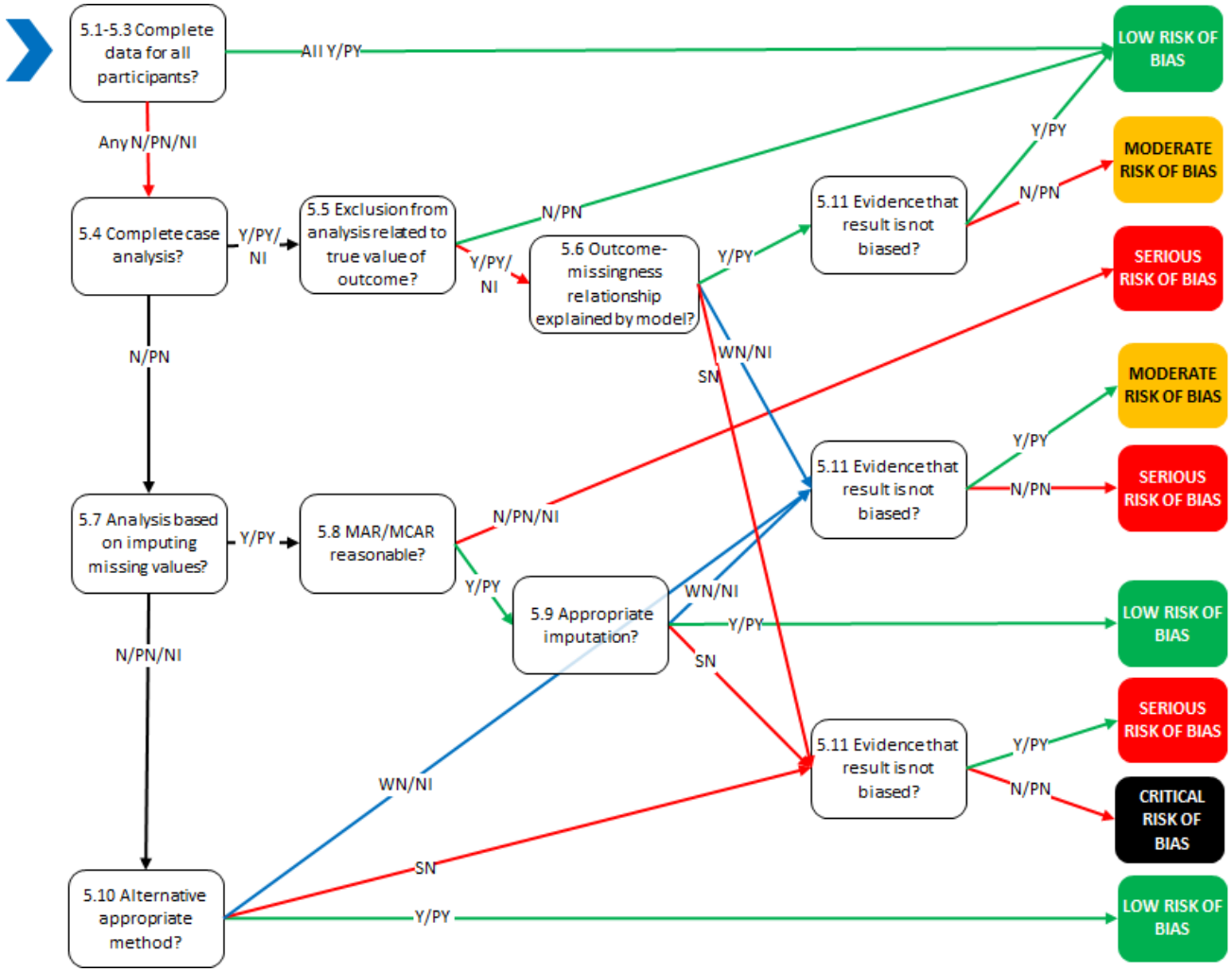
Considerations of bias due to missing data depend on how the analysis accounted for the missing data. Different signalling questions should be answered depending on three types of analysis. The first is that a **complete case analysis**, restricted to participants with complete data on all the intervention, outcome and confounding variables, was performed. In this situation, an important consideration is whether missingness of individual participants from the analysis is related to the true value of the outcome for those participants. The second is that missing data were **imputed**, which means that estimated or assumed values were assigned to participants with missing data. Imputed data should not lead to bias if the data are 'missing at random' (see the elaboration for signalling question 5.8) and an appropriate imputation method is applied. Other types of analysis are addressed by a separate, general, signalling question. The final signalling question asks whether sensitivity analyses were performed that demonstrated that the impact of missing data is minimal.

Signaling questions	Elaboration	Response options
5.1 Were complete data on intervention status available for all, or nearly all, participants?	"Of 156 eligible participants, 32 were missing year one assessments, one was missing an initial assessment, and eight transferred care prior to the year one follow-up."	Y
5.2 Were complete data on the outcome available for all, or nearly all, participants?	"Of 156 eligible participants, 32 were missing year one assessments, one was missing an initial assessment, and eight transferred care prior to the year one follow-up."	N
5.3 Were complete data on important confounding variables available for all, or nearly all, participants?	Important cointerventions were not available	N
5.4 If N/PN/NI to 5.1, 5.2 or 5.3: Is the result based on a complete case analysis?		Y
5.5 If Y/PY/NI to 5.4: Was exclusion from the analysis because of missing data (in intervention, confounders or the outcome) likely to be related to the true value of the outcome?	Loss to follow up or transferral of care may be associated with nonadherence or worse outcomes	Y
5.6 If Y/PY/NI to 5.5: Is the relationship between the outcome and missingness likely to be	-	SN (No, and bias is likely to be substantial)

explained by the variables in the analysis model?		
5.7 If N/PN to 5.4: Was the analysis based on imputing missing values?	-	NA
5.8 If Y/PY to 5.7: Is it reasonable to assume that data were 'missing at random' (MAR) or 'missing completely at random' (MCAR)?	-	NA
5.9 If Y/PY to 5.8: Was imputation performed appropriately?	-	NA
5.10 If N/PN/NI to 5.7: Was an appropriate alternative method used to correct for bias due to missing data?	-	NA
5.11 If PN/N/NI to 5.1, 5.2 or 5.3 AND (Y/PY/NI to 5.5 OR (Y/PY to 5.8 AND WN/SN/NI to 5.9) OR WN/SN/NI to 5.10): Is there evidence that the result was not biased by missing data?	The missingness could be associated with outcomes and the authors did not explore or evaluate this factor.	PN
Risk of bias judgement	See algorithm.	Critical
Optional: What is the predicted direction of bias due to missing data?	-	Favours intervention / Favours

		comparator / Towards null /Away from null / Unpredictable

Algorithm for reaching default risk of bias judgement:



6. Bias in measurement of the outcome

Guidance notes

Bias may be introduced if outcomes are misclassified or measured with error. Misclassification or measurement error of outcomes may be non-differential or differential.

Non-differential measurement error is unrelated to the intervention received. It can be systematic (for example when measurement of blood pressure is consistently 5 units too high in every participant) – in which case it will not affect precision or cause bias; or it can be random (for example when measurement of blood pressure is sometimes too high and sometimes too low in a manner that does not depend on the intervention or the outcome) – in which case it will affect precision without causing bias.

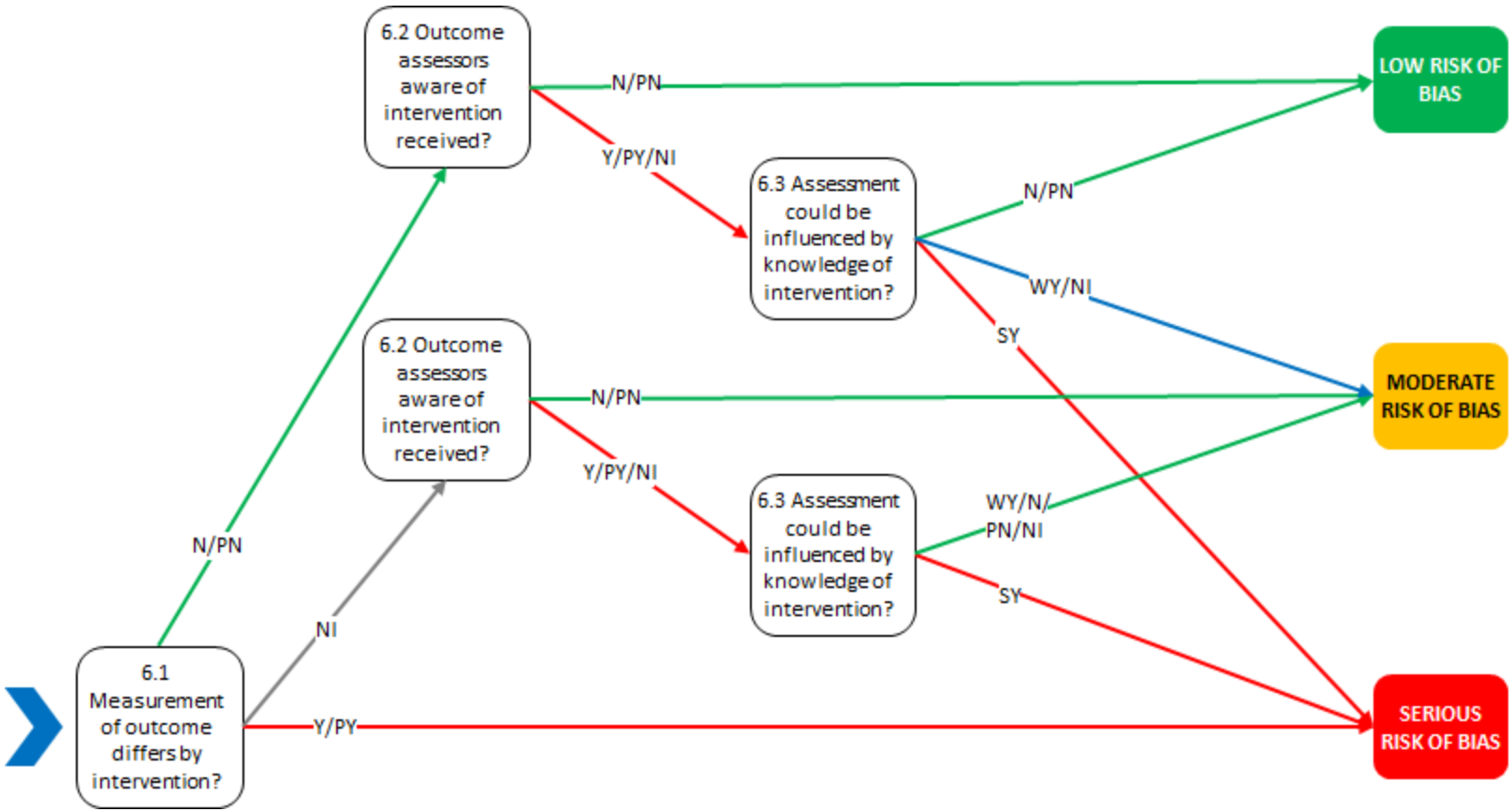
Differential measurement error is measurement error related to intervention received. It will bias the intervention-outcome relationship. This is often referred to as detection bias. Examples of situations in which detection bias can arise are (i) if outcome assessors are aware of intervention received (particularly when the outcome is subjective); (ii) different methods (or intensities of observation) are used to assess outcomes of participants receiving different interventions; and (iii) measurement errors are related to intervention received (or to a confounder of the intervention-outcome relationship).

Blinding of outcome assessors aims to prevent systematic differences in measurements according to intervention received. However, blinding is frequently not possible or not performed for practical reasons.

Signalling questions	Elaboration	Response options
6.1 Could measurement or ascertainment of the outcome have differed between intervention groups?		PN

6.2 Were outcome assessors aware of the intervention received by study participants?	No blinding: the study was based on clinical practice	Y
6.3 If Y/PY/NI to 6.2: Could assessment of the outcome have been influenced by knowledge of the intervention received?	The outcomes including body dissatisfaction, anxiety, depression, and quality of life could all be influenced by the knowledge of the intervention received	SY (yes, to a large extent)
Risk of bias judgement	See algorithm.	Serious
Optional: What is the predicted direction of bias in measurement of outcomes?	If the likely direction of bias can be predicted, it is helpful to state this. The direction might be characterized as being in favour of the intervention, as being in favour of the comparator, or as towards (or away from) the null.	Favours intervention / Favours comparator / Towards null /Away from null / Unpredictable

Algorithm for reaching default risk of bias judgement:



7. Bias in selection of the reported result

Guidance notes

Selective reporting can arise for both harms and benefits of an intervention, although the motivations (and direction of bias) underlying selective reporting of effect estimates for harms and benefits may differ. Selective reporting may arise, for example, from a desire for findings to be newsworthy (or sufficiently noteworthy to merit publication), or from commercial considerations, or from a desire to demonstrate that there is not evidence of a harmful effect of an intervention.

Selective outcome reporting occurs when the effect estimate for an outcome measurement was selected from among analyses of multiple outcome measurements for the outcome domain. Examples include: use of multiple measurement instruments (e.g. pain scales) and reporting only the most favourable result; reporting only the most favourable subscale (or a subset of subscales) for an instrument when measurements for other subscales were available; reporting only one or a subset of time points for which the outcome was measured.

Selective analysis reporting occurs when results are selected from effects estimated in multiple ways: e.g. carrying out analyses of both change scores and post-intervention scores adjusted for baseline; multiple analyses of a particular measurement with and without transformation; multiple analyses of a particular outcome with and without adjustment for potential confounders (or with adjustment for different sets of potential confounders); multiple analyses of a particular outcome with and without, or with different, methods to take account of missing data; a continuously scaled outcome converted to categorical data with different cut-points; multiple composite outcomes analysed for one outcome domain, but results were reported only for one (or a subset) of the composite outcomes. (Reporting an effect estimate for an unusual composite outcome might be evidence of such selective reporting.)

Selection of a subgroup from a larger cohort: The cohort for analysis may have been selected from a larger cohort for which data were available on the basis of a more interesting finding. Subgroups defined in unusual ways (e.g. an unusual classification of subgroups by dose or dose frequency) may provide evidence of such selective reporting.

The best evidence that results were not selectively reported is available if a pre-specified, publicly available analysis plan is available (e.g. from a link in a publication or from an online platform) and is in line with the reported results. Protocols for non-

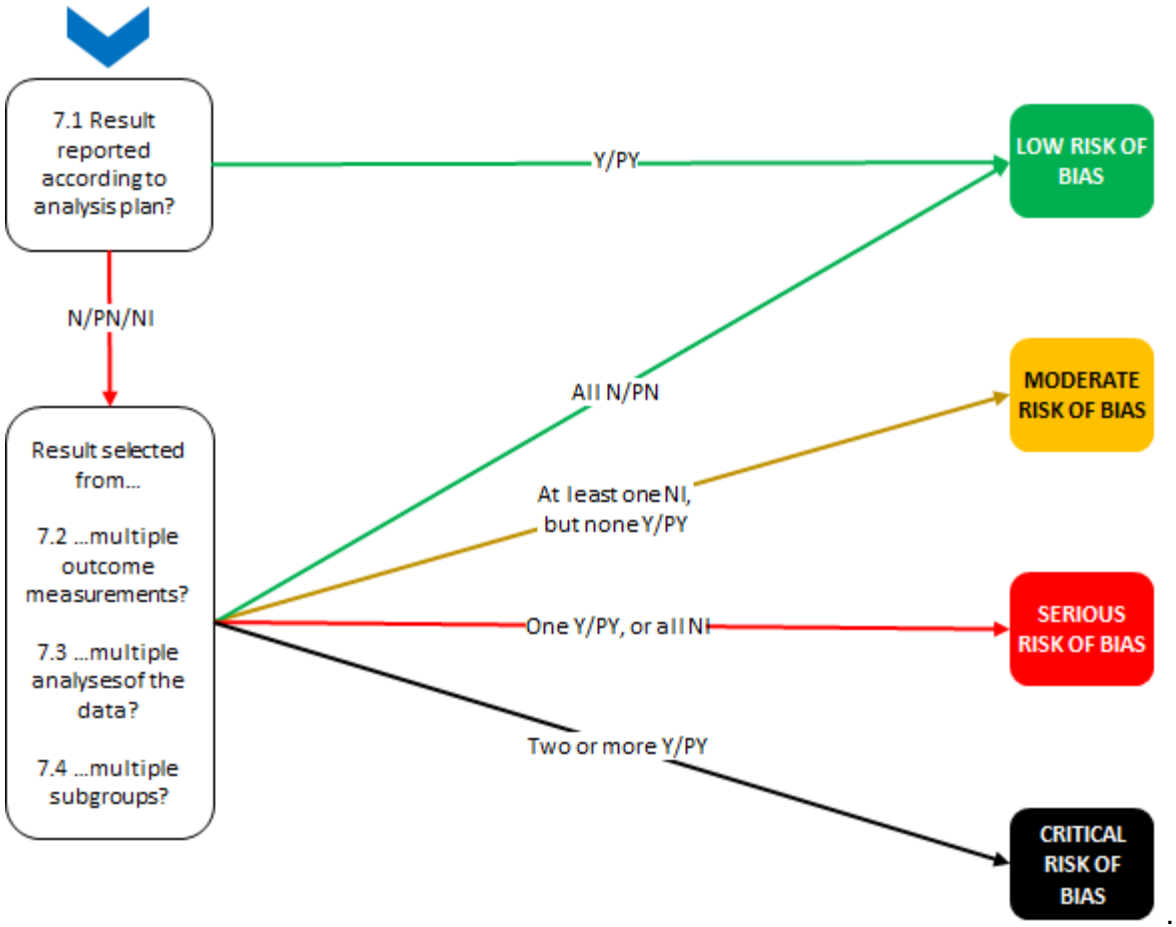
randomized studies are increasingly being registered, although there is inconsistency across platforms (Malmsiø et al, 2022). An analysis plan that is sufficiently detailed to permit full assessment of selective reporting may seldom be available for observational studies. In the absence of a protocol or analysis plan, clues can sometimes be gained by comparing Methods sections with Results sections.

Malmsiø D, Frost A, Hróbjartsson A. A scoping review finds that guides to authors of protocols for observational epidemiological studies varied highly in format and content. J Clin Epidemiol. 2022 Dec 20;154:156-166. doi: 10.1016/j.jclinepi.2022.12.012.

Signalling questions	**Elaboration**	**Response options**
7.1 Was the result reported in accordance with an available, pre-determined analysis plan?		NI
Is the numerical result being assessed likely to have been selected, on the basis of the results, from...		
7.2 ... multiple outcome *measurements* (e.g. scales, definitions, time points) within the outcome domain?		NI
7.3 ... multiple *analyses* of the data?	The authors reported paired t test and regression analyses. It is unclear whether they selectively reported the results of multiple	NI

	analyses. They conducted sensitivity analysis of excluding 14 participants included in a previous study, but reported only the results with these participants after claiming the analyses were similar.	
7.4 ... multiple *subgroups*?		NI
Risk of bias judgement	See algorithm.	Moderate
Optional: What is the predicted direction of bias in selection of the reported result?	If the likely direction of bias can be predicted, it is helpful to state this. The direction might be characterized as being in favour of the intervention, as being in favour of the comparator, or as towards (or away from) the null.	Favours intervention / Favours comparator / Towards null /Away from null / Unpredictable

Algorithm for reaching default risk of bias judgement:



Overall risk of bias

Guidance notes

ROBINS-I defaults to setting the overall risk of bias for a result to be equal to the risk-of-bias judgement for the domain with the greatest risk of bias. For example, if the 'worst' judgement across domains is of serious risk of bias, then the result would be judged as at serious risk of bias overall. However, the user may override this to judge the result to be at greater risk of bias if there are problems in several domains. For example, if several domains are assessed to be at serious risk of bias, and it is considered that these problems are likely to be compounded, then it may be reasonable to judge the result to be at critical risk of bias overall. Predicting the direction of bias overall may be difficult. Risk-of-bias judgements for the individual domains might be used to inform the influence of that domain to the likely direction of bias overall.

Overall risk of bias	See algorithm.	Critical risk
What is the predicted direction of bias?	If the likely direction of bias can be predicted, it is helpful to state this. The direction might be characterized as being in favour of the intervention, as being in favour of the comparator, or as towards (or away from) the null. Alternatively, if the direction is driven by bias due to confounding, the direction may be an upwards bias (overestimate the effect) or a downward bias (underestimate the effect).	Upward bias (overestimate the effect) / Downward bias (underestimate the effect) / Favours intervention / Favours comparator / Towards null /Away from null / Unpredictable

Algorithm for reaching overall risk of bias judgement:

Judgement	**Interpretation**	**How reached**
Low risk of bias except for concerns about uncontrolled confounding	There is the possibility of uncontrolled confounding that has not been controlled for (given the observational nature of the study), but otherwise little or no concern about bias in the result	*Low risk of bias except for concerns about uncontrolled confounding* in Domain 1 and *Low risk of bias* in all other domains
Moderate risk of bias	There is some concern about bias in the result, although it is not clear that there is an important risk of bias	At least one domain is at *Moderate risk of bias*, but no domains are at *Serious risk of bias* or *Critical risk of bias*
Serious risk of bias	The study has some important problems: characteristics of the study give rise to a serious risk of bias in the result	At least one domain is at *Serious risk of bias*, but no domains are at *Critical risk of bias* OR Several domains are at *Moderate*, leading to an additive judgement of *Serious risk of bias*
Critical risk of bias	The study is very problematic: characteristics of the study give rise to a critical risk of bias in the result, such that the result should generally be excluded from evidence syntheses.	At least one domain is at *Critical risk of bias* OR Several domains are at *Serious risk of bias*, leading to an additive judgement of *Criticial risk of bias*

Nunes-Moreno et al. (2025)

This study investigated the association between gender-affirming hormone therapy (GAHT) or gonadotropin-releasing hormone agonists (GnRHa) and emergency department or inpatient diagnoses of suicidality among children and adolescents with gender dysphoria, using the PEDSnet electronic health record network. This study is another observational study examining the treatment effect of puberty blockers (PBs) and CSH in mitigating adverse mental health outcomes among children and adolescents with gender dysphoria. With Cox regression models, the authors reported that CSH was associated with a statistically significant reduction in suicidality risk (hazard ratio [HR] = 0.564 [95% CI 0.36–0.89]), whereas PBs use showed a non-significant trend toward reduced risk (HR = 0.79 [0.47–1.31]). We evaluated the potential risk of bias with these results.

Risk of bias due to confounding: Critical

Bias due to confounding was judged to be at critical risk, as important baseline mental health conditions, family and psychological support, and cointerventions were not controlled.

Bias in classification of interventions and bias in selection of participants: Low

Bias in classification of interventions and selection of participants was rated as low risk, given consistent recording of prescriptions in the electronic health record and reasonable alignment between intervention assignment and follow-up.

Bias due to deviations from intended interventions: Low

Bias due to deviations from intended interventions was also low risk, as deviations were unlikely to materially affect outcomes.

Bias due to missing data: Serious

In contrast, bias due to missing data was judged serious, as confounder data such as baseline mental health and family support were incomplete.

Bias in measurement of outcomes: Serious

Similarly, outcome measurement was at serious risk of bias, since suicidality diagnoses depend on presentation to emergency or inpatient settings and may be under-detected or differentially recorded between groups.

Bias in selection of reported results: Serious

Bias in selection of reported results was also considered serious, given exploratory reporting of GnRHa analyses after non-significant primary findings.

Overall ROBINS-I judgement

Assessment of risk of bias using the *ROBINS-I V2 Tool* (2024) identified concerns across several domains. Taken together, the overall risk of bias for the GAHT and GnRHa results was assessed as critical, reflecting the unresolved confounding and multiple serious risks across domains. Although the study leverages a large multicenter dataset, this study has similar limitations as previously reported observational studies. Consideration of this study would not sway the conclusion of systematic reviews on PBs and CSH, nor the conclusion of the overview of the systematic reviews.

The ROBINS-I V2 tool: Nunes-Moreno

At planning stage: list confounding factors

P1. List the important confounding factors relevant to all or most studies on this topic. Specify whether these are particular to specific intervention-outcome combinations.

Guidance notes

A confounding factor is a prognostic factor that predicts the interventions received. Important confounding factors are those that have the potential to introduce material bias into an estimated effect. Factors that are expected to have only very weak associations with the intervention or with the outcome, such that failure to account for them in the analysis will not have a material impact on the estimated effect of intervention on outcome, need not be considered here. Important confounding factors should be pre-specified at the planning stage, for example in the protocol of a systematic review that will include studies of the effects of interventions. The identification of potential confounding factors requires content knowledge and may usefully be informed by examination of relevant literature. Important confounding factors should be specified at the level of the broad research question (e.g. using a single list of confounding factors for a systematic review). This broad question may cover several specific interventions and/or outcomes. If confounding factors are specific to particular intervention-outcome combinations, then this should be stated.

Characteristics including natal sex, age of gender dysphoria diagnosis, starting age of intervention/duration of gender dysphoria diagnosis before treatment

Comorbidities such as anxiety, depression, baseline suicidality, ADHD, etc.

Co-interventions such as psychological support, family support, social transition, surgery

For each study result: preliminary considerations

Guidance notes

The following questions should be answered only for the specific result that is being evaluated for the current ROBINS-I assessment.

In case of multiple alternative analyses being presented, it is important to specify the numeric result (e.g. RR = 1.52 (95% CI 0.83 to 2.77) and/or a reference (e.g. to a table, figure or paragraph) that uniquely defines the result being assessed.

Some characteristics of a study or a result may lead directly to the result being at critical risk of bias, and so make detailed risk-of-bias assessments unnecessary. A series of preliminary questions in this section aim to identify such situations.

Two preliminary questions are used to examine whether there is a need to examine time-varying confounding in the first domain of the tool (Bias due to confounding). If participants could switch between intervention groups then associations between intervention and outcome may be biased by time-varying confounding. This occurs when prognostic factors influence switches between intended interventions. For example, in a cohort study of the effect of antiretroviral therapy (ART) on rates of AIDS and death in people with HIV, follow-up time for each participant was split according to receipt of ART. Because CD4 counts during follow-up influenced the decision to start ART, CD4 count was a time-varying confounder.

The **target randomized trial specific to the study** is a hypothetical randomized trial, which need not be ethical or feasible, that compares the health effects of the same interventions, conducted with the same eligibility criteria as the non-randomized study. In general, such target trials will not use blinding of participants or of health professionals administering interventions.

If multiple assessors will implement ROBINS-I independently, the questions in this section should be agreed between all assessors before each assessor works individually through the risk-of-bias assessment itself.

A. Specify the result being assessed for risk of bias

Guidance notes (specifying the numerical result)

A ROBINS-I assessment of risk of bias is specific to a particular study result. This is because different results from the same study may be at importantly different risks of bias (consider, for example, an unadjusted estimate of intervention effect compared with an estimate that is adjusted for numerous important confounding factors). Consequently, it may be necessary to undertake several ROBINS-I assessments of different results from the same study. If the study presents multiple alternative analyses, specify the numerical result (e.g. RR=1.52 (95% CI 0.83 to 2.77)) and/or a reference (e.g. to a table, figure or paragraph) that uniquely defines the result being assessed.

A1. Specify the numerical result being assessed

hazard ratio [HR] = 0.564, 95% confidence interval [95% CI: 0.36–0.89], p = 0.0137 -- Among TGD youth prescribed GAHT during our study period, there was a 43.6% reduction in risk of an ED or inpatient diagnosis of suicidality compared with those never prescribed GAHT during our study period or before GAHT initiation.
HR = 0.79 [0.47–1.31], p = 0.357 -- TGD youth who were prescribed GnRHa therapy had a nonstatistically significant reduction in ED or inpatient suicidality diagnoses compared with those never prescribed GnRHa

A2. Provide further details about this result (for example, location in the study report, reason it was chosen) [optional]

B. Decide whether to proceed with a risk-of-bias assessment

Guidance notes (whether to proceed with a risk-of-bias assessment)
Some characteristics of a study or a result may lead directly to the result being at critical risk of bias, and so make detailed risk-of-bias assessments unnecessary. The questions in this section aim to identify such situations.

B1 Did the authors make any attempt to control for confounding?	Confounding is a substantial problem in most non-randomized studies, and it is usually important to control for the important confounding factors.	Y for GAHT analysis; Y for GnRHa analysis
B2 If N/PN to B1: Is there sufficient potential for confounding that an unadjusted result should not be considered further?	If there is sufficient potential for confounding that an unadjusted result should not be considered further, then the result is judged to be at 'Critical risk of bias'.	

B3 Was the method of measuring the outcome inappropriate?	This question aims to identify methods of outcome measurement (data collection) that are unsuitable for the outcome they are intended to evaluate. This enables a rapid assessment that a result should be regarded as at 'Critical risk of bias'. The question does not aim to assess whether the choice of outcome being evaluated was *sensible* (e.g. because it is a surrogate or proxy for the main outcome of interest). In most circumstances, for pre-specified outcomes, the answer to this question will be 'N' or 'PN'. Answer 'Y or 'PY' if the method of measuring the outcome is inappropriate, for example because: (1) important ranges of outcome values fall outside levels that are detectable using the measurement method; or (2) the measurement instrument has been demonstrated to have such poor reliability or validity that estimates of the relationship between intervention and the measured outcome are not useful. (3) The measurement method differed substantially between people in the intervention and comparator groups, so that differences between the groups are not interpretable.	PN for GAHT analysis; PN for GnRHa analysis Both analyses used emergency department (ED) or inpatient visit for suicidality as the outcome

If the answer to either B2 or B3 is 'Yes' or 'Probably yes', the result should be considered to be at 'Critical risk of bias' and no further assessment is required.

C. Specify the analysis in the current study for which results are being assessed for risk of bias

Specify the outcome to which this result relates.

hazard ratio [HR] = 0.564, 95% confidence interval [95% CI: 0.36–0.89], p = 0.0137 -- Among TGD youth prescribed GAHT during our study period, there was a 43.6% reduction in risk of an ED or inpatient diagnosis of suicidality compared with those never prescribed GAHT during our study period or before GAHT initiation. HR = 0.79 [0.47–1.31], p = 0.357 -- TGD youth who were prescribed GnRHa therapy had a nonstatistically significant reduction in ED or inpatient suicidality diagnoses compared with those never prescribed GnRHa

C1. Specify the participant group on which this result was based.

For GAHT analysis: youth with gender dysphoria and prescribed GAHT (n = 1020) vs those with gender dysphoria but without GAHT (n = 2294) For GnRHa analysis: youth with gender dysphoria and prescribed GnRHa (n = 456) vs those with gender dysphoria but without GnRHa (n = 2865)

C2 to C3. Determine whether there is a need to consider time-varying confounding.

C2. Was the analysis based on splitting participants' follow up time according to intervention received, or was follow-up censored when participants in one group switched to another group (e.g. when comparison group participants started the intervention)?

☐	Yes for both analyses – "prescribed GAHT or GnRHa 'during the study period'"	Use Variant A of Domain 1 Proceed to next question

C3. If **Y** to C2, were intervention discontinuations or switches likely to be related to factors that are predictive of the outcome?

☐	Yes for both analyses – they may discontinue treatment if they do not identify as transgender, or they may be more likely to receive treatment (switch) if their family or their healthcare providers consider it beneficial or would decrease risk of suicidality	Use Variant A of Domain 1 Use Variant B of Domain 1

D. Specify a (hypothetical) target randomized trial specific to the study

Guidance notes

Evaluations of risk of bias are facilitated by considering the non-randomized study as an attempt to emulate a pragmatic randomized trial, which we refer to as the **target trial**. The first part of a ROBINS-I assessment for a particular study is to specify a target trial - the hypothetical randomized trial whose results should be the same as those from the non-randomized study under consideration, in the absence of bias. Its key characteristics are the types of participant (including exclusion/inclusion criteria) and

descriptions of the intervention strategy and comparator strategy. These issues were considered in more detail by Hernán (2016). Differences between the target trial for the individual non-randomized study and the generic research question of the review relate to issues of heterogeneity and/or generalizability rather than risk of bias.

Because it is hypothetical, ethics and feasibility need not be considered when specifying the target trial. For example there would be no objection to a target trial that compared individuals who did and did not start smoking, even though such a trial would be neither ethical nor feasible in practice.

Selection of a patient group that is eligible for a target trial may require detailed consideration, and lead to exclusion of many patients. For example, Magid et al (2010) studied the comparative effectiveness of ACE inhibitors compared to beta-blockers as second-line treatments for hypertension. From an initial cohort of 1.6m patients, they restricted the analysis population to (1) persons with incident hypertension, (2) who were initially treated with a thiazide agent, and (3) who had one of the two drugs of interest added as a second agent for uncontrolled hypertension, and (4) who did not have a contraindication to either drug. Their "comparative effectiveness" cohort included 15,540 individuals: less than 1% of the original cohort.

A note on terminology: Throughout ROBINS-I V2, we refer regularly to "intervention" and "comparator". The comparator may be an alternative active intervention, a control condition or no intervention at all.

We sometimes refer to the "intervention strategy" and "comparator strategy", because an intervention typically consists of a package of care or procedures, and may be implemented over a period of time rather than on a single occasion. Specification of the whole strategy of interest is particularly important when interest is in a 'per protocol' effect.

In non-randomized studies, assignment to the intervention or comparator is inferred from the recorded intervention for each participant. This is in contrast to randomized trials, in which participants are randomly assigned to the intervention or comparator. We refer to the participants assigned to each strategy as the "intervention group" and "comparator group".

Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology* 2016;183:758-64; doi:10.1093/aje/kwv254.

Magid DJ, Shetterly SM, Margolis KL, Tavel HM, O'Connor PJ, Selby JV, Ho PM. Comparative effectiveness of angiotensin-converting enzyme inhibitors versus beta-blocker as second-line therapy for hypertension. *Circulation: Cardiovascular Quality and Outcomes* 2010;3:453-458; doi:10.1161/CIRCOUTCOMES.110.940874.

D1. Specify the participants and eligibility criteria	youth with a diagnosis of gender dysphoria for GAHT youth with a diagnosis of gender dysphoria, at Tanner stage 2 for GnRHa
D2. Specify the intervention strategy	GAHT for GAHT analysis GnRHa for GnRHa analysis
D3. Specify the comparator strategy	Placebo for both

E. Decide on the effect of interest

E1. Is your aim for this study…?

☐ to assess the intention-to-treat effect (the effect of *assignment to* an intervention strategy or comparator strategy)

E2. **If the aim is to assess a per-protocol effect**, briefly define the changes to the intervention or comparator strategies that will be considered to be protocol deviations and, optionally, those changes that will not be considered. For example, the protocol deviations considered could be: “Starting intervention among comparator group participants, while acceptable changes could be “stopping intervention because of intervention-related toxicities occur or disease progression” or “changes to intervention after the trial baseline”.

F. Information sources

Guidance notes

Evaluation of a study should be based on the maximum possible amount of available information. In addition to published papers describing a study's methods and results, such information may be derived from the study protocol, unpublished reports or through correspondence with the study investigators.

Which of the following sources have you obtained to help you inform your risk of bias judgements (tick as many as apply)?

- ☐ Journal article(s)
- ☐ Study protocol
- ☐ Statistical analysis plan (SAP)
- ☐ Non-commercial registry record (e.g. ClinicalTrials.gov record)
- ☐ Company-owned registry record (e.g. GSK Clinical Study Register record)
- ☐ "Grey literature" (e.g. unpublished thesis)
- ☐ Conference abstract(s)
- ☐ Regulatory document (e.g. Clinical Study Report, Drug Approval Package)
- ☐ Individual participant data
- ☐ Research ethics application
- ☐ Grant database summary (e.g. NIH RePORTER, Research Councils UK Gateway to Research)
- ☐ Personal communication with investigator
- ☐ Personal communication with sponsor

Please specify any additional sources not listed above

Evaluation of confounding factors

Complete a row for each important confounding factor listed in advance (subsection (i) below); and either relevant to the setting of this particular study or identified by the study authors (subsection (ii)). **"Important" confounding factors are those for which, in the context of this study, adjustment is expected to lead to a meaningful change in the estimated effect of the intervention.**

Guidance notes

Confounding is of fundamental importance to the analysis and interpretation of non-randomized studies of the effect of interventions on outcomes. ROBINS-I addresses two types of confounding: baseline confounding and time-varying confounding.

Baseline confounding occurs when one or more prognostic factors, present before the start of the intervention, predict intervention received. Appropriate methods to control for confounders measured at baseline include stratification, regression, matching, standardization, and inverse probability weighting. The analysis may control for individual variables or for estimated propensity scores (inverse probability weighting is based on a function of the propensity score).

Time-varying confounding needs to be considered in studies that partition follow-up time for individual participants according to intervention received.

We use the term **confounding factor** for each broad source of potential confounding. It may not be possible to measure a factor well, and we distinguish between the confounding factor and the **variables** used to measure it. These variables may be used, for example, as covariates in a regression analysis.

In the context of a particular study, variables need not be included in the analysis: (a) if they are not associated with the outcome, conditional on intervention received (noting that lack of a statistically significant association is not evidence of a lack of association); (b) if they are not associated with intervention; (c) if adjustment makes no or minimal difference to the estimated effect of intervention on outcome; (d) because the confounder was addressed in the study design, for example by restricting to individuals with the same value of the confounder; (e) because a negative control demonstrates that there was unlikely to have

been confounding due to this variable or that uncontrolled confounding was likely to be minimal; or (f) because external evidence suggests that controlling for the variable is not necessary in the context of the study being assessed.

In some studies, researchers may include a very large set of potential confounding variables in an analysis without considering their associations with outcome and intervention. Users of ROBINS-I should focus on (i) the confounding factors they determined a priori to be important and (ii) other factors for which adjustment is expected to lead to an important change in the estimated effect of the intervention on the outcome in the context of the current study.

Users of ROBINS-I should evaluate the confounding factors that they prespecified as important for the intervention-outcome relationship under study. The tool also allows the user to evaluate a second list of any further confounding factors that are either relevant to the setting of this particular study or which the study authors identified as potentially important. It is likely that new ideas relating to confounding and other potential sources of bias will be identified after the drafting of the review protocol, and even after piloting data collection from studies selected for inclusion in the systematic review. For example, such issues may be identified because they are mentioned in the introduction and/or discussion of one or more papers. This could be addressed in practice by explicitly recording whether potential confounders or other sources of bias are mentioned in the paper.

In very rare situations it is possible that no confounding factors are present, either because interventions received are known to be unrelated to any prognostic factors for the outcome of interest, or because no such prognostic factors exist. In such situations, the risk of bias due to confounding may be assessed as low.

The purpose of this preliminary assessment of confounding factors is to review the extent to which the result being assessed was controlled for confounding, considering both the prespecified confounding factors and any further confounding factors identified as important in the context of the study being assessed. This enables users of ROBINS-I to answer the signalling questions for the Domain 1 assessment (Risk of bias due to confounding). "Important" confounding factors are those for which, in the context of this study, adjustment is expected to lead to an important change in the estimated effect of the intervention.

The preliminary assessment consists of the following steps for each confounding factor.

- determine which variables (if any) were measured for the factor;
- determine which of these variables were controlled for in the analysis;
- for variables that were not controlled for, look for evidence that controlling for the variable was not necessary in this particular study;
- determine whether the confounding factor was measured validly and reliably by the variables used to measure it (this is assessed at the level of the confounding factor rather than the level of the individual variables used to measure the factor);
- determine the likely direction of bias if the analysis fails to adjust for this variable (alone).

The direction of bias, if the analysis fails to adjust for a particular variable (alone), will be that the effect estimate is biased *upwards* or biased *downwards*. For example, if older age predicts that a particular intervention is more likely to be received and the outcome is mortality, then this confounding would bias the estimated effect downwards: unless we adjust for age the intervention will appear more positively associated with higher mortality than it should. In the presence of *positive confounding* (the confounder is positively associated with both intervention and outcome, or negatively associated with both intervention and outcome), the bias will be upwards. In the presence of *negative confounding* (the confounder is positively associated with intervention and negatively associated with outcome, or vice versa), the bias will be downwards.

(i) Important confounding factors listed in advance [for both analyses]						
Confounding factor	Measured variable(s) for this factor, if any	Was this variable (or were these variables) controlled for in the analysis? (Y / N)	If this confounding factor was controlled for, was it measured validly and reliably by this variable (or these variables)?* (NA / Y / PY / PN / N / NI)	If this confounding factor was not controlled for, is there evidence that controlling for it was unnecessary?** (NA / Y / PY / PN / N)	OPTIONAL: Is failure to adjust for this confounding factor expected to bias the effect estimate upwards or downwards? (Upward bias (overestimate the intervention effect) / Downward bias (underestimate the intervention effect) / No information or unpredictable)	Comments
Natal sex	Electronic health record sex	Y	Y			
Age of gender dysphoria diagnosis	Age at first diagnosis of gender dysphoria	Y	Y			
Starting age of intervention/duration of gender dysphoria	N	N		N		

Comorbidities	N	N		N		Baseline mental health is important factor influencing prognosis
Baseline anxiety	N	N		N		It is one major example of comorbidities
Baseline depression	N	N		N		It is one major example of comorbidities
Baseline suicidality	N	N		N		It is one major example of comorbidities
Psychological support	Behavioral health	N		N		Table 4, if any, is the proof that

	provider encounter					this factor should be controlled
Family support	N	N		N		
Social transition	N	N		N		
Surgery	N	N		N		

(ii) Additional important confounding factors relevant to the setting of this particular study, or identified by the study authors [for both analyses]						
Confounding factor	Measured variable(s) for this factor, if any	Was this variable (or were these variables) controlled for in the analysis? (Y / N)	If this confounding factor was controlled for, was it measured validly and reliably by this variable (or these variables)?* (NA / Y / PY / PN / N / NI)	If this confounding factor was not controlled for, is there evidence that controlling for it was unnecessary?** (NA / Y / PY / PN / N)	OPTIONAL: Is failure to adjust for this confounding factor expected to bias the effect estimate upwards or downwards? (Upward bias (overestimate the intervention effect) / Downward bias (underestimate the intervention effect) / No information or unpredictable)	Comments
Race	Race	Y	Y			

Health insurance	Type of health insurance	Y	Y			
Cointervention of GnRHa (For GAHT analysis)	Indication of GnRHa prescription	Y	Y			
Cointervention of GAHT (For GnRHa analysis)	Indication of GAHT prescription	Y	Y			

* “Validity” refers to whether the confounding variable or variables accurately measure the confounding factor, while “reliability” refers to the precision of the measurement (more measurement error means less reliability).

** In the context of a particular study, variables need not be included in the analysis: (a)) if they are measured validly and reliably and are not associated with the outcome, conditional on intervention (noting that lack of a statistically significant association is not evidence of a lack of association; (b) if they are measured validly and reliably and are not associated with intervention; (c) if they are measured validly and reliably and adjustment makes no or minimal difference to the estimated effect of the primary parameter; (d) because the confounder was addressed in the study design, for example by restricting to individuals with the same value of the confounder; (e) because a negative control demonstrates that there was unlikely to have been confounding due to this variable or that uncontrolled confounding was likely to be minimal; or (f) because external evidence suggests that controlling for the variable is not necessary in the context of the study being assessed”.

Risk of bias assessment

Responses underlined in green are potential markers for low risk of bias, and responses in red are potential markers for a risk of bias. Where questions relate only to sign posts to other questions, no formatting is used.

1. Bias due to confounding

Guidance notes

The questions in this domain focus on the confounding factors that were identified as important in the preliminary evaluation in section E.

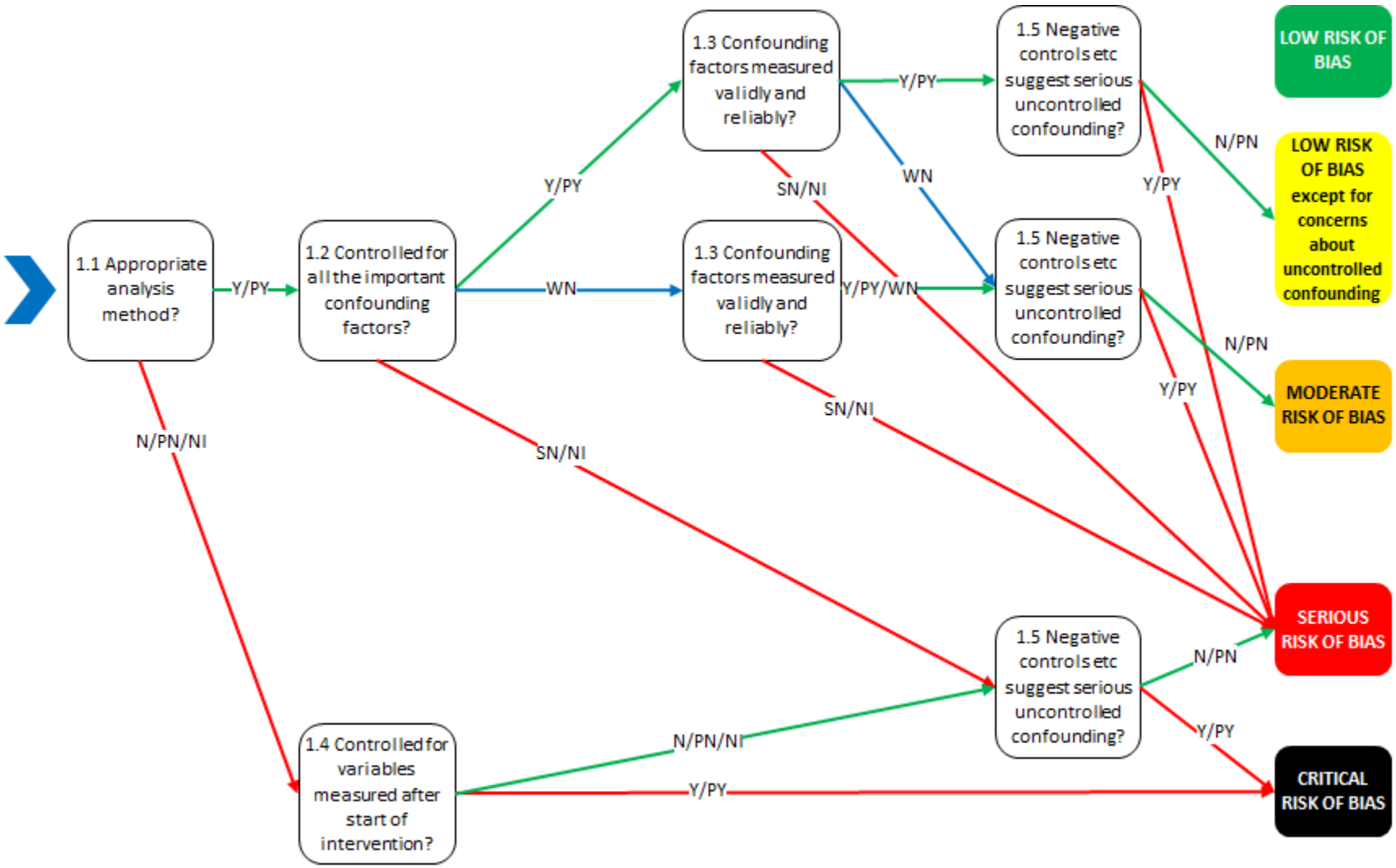
We use the term uncontrolled confounding to refer to confounding that was not controlled by the design or analysis of the study – and is therefore likely to bias the estimated effect of intervention. This may arise because (i) confounding factors were not (or could not) be measured; (ii) variables used to measure confounding factors were insufficient to characterize the confounding factor; or (iii) variables that characterize the confounding factor were measured but not included in the analysis.

Domain 1, Variant B (the analysis was based on splitting participants' follow up time according to intervention received, so both baseline and time-varying confounding need to be addressed – Y to C2 and Y to C3)

Signalling questions	Elaboration	Response
1.1 Did the authors use an analysis method that was appropriate to control for time-varying as well as baseline confounding?	The important confounding factors are those specified in the *Preliminary consideration of confounding factors*. Important baseline mental health conditions were not controlled. Cointerventions were not controlled. "An Anderson-Gill counting process regression model, using the robust sandwich variance estimator, was used to model recurrent events (ED or inpatient diagnosis for suicidality), which were assumed to be independent of each other" – this may not be appropriate to control for time-varying confounding	N for both analyses
1.2 If Y/PY to 1.1: Did the authors control for all the important baseline and time-varying confounding factors for which this was necessary?	-	NA
1.3 If Y/PY/WN to 1.2: Were confounding factors that were controlled for (and for which control was necessary) measured validly and reliably by the variables available in this study?	-	NA
1.4 If N/PN/NI to 1.1: Did the authors control for time-varying factors or other variables measured after the start of intervention?		PN

1.5 Did the use of negative controls, or other considerations, suggest serious unmeasured confounding?	Table 4 indeed suggested that unmeasured confounding factors exist.	PY for both analyses
Risk of bias judgement		Critical
Optional: What is the predicted direction of bias due to confounding?		

Algorithm for reaching default risk of bias judgement:

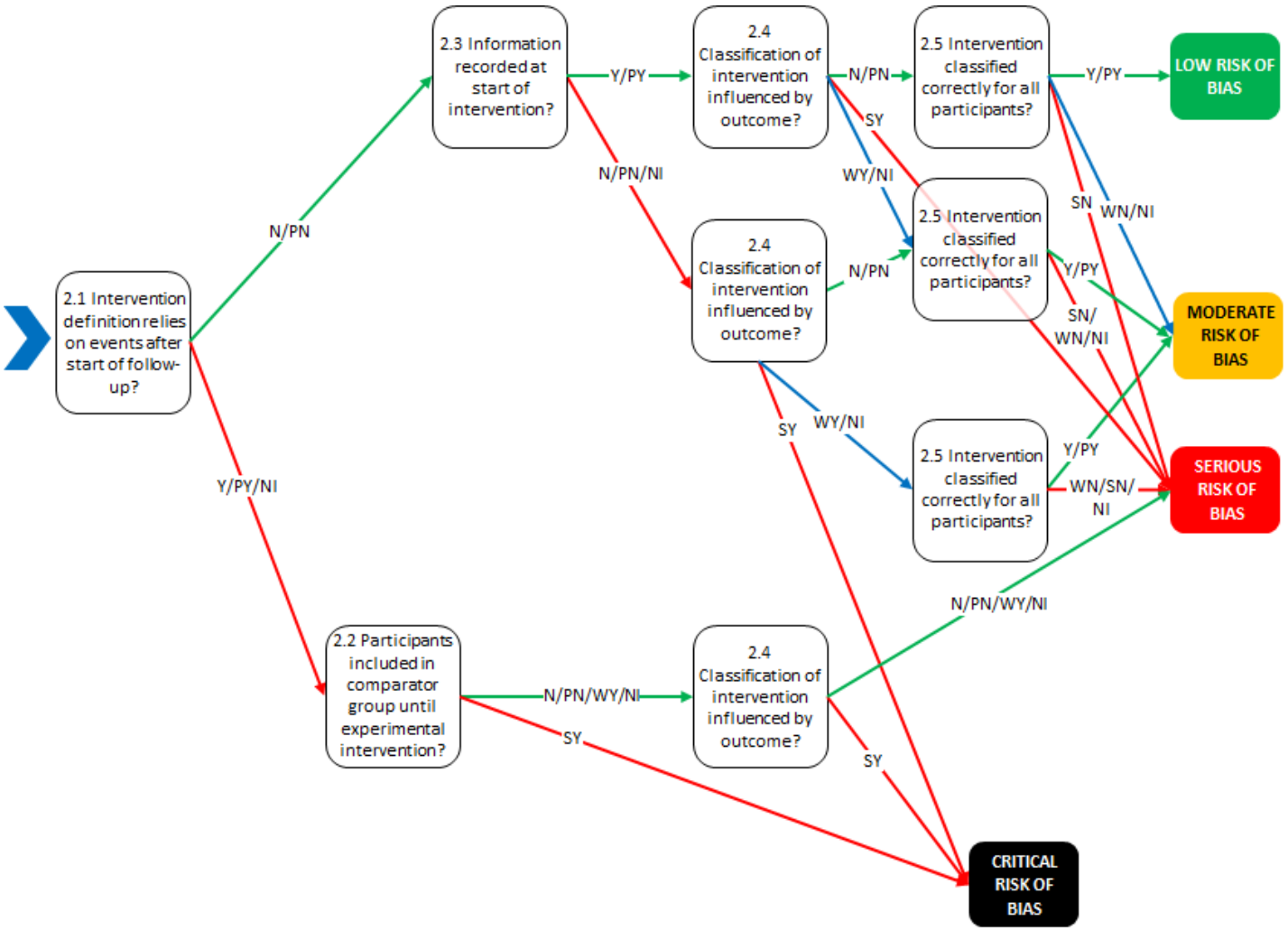


2. Bias in classification of interventions

Signalling questions	Elaboration	Response options
Questions about immortal time bias arising from definition of intervention groups		
2.1 Did assignment of participants to the intervention group or the comparator group rely on events or measurements that occurred after the start of follow up?	Once a participant received the treatment, this person would be assigned to the corresponding group and the follow up started.	PN for both analyses
2.2 If Y/PY to 2.1: Were participants included in the comparator group until they fulfilled the definition of the intervention (or vice versa)?	-	NA
Questions about differential misclassification		
2.3 If N/PN to 2.1: Was all information used to classify intervention and comparator groups recorded at or before the time the interventions started?	According to electronic health record (EHR), it happened at the same time	Y
2.4 Was classification of intervention status influenced by knowledge of the outcome or risk of the outcome?	EHR	PN
Question about non-differential misclassification		

2.5 If N/PN to 2.1 and WY/N/PN/NI 2.4: Was intervention status classified correctly for all, or nearly all, participants?	Probably nearly all	PY
Risk of bias judgement	See algorithm.	Low
Optional: What is the predicted direction of bias in classification of interventions?	-	Favours intervention / Favours comparator / Towards null /Away from null / Unpredictable

Algorithm for reaching default risk of bias judgement:



3. Bias in selection of participants into the study (or into the analysis)

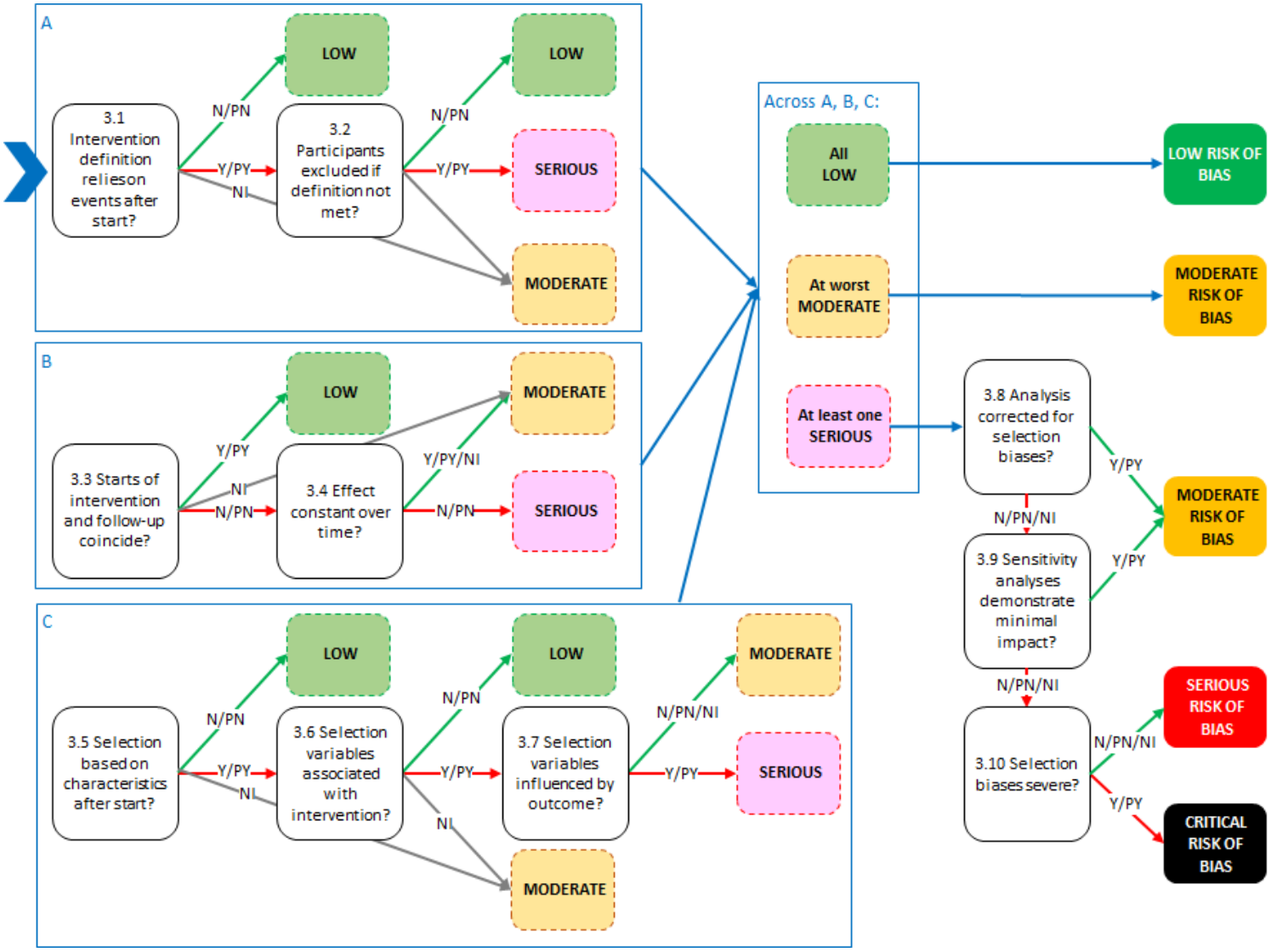
In the target trial, start of follow up is the time at which participants meet eligibility criteria and are assigned to interventions. In answering the signalling questions for this domain, consider what is the start of follow up in the study under consideration, for both the intervention and comparison groups.

Signalling questions	**Elaboration**	**Response options**
A. Questions about immortal time bias arising from definition of intervention groups		
3.1 (=2.1) Did assignment of participants to the intervention group or the comparator group rely on events or measurements that occurred after the start of follow up?	Once a participant received the treatment, this person would be assigned to the corresponding group and the follow up started.	PN for both analyses
3.2 If Y/PY to 3.1: Were participants excluded after the start of follow-up because they did not meet the definition of either the intervention or the comparator?	-	NA
B. Questions about prevalent user bias		
3.3 Were start of follow up and start of intervention the same for most participants?		PY
3.4 If N/PN to 3.3: Is the effect of intervention expected to be constant over the time period studied?	-	NA

C. Questions about other types of selection bias		
3.5 Was selection of participants into the study (or into the analysis) based on participant characteristics observed after the start of intervention (additional to the situations addressed in 3.1 and 3.3)?	Those with an ED or inpatient visit for suicidality within 30 days of their first PEDSnet visit were excluded	Y
3.6 If Y/PY to 3.5: Were the post-intervention variables that influenced selection likely to be associated with intervention?		PN
3.7 If Y/PY to 3.6: Were the post-intervention variables that influenced selection likely to be influenced by the outcome or a cause of the outcome?		NA
D. Questions about analysis, sensitivity analyses and severity of the problem		
3.8 If Y/PY to 3.2, N/PN 3.4 or Y/PY to 3.7: Is it likely that the analysis corrected for all of the potential selection biases identified in 3.1-3.2, 3.3-3.4 or 3.5-3.7 above?	-	NA
3.9 If N/PN to 3.8: Did sensitivity analyses demonstrate that the likely impact of the	-	NA

potential selection biases identified in 3.1-3.2, 3.3-3.4 or 3.5-3.7 above was minimal?		
3.10 If N/PN to 3.9: Were potential selection biases identified in 3.1-3.2, 3.3-3.4 or 3.5-3.7 above sufficiently severe that the result should not be included in a quantitative synthesis?	-	NA
Risk of bias judgement	See algorithm.	Low
Optional: What is the predicted direction of bias in selection of participants into the study?	-	Favours intervention / Favours comparator / Towards null /Away from null / Unpredictable

Algorithm for reaching default risk of bias judgement



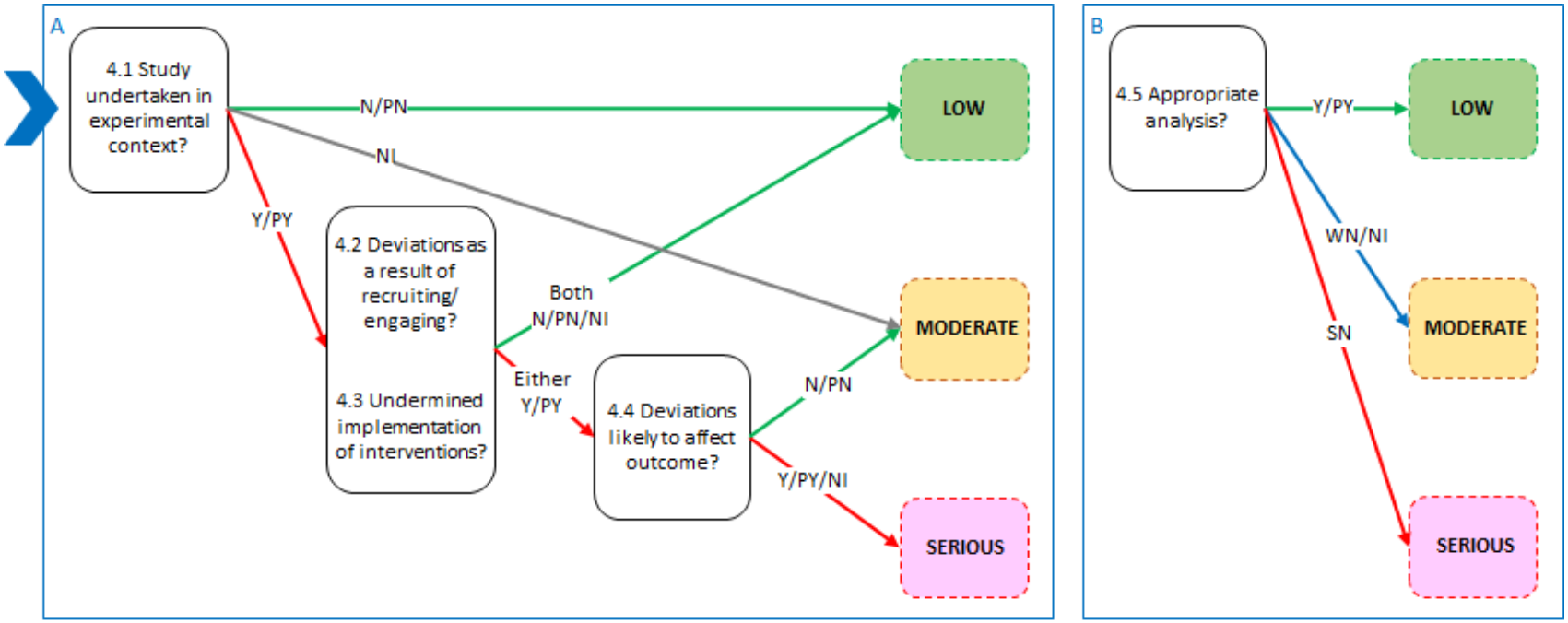
4. Bias due to deviations from intended interventions

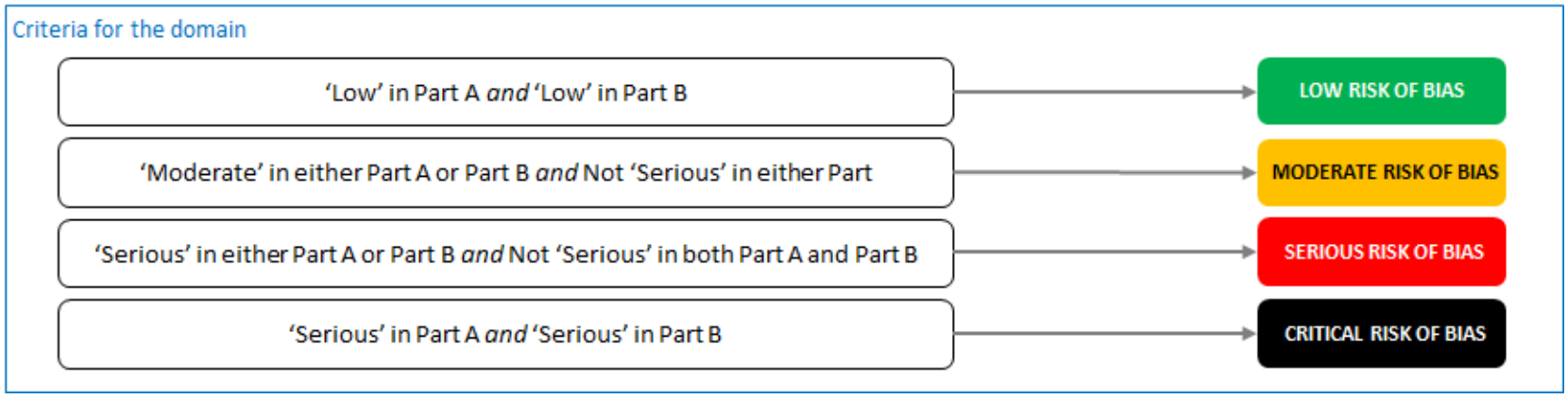
Domain 4, Variant A: Effect of assignment to intervention

Signalling questions	Elaboration	Response options
4.1 Was the study undertaken in an experimental context?	EHR review	N
4.2. If Y/PY to 4.1: Did participants deviate from the intended intervention as a result of the processes of recruiting and engaging them in the study?	-	NA
4.3. If Y/PY to 4.1: Did study personnel consciously or unconsciously undermine implementation of the intended interventions?	-	NA
4.4. If Y/PY/NI to 4.2 or 4.3: Were these deviations from intended intervention likely to have affected the outcome?	Cointervention, or non-compliant may have happened	PY
4.5. Was an appropriate analysis used to estimate the effect of assignment to intervention?		PY

Risk of bias judgement	See algorithm.	Low
Optional: What is the predicted direction of bias in classification of interventions?	If the likely direction of bias can be predicted, it is helpful to state this. The direction might be characterized as being in favour of the intervention, as being in favour of the comparator, or as towards (or away from) the null.	Favours intervention / Favours comparator / Towards null /Away from null / Unpredictable

Algorithm for reaching default risk of bias judgement (effect of assignment to intervention):





5. Bias due to missing data

Guidance notes

Missing outcome data may arise, among other reasons, through attrition (loss to follow up), missed appointments and incomplete data collection. Additionally, in non-randomized studies data may be missing for characteristics including interventions received and confounders.

A general rule for consideration of bias due to missing data is that we they should consider biases introduced by the missing data, compared with the effect estimate from an analysis in which all the data we intended to collect were available. Unfortunately, a single threshold for an acceptable proportion of missing data cannot meaningfully be defined. For example, a result based on 95% complete outcome data might be biased if the outcome was rare and if reasons for missing outcome data were strongly related to intervention group. Therefore, the potential for bias due to missing data should be assessed unless complete data on intervention status, the outcome and confounding variables were available for all, or nearly all, participants.

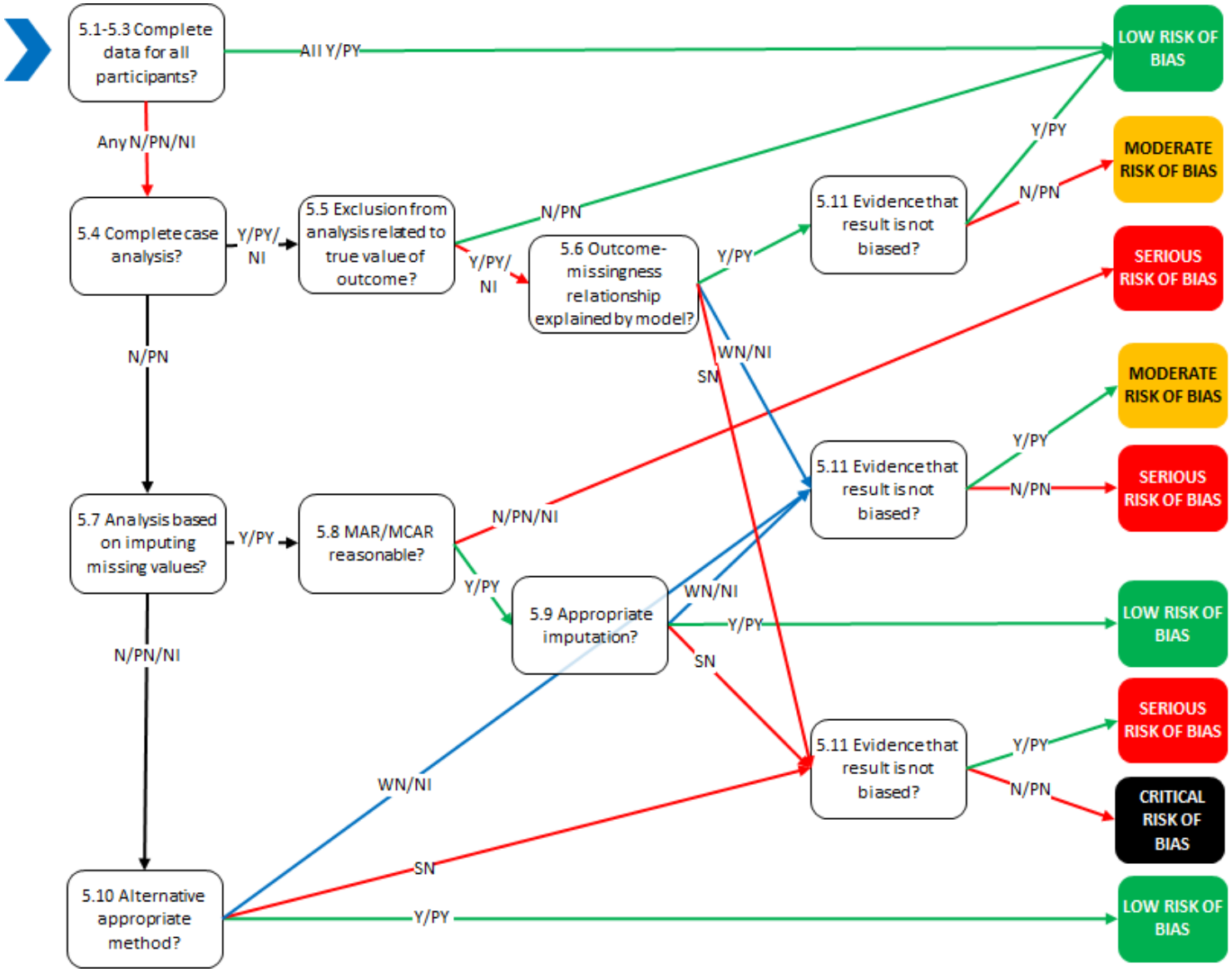
Considerations of bias due to missing data depend on how the analysis accounted for the missing data. Different signalling questions should be answered depending on three types of analysis. The first is that a **complete case analysis**, restricted to participants with complete data on all the intervention, outcome and confounding variables, was performed. In this situation, an important consideration is whether missingness of individual participants from the analysis is related to the true value of the outcome for those participants. The second is that missing data were **imputed**, which means that estimated or assumed values were assigned to participants with missing data. Imputed data should not lead to bias if the data are 'missing at random' (see the elaboration for signalling question 5.8) and an appropriate imputation method is applied. Other types of analysis are addressed by a separate, general, signalling question. The final signalling question asks whether sensitivity analyses were performed that demonstrated that the impact of missing data is minimal.

Signaling questions	Elaboration	Response options
5.1 Were complete data on intervention status available for all, or nearly all, participants?	Based on EHR	Y
5.2 Were complete data on the outcome available for all, or nearly all, participants?	Based on EHR	Y
5.3 Were complete data on important confounding variables available for all, or nearly all, participants?	Important baseline mental health conditions or cointerventions were not available	N
5.4 If N/PN/NI to 5.1, 5.2 or 5.3: Is the result based on a complete case analysis?		Y
5.5 If Y/PY/NI to 5.4: Was exclusion from the analysis because of missing data (in intervention, confounders or the outcome) likely to be related to the true value of the outcome?	loss to follow up or withdrawal generally low	PN
5.6 If Y/PY/NI to 5.5: Is the relationship between the outcome and missingness likely to be	-	NA

explained by the variables in the analysis model?		
5.7 If N/PN to 5.4: Was the analysis based on imputing missing values?	-	NA
5.8 If Y/PY to 5.7: Is it reasonable to assume that data were 'missing at random' (MAR) or 'missing completely at random' (MCAR)?	-	NA
5.9 If Y/PY to 5.8: Was imputation performed appropriately?	-	NA
5.10 If N/PN/NI to 5.7: Was an appropriate alternative method used to correct for bias due to missing data?	-	NA
5.11 If PN/N/NI to 5.1, 5.2 or 5.3 AND (Y/PY/NI to 5.5 OR (Y/PY to 5.8 AND WN/SN/NI to 5.9) OR WN/SN/NI to 5.10): Is there evidence that the result was not biased by missing data?	No details	PN
Risk of bias judgement	See algorithm.	Serious
Optional: What is the predicted direction of bias due to missing data?	-	Favours intervention / Favours

		comparator / Towards null /Away from null / Unpredictable

Algorithm for reaching default risk of bias judgement:



6. Bias in measurement of the outcome

Guidance notes

Bias may be introduced if outcomes are misclassified or measured with error. Misclassification or measurement error of outcomes may be non-differential or differential.

Non-differential measurement error is unrelated to the intervention received. It can be systematic (for example when measurement of blood pressure is consistently 5 units too high in every participant) – in which case it will not affect precision or cause bias; or it can be random (for example when measurement of blood pressure is sometimes too high and sometimes too low in a manner that does not depend on the intervention or the outcome) – in which case it will affect precision without causing bias.

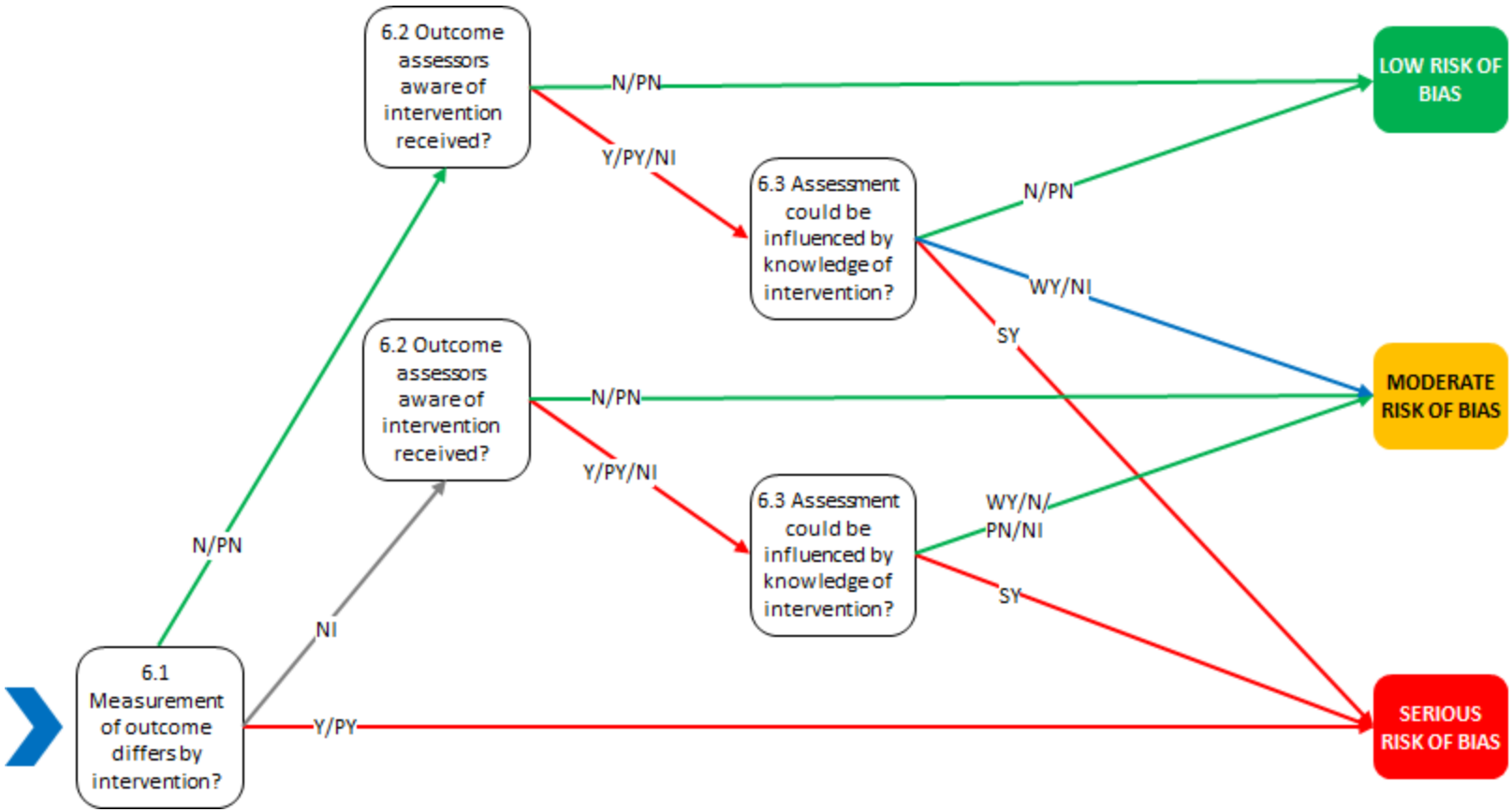
Differential measurement error is measurement error related to intervention received. It will bias the intervention-outcome relationship. This is often referred to as detection bias. Examples of situations in which detection bias can arise are (i) if outcome assessors are aware of intervention received (particularly when the outcome is subjective); (ii) different methods (or intensities of observation) are used to assess outcomes of participants receiving different interventions; and (iii) measurement errors are related to intervention received (or to a confounder of the intervention-outcome relationship).

Blinding of outcome assessors aims to prevent systematic differences in measurements according to intervention received. However, blinding is frequently not possible or not performed for practical reasons.

Signalling questions	**Elaboration**	**Response options**
6.1 Could measurement or ascertainment of the outcome have differed between intervention groups?	Participants may be more likely to present to ED or clinic due to suicidality depending on the treatment and co-interventions that they have received	Y

6.2 Were outcome assessors aware of the intervention received by study participants?	EHR	PN
6.3 If Y/PY/NI to 6.2: Could assessment of the outcome have been influenced by knowledge of the intervention received?	-	NA
Risk of bias judgement	See algorithm.	Serious
Optional: What is the predicted direction of bias in measurement of outcomes?	If the likely direction of bias can be predicted, it is helpful to state this. The direction might be characterized as being in favour of the intervention, as being in favour of the comparator, or as towards (or away from) the null.	Favours intervention / Favours comparator / Towards null /Away from null / Unpredictable

Algorithm for reaching default risk of bias judgement:



7. Bias in selection of the reported result

Guidance notes

Selective reporting can arise for both harms and benefits of an intervention, although the motivations (and direction of bias) underlying selective reporting of effect estimates for harms and benefits may differ. Selective reporting may arise, for example, from a desire for findings to be newsworthy (or sufficiently noteworthy to merit publication), or from commercial considerations, or from a desire to demonstrate that there is not evidence of a harmful effect of an intervention.

Selective outcome reporting occurs when the effect estimate for an outcome measurement was selected from among analyses of multiple outcome measurements for the outcome domain. Examples include: use of multiple measurement instruments (e.g. pain scales) and reporting only the most favourable result; reporting only the most favourable subscale (or a subset of subscales) for an instrument when measurements for other subscales were available; reporting only one or a subset of time points for which the outcome was measured.

Selective analysis reporting occurs when results are selected from effects estimated in multiple ways: e.g. carrying out analyses of both change scores and post-intervention scores adjusted for baseline; multiple analyses of a particular measurement with and without transformation; multiple analyses of a particular outcome with and without adjustment for potential confounders (or with adjustment for different sets of potential confounders); multiple analyses of a particular outcome with and without, or with different, methods to take account of missing data; a continuously scaled outcome converted to categorical data with different cut-points; multiple composite outcomes analysed for one outcome domain, but results were reported only for one (or a subset) of the composite outcomes. (Reporting an effect estimate for an unusual composite outcome might be evidence of such selective reporting.)

Selection of a subgroup from a larger cohort: The cohort for analysis may have been selected from a larger cohort for which data were available on the basis of a more interesting finding. Subgroups defined in unusual ways (e.g. an unusual classification of subgroups by dose or dose frequency) may provide evidence of such selective reporting.

The best evidence that results were not selectively reported is available if a pre-specified, publicly available analysis plan is available (e.g. from a link in a publication or from an online platform) and is in line with the reported results. Protocols for non-

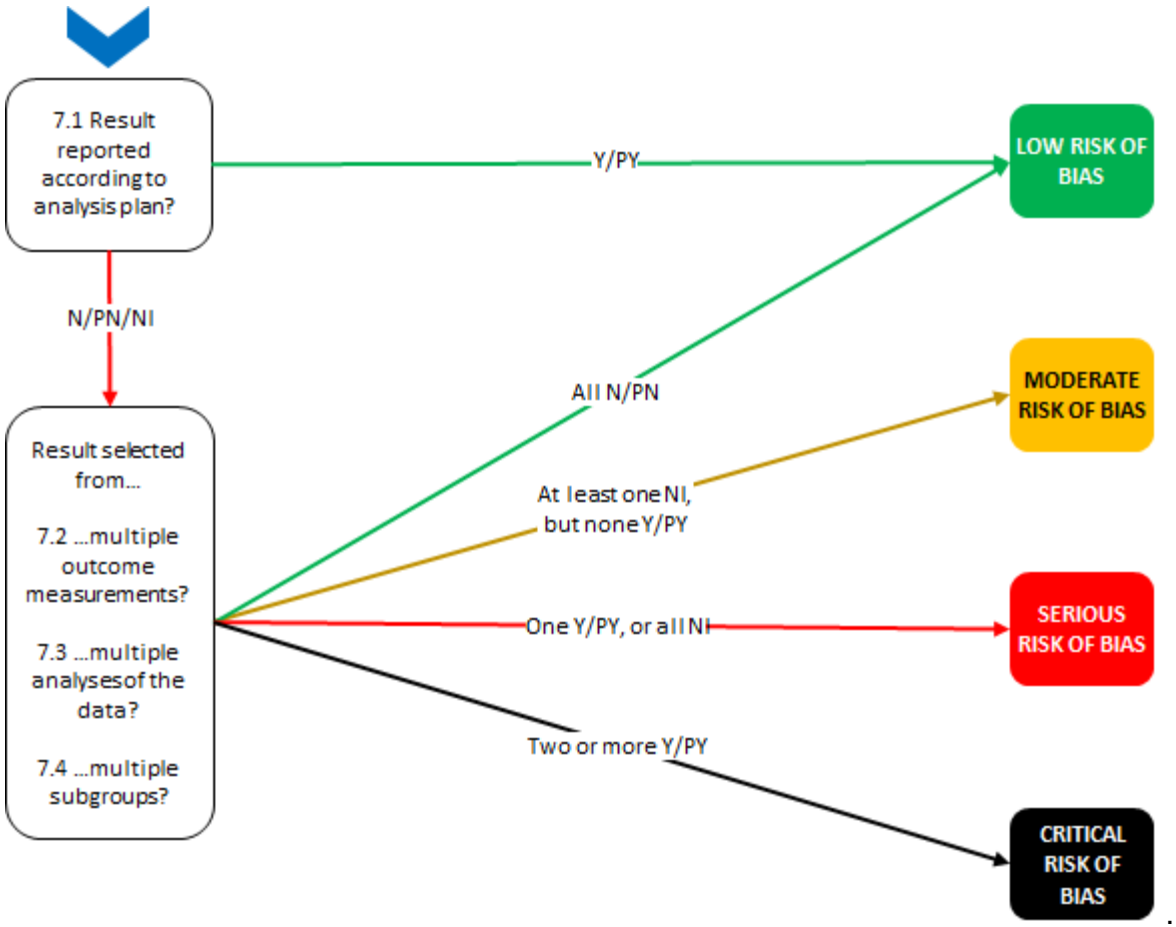
randomized studies are increasingly being registered, although there is inconsistency across platforms (Malmsiø et al, 2022). An analysis plan that is sufficiently detailed to permit full assessment of selective reporting may seldom be available for observational studies. In the absence of a protocol or analysis plan, clues can sometimes be gained by comparing Methods sections with Results sections.

Malmsiø D, Frost A, Hróbjartsson A. A scoping review finds that guides to authors of protocols for observational epidemiological studies varied highly in format and content. J Clin Epidemiol. 2022 Dec 20;154:156-166. doi: 10.1016/j.jclinepi.2022.12.012.

Signalling questions	**Elaboration**	**Response options**
7.1 Was the result reported in accordance with an available, pre-determined analysis plan?		NI
Is the numerical result being assessed likely to have been selected, on the basis of the results, from...		
7.2 ... multiple outcome *measurements* (e.g. scales, definitions, time points) within the outcome domain?		NI
7.3 ... multiple *analyses* of the data?	Selection on the basis of the results arises from a desire for findings to be newsworthy, sufficiently noteworthy to merit	PY

	publication, or to confirm a prior hypothesis. For GAHT, the authors reported only the main results. But for GnRHa, after non-significant results, the authors reported further analysis of comparing GAHT versus no GAHT among people receiving GnRHa.	
7.4 ... multiple *subgroups*?		NI
Risk of bias judgement	See algorithm.	Serious
Optional: What is the predicted direction of bias in selection of the reported result?	If the likely direction of bias can be predicted, it is helpful to state this. The direction might be characterized as being in favour of the intervention, as being in favour of the comparator, or as towards (or away from) the null.	Favours intervention / Favours comparator / Towards null /Away from null / Unpredictable

Algorithm for reaching default risk of bias judgement:



Overall risk of bias

Guidance notes

ROBINS-I defaults to setting the overall risk of bias for a result to be equal to the risk-of-bias judgement for the domain with the greatest risk of bias. For example, if the 'worst' judgement across domains is of serious risk of bias, then the result would be judged as at serious risk of bias overall. However, the user may override this to judge the result to be at greater risk of bias if there are problems in several domains. For example, if several domains are assessed to be at serious risk of bias, and it is considered that these problems are likely to be compounded, then it may be reasonable to judge the result to be at critical risk of bias overall. Predicting the direction of bias overall may be difficult. Risk-of-bias judgements for the individual domains might be used to inform the influence of that domain to the likely direction of bias overall.

Overall risk of bias	See algorithm.	Critical risk
What is the predicted direction of bias?	If the likely direction of bias can be predicted, it is helpful to state this. The direction might be characterized as being in favour of the intervention, as being in favour of the comparator, or as towards (or away from) the null. Alternatively, if the direction is driven by bias due to confounding, the direction may be an upwards bias (overestimate the effect) or a downward bias (underestimate the effect).	Upward bias (overestimate the effect) / Downward bias (underestimate the effect) / Favours intervention / Favours comparator / Towards null /Away from null / Unpredictable

Algorithm for reaching overall risk of bias judgement:

Judgement	Interpretation	How reached
Low risk of bias except for concerns about uncontrolled confounding	There is the possibility of uncontrolled confounding that has not been controlled for (given the observational nature of the study), but otherwise little or no concern about bias in the result	*Low risk of bias except for concerns about uncontrolled confounding* in Domain 1 and *Low risk of bias* in all other domains
Moderate risk of bias	There is some concern about bias in the result, although it is not clear that there is an important risk of bias	At least one domain is at *Moderate risk of bias*, but no domains are at *Serious risk of bias* or *Critical risk of bias*
Serious risk of bias	The study has some important problems: characteristics of the study give rise to a serious risk of bias in the result	At least one domain is at *Serious risk of bias*, but no domains are at *Critical risk of bias* OR Several domains are at *Moderate*, leading to an additive judgement of *Serious risk of bias*
Critical risk of bias	The study is very problematic: characteristics of the study give rise to a critical risk of bias in the result, such that the result should generally be excluded from evidence syntheses.	At least one domain is at *Critical risk of bias* OR Several domains are at *Serious risk of bias*, leading to an additive judgement of *Criticial risk of bias*

Bibliography

Aaron, D. G., & Konnoth, C. (2025). The future of gender-affirming care—A law and policy perspective on the Cass Review. *New England Journal of Medicine*, *392*(6), 526–528.

Abbruzzese, E., Levine, S. B., & Mason, J. W. (2023). The myth of “reliable research” in pediatric gender medicine: A critical evaluation of the Dutch studies—and research that has followed. *Journal of Sex and Marital Therapy*, *49*(6), 673–699.

Achille, C., Taggart, T., Eaton, N. R., Osipoff, J., Tafuri, K., Lane, A., & Wilson, T. A. (2020). Longitudinal impact of gender-affirming endocrine intervention on the mental health and well-being of transgender youths: preliminary results. *International Journal of Pediatric Endocrinology*, *2020*(8).

Akl, E. A., Hakoum, M., Khamis, A., Khabsa, J., Vassar, M., & Guyatt, G. (2022). A framework is proposed for defining, categorizing, and assessing conflicts of interest in health research. *Journal of Clinical Epidemiology*, *149*, 236–243.

American Academy of Pediatrics. (May 1, 2025). AAP statement on the HHS report Treatment for Pediatric Gender Dysphoria. https://web.archive.org/web/20250927063305/https://www.aap.org/en/news-room/news-releases/aap/2025/aap-statement-on-hhs-report-treatment-for-pediatric-gender-dysphoria/

American Psychiatric Association. (May 1, 2025). APA statement on access to treatment for transgender, gender diverse, and nonbinary people. https://web.archive.org/web/20250502165038/https://updates.apaservices.org/statement-on-access-to-treatment-for-transgender-gender-diverse-and-nonbinary-people

Ashley, F. (2023). Interrogating gender-exploratory therapy. *Perspectives on Psychological Science*, *18*(2), 472–481.

Baxendale, S. (2025). How to be a better doctor: Recognizing how cognitive biases shape—and distort—clinical evidence. *British Journal of Hospital Medicine*, *86*(2), 1–14.

Biggs, M. (2020). Puberty blockers and suicidality in adolescents suffering from gender dysphoria. *Archives of Sexual Behavior*, *49*, 2227–2229.

Block, J. (2024). Gender medicine in the US: how the Cass review failed to land. *BMJ*, *385*, q1141.

Borah, L., Zebib, L., Sanders, H. M., Lane, M., Stroumsa, D., & Chung, K. C. (2023). State restrictions and geographic access to gender-affirming care for transgender youth. *JAMA*, *330*(4), 375–378.

Budge, S. L., Abreu, R. L., Flinn, R. E., Donahue, K. L., Estevez, R., Olezeski, C. L., ... & Allen, B. J. (2024). Gender affirming care is evidence based for transgender and gender-diverse youth. *Journal of Adolescent Health*, *75*(6), 851–853.

Bumphenkiatikul, T., Sakpetch, T., Sungnuch, P., Huang, X., Hataiyusuk, S., & Wainipitapong, S. (2025). Non-medical gender affirming practices among transgender individuals: A systematic review on the health implications of chest binding and genital tucking. *International Journal of Sexual Health.* https://doi.org/10.1080/19317611.2025.2560416

Button, K. S., Kounali, D., Thomas, L., Wiles, N. J., Peters, T. J., Welton, N. J., ... & Lewis, G. (2015). Minimal clinically important difference on the Beck Depression Inventory-II according to the patient's perspective. *Psychological Medicine*, *45*(15), 3269–3279.

Cass, H. (2024). *Independent Review of Gender Identity Services for Children and Young People: Final Report*. https://webarchive.nationalarchives.gov.uk/ukgwa/20250310143933/https://cass.independent-review.uk/home/publications/final-report/

Chelliah, P., Lau, M., & Kuper, L. E. (2024). Changes in gender dysphoria, interpersonal minority stress, and mental health among transgender youth after one year of hormone therapy. *Journal of Adolescent Health*, *74*(6), 1106–1111.

Chen, D., Berona, J., Chan, Y.-M., Ehrensaft, D., Garofalo, R., Hidalgo, M. A., Rosenthal, S. M., Tishelman, A. C., & Olson-Kennedy, J. (2023). Psychosocial functioning in transgender youth after 2 years of hormones. *New England Journal of Medicine*, 388(3), 240–250.

Cheung, C. R., Abbruzzese, E., Lockhart, E., Maconochie, I. K., & Kingdon, C. C. (2025). Gender medicine and the Cass Review: Why medicine and the law make poor bedfellows. *Archives of Disease in Childhood*, *110*, 251–255.

Chiles v. Salazar (No. 24-539), *Dr. Jack L. Turban and Dr. Lisa R. Fortuna*, Amicus curiae brief (Supreme Court of the United States, August 26, 2025). https://web.archive.org/web/20251002163018/https://www.supremecourt.gov/DocketPDF/24/24-539/370817/20250826173111769_24-539%20Amicus%20Brief.pdf

Coalition for the Advancement and Application of Psychological Science. (2021, July 26). Position statement on rapid onset gender dysphoria (ROGD). https://web.archive.org/web/20251014213947/https://www.caaps.co/rogd-statement

Cochrane. 2024. ROBINS-I V2 tool. https://web.archive.org/web/20250829000548/https://www.riskofbias.info/welcome/robins-i-v2

Cohn, J. 2025. Censorship of essential debate in gender medicine research. *Journal of Controversial Ideas*, *5*(2), 3.

Committee on Bioethics, Katz, A. L., Macauley, R. C., Mercurio, M. R., Moon, M. R., Okun, A. L., Opel, D. J., & Statter, M. B. (2016). Informed consent in decision-making in pediatric practice. *Pediatrics*, *138*(2), e20161484.

Commonwealth of Massachusetts. (2017). 130 Code of Massachusetts Regulations, §450.204: Medical necessity.

https://web.archive.org/web/2/https://www.law.cornell.edu/regulations/massachusetts/130-CMR-450-204

Coverdale, J. H., Aggarwal, R., Balon, R., Beresin, E. V., Guerrero, A. P., Louie, A. K., ... & Brenner, A. M. (2024). Practical advice for preventing problems when referencing the literature. *Academic Psychiatry*, *48*(1), 5–9.

deMayo, B. E., Gallagher, N. M., Leshin, R. A., & Olson, K. R. (2025). Stability and change in gender identity and sexual orientation across childhood and adolescence. *Monographs of the Society for Research in Child Development*, *90*(1–3).

Dopp, A. R., Peipert, A., Buss, J., De Jesus-Romero, R., Palmer, K., & Lorenzo-Luaces, L. (2024). *Interventions for gender dysphoria and related health problems in transgender and gender-expansive youth: A systematic review of benefits and risks to inform practice, policy, and research*. RAND. https://web.archive.org/web/20250501163212/https://www.rand.org/pubs/research_reports/RRA3223-1.html

Dowshen, N., Baker, K., Garofalo, R., Chen, D., Inwards-Breland, D. J., Sequeira, G., ... & McNamara, M. (2025). A critical scientific appraisal of the Health and Human Services report on pediatric gender dysphoria. *Journal of Adolescent Health*, *77*(3), 342–345.

Eguale, T., Buckeridge, D. L., Verma, A., Winslade, N. E., Benedetti, A., Hanley, J. A., & Tamblyn, R. (2016). Association of off-label drug use and adverse drug events in an adult population. *JAMA Internal Medicine*, *176*(1), 55.

Ehrensaft, D. (1992). Preschool child sex abuse: The aftermath of the Presidio case. *American Journal of Orthopsychiatry*, *62*, 234–244.

Ehrensaft, D. (2016, February 26). A developmental perspective of transgender and gender nonconforming youth and a collaborative model of care. Jon E. Nadherny/Calciano Memorial Youth Symposium. https://youtu.be/30rEjumFaDY, 1:01–2:12.

Figliola, J. (2025, April 18). The dangerous and muddled logic of gender medicine. *City Journal*. https://web.archive.org/web/20250501184450/https://www.city-journal.org/article/gender-medicine-trans-identifying-minors-wpath

Georges, E., Brown, E. C., & Cohen, R. S. (2024). Prohibition of gender-affirming care as a form of child maltreatment: reframing the discussion. *Pediatrics*, *153*(1), e2023064292.

Ghorayshi, A. (2024, May 13). Hilary Cass says U.S. doctors are 'out of date' on youth gender medicine. *New York Times*. https://web.archive.org/web/20250501184519/https://www.nytimes.com/2024/05/13/health/hilary-cass-transgender-youth-puberty-blockers.html

Guyatt, G. H., Oxman, A. D., Kunz, R., Atkins, D., Brozek, J., Vist, G., ... & Schünemann, H. J. (2011). GRADE guidelines: 2. Framing the question and deciding on important outcomes. *Journal of Clinical Epidemiology*, *64*(4), 395-400.

Guyatt, G. H., Rennie, D., Meade, M., & Cook, D. (Eds.). (2015). *Users' Guides to the Medical Literature: Essentials of Evidence-Based Clinical Practice* (3rd edition). McGraw-Hill Education.

Halim, M. L. D., Ruble, D. N., Tamis-LeMonda, C. S., Shrout, P. E., & Amodio, D. M. (2017). Gender attitudes in early childhood: Behavioral consequences and cognitive antecedents. *Child Development*, *88*(3), 882–899.

Hembree, W. C., Cohen-Kettenis, P. T., Gooren, L., Hannema, S. E., Meyer, W. J., Murad, M. H., Rosenthal, S. M., Safer, J. D., Tangpricha, V., & T'Sjoen, G. G. (2017). Endocrine treatment of gender-dysphoric/gender-incongruent persons: An Endocrine Society clinical practice guideline. *Endocrine Practice*, *23*(12), 1437–1437.

Hembree, W. C., Cohen-Kettenis, P., Delemarre-van de Waal, H. A., Gooren, L. J., Meyer, W. J., Spack, N. P., Tangpricha, V., & Montori, V. M. (2009). Endocrine

treatment of transsexual persons: An Endocrine Society clinical practice guideline. *Journal of Clinical Endocrinology and Metabolism*, *94*(9), 3132–3154.

Higgins, J. P. T., Chandler, J., Cumpston, J., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane Handbook for Systematic Reviews of Interventions* (2nd ed.). John Wiley & Sons Ltd.

Hoffmann-Eßer, W., Siering, U., Neugebauer, E. A., Brockhaus, A. C., McGauran, N., & Eikermann, M. (2018). Guideline appraisal with AGREE II: Online survey of the potential influence of AGREE II items on overall assessment of guideline quality and recommendation for use. *BMC Health Services Research*, *18*(1), 143.

Horton, C. (2025). The Cass Review: Cis-supremacy in the UK's approach to healthcare for trans children. *International Journal of Transgender Health*, *26*(4), 1120-1144.

Huit, T. Z., Coyne, C., & Chen, D. (2024). State of the science: Gender-affirming care for transgender and gender diverse youth. *Behavior Therapy*, *55*(6), 1335–1347.

Karlsson, J., Beaufils, P. (2013). Legitimate division of large data sets, salami slicing and dual publication, where does a fraud begin? *Knee Surgery, Sports Traumatology, Arthroscopy*, *21*, 751–752.

Kellner, S., & Bascom, E. (May 15, 2025). Experts scrutinize HHS report on gender-affirming care for minors. *Healio*. https://web.archive.org/web/20250515205933/https://www.healio.com/news/pediatrics/20250515/experts-scrutinize-hhs-report-on-genderaffirming-care-for-minors

Kepp, K. P., Aavitsland, P., Ballin, M., Balloux, F., Baral, S., Bardosh, K., Bauchner, H., Bendavid, E., Bhopal, R., Blumstein, D. T., Boffetta, P., Bourgeois, F., Brufsky, A., Collignon, P. J., Cripps, S., Cristea, I. A., Curtis, N., Djulbegovic, B., Faude, O., … Ioannidis, J. P. A. (2024). Panel stacking is a threat to consensus statement validity. *Journal of Clinical Epidemiology*, *173*, 111428.

Kincaid, E. (2025, April 18). Medical societies call for BMJ to retract 'misleading and irresponsible' guideline. *Retraction Watch*.

https://web.archive.org/web/20250419000709/https://retractionwatch.com/2025/04/18/bmj-clinical-guideline-spine-pain-medical-societies-call-for-retraction/

Kingdon, C., Stingelin-Giles, N., & Cass, H. (2025). The Cass Review: Distinguishing fact from fiction. *American Journal of Bioethics*, *25*(6), 5–10.

Kuper, L. E., Stewart, S., Preston, S., Lau, M., & Lopez, X. (2020). Body dissatisfaction and mental health outcomes of youth on gender-affirming hormone therapy. *Pediatrics*, *145*(4).

Labcorp. (2024). Estrogens, total. https://web.archive.org/web/20240913082123/https://www.labcorp.com/tests/004549/estrogens-total

Labcorp. (2025a). Estradiol, sensitive, LC/MS. https://web.archive.org/web/20250422164130/https://www.labcorp.com/tests/140244/estradiol-sensitive-lc-ms

Labcorp. (2025b). Testosterone, total, women, children, and hypogonadal males, LC/MS-MS. https://web.archive.org/web/20250818083950/https://www.labcorp.com/tests/070001/testosterone-total-women-children-and-hypogonadal-males-lc-ms-ms

Leung, P. T., Macdonald, E. M., Stanbrook, M. B., Dhalla, I. A., & Juurlink, D. N. (2017). A 1980 letter on the risk of opioid addiction. *New England Journal of Medicine*, *376*(22), 2194–2195.

Ludvigsson, J. F., Adolfsson, J., Höistad, M., Rydelius, P.-A., Kriström, B., & Landén, M. (2023). A systematic review of hormone treatment for children with gender dysphoria and recommendations for research. *Acta Paediatrica*, *112*(11), 2279–2292.

Madrigal-Borloz, V. (2020). *Practices of so-called “conversion therapy”: Report of the independent expert on protection against violence and discrimination based on sexual orientation and gender identity*. United Nations.

https://web.archive.org/web/20250222022726/https://docs.un.org/en/A/HRC/44/53

Matheny Antommaria, A. H., Kelleher, M., & Peterson, R. J. (2025). Quality of evidence and strength of recommendations in American Academy of Pediatrics' guidelines. *Pediatrics*, *155*(4), e2024067836

McDeavitt, K. (2024). Paediatric gender medicine: longitudinal studies have not consistently shown improvement in depression or suicidality. *Acta Paediatrica*, *113*(8), 1757–1771.

McDeavitt, K., Cohn, J., & Levine, S. B. (2025). Critiques of the Cass Review: Fact-checking the peer-reviewed and grey literature. *Journal of Sex and Marital Therapy*, *51*(2), 175–199.

McGregor, K., McKenna, J. L., Williams, C. R., Barrera, E. P., & Boskey, E. R. (2024). Association of pubertal blockade at Tanner 2/3 with psychosocial benefits in transgender and gender diverse youth at hormone readiness assessment. *Journal of Adolescent Health*, *74*(4), 801–807.

McNamara, M., Abdul-Latif, H., Boulware, S. D., Kamody, R., Kuper, L., Olezeski, C., … & Alstott, A. L. (2022, July 8). A critical review of the June 2022 Florida Medicaid report on the medical treatment of gender dysphoria. https://web.archive.org/web/20241205041150/https://files-profile.medicine.yale.edu/documents/c11e1419-a122-4b2f-87a8-cc4c9fbf57a4

McNamara, M., Baker, K., Connelly, K., Janssen, A., Olson-Kennedy, J., Pang, K. C., Scheim, A., Turban, J., & Alstott, A. (2024). *An Evidence-Based Critique of "The Cass Review" on Gender-affirming Care for Adolescent Gender Dysphoria*. https://web.archive.org/web/20250501184400/https://law.yale.edu/sites/default/files/documents/integrity-project_cass-response.pdf.

McNamara, M., Kronish, A., Birkhead, D., & Mehringer, J. (2025). Preserving pediatric best practice in challenging times. *JAMA Pediatrics*, *179*(9), 939–940.

McNamara, M., Lepore, C., & Alstott, A. (2022). Protecting transgender health and challenging science denialism in policy. *New England Journal of Medicine*, *387*(21), 1919–1921.

McNamara, M., Lepore, C., Alstott, A., Kamody, R., Kuper, L., Szilagyi, N., ... & Olezeski, C. (2022). Scientific misinformation and gender affirming care: tools for providers on the front lines. *Journal of Adolescent Health*, *71*(3), 251–253.

Meade, N. G., Lepore, C., Olezeski, C. L., & McNamara, M. (2024). Understanding and addressing disinformation in gender-affirming health care bans. *Transgender Health*, *9*(4), 281–287.

Miroshnychenko, A., Ibrahim, S., Roldan, Y., Kulatunga-Moruzi, C., Montante, S., Couban, R., ... & Brignardello-Petersen, R. (2025). Gender affirming hormone therapy for individuals with gender dysphoria aged< 26 years: a systematic review and meta-analysis. *Archives of Disease in Childhood*, *110*(6), 437–445.

Miroshnychenko, A., Roldan, Y., Ibrahim, S., Kulatunga-Moruzi, C., Montante, S., Couban, R., Guyatt, G., & Brignardello-Petersen, R. (2025). Puberty blockers for gender dysphoria in youth: A systematic review and meta-analysis. *Archives of Disease in Childhood*, *110*(6), 429–436.

Miroshnychenko, A., Roldan, Y. M., Ibrahim, S., Kulatunga-Moruzi, C., Dahlin, K., Montante, S., Couban, R., Guyatt, G., & Brignardello-Petersen, R. (2025). Mastectomy for individuals with gender dysphoria below 26 years of age: A systematic review and meta-analysis. *Plastic and Reconstructive Surgery*, *155*(6), 915 –923.[1]

National Institute for Health and Care Excellence. (2020a). *Evidence Review: Gender-Affirming Hormones for Children and Adolescents with Gender Dysphoria*. National Institute for Health and Care Excellence. https://webarchive.nationalarchives.gov.uk/ukgwa/20250310143633/https://cass.i

[1] This is cited in the Review as Miroshnychenko (2024).

ndependent-review.uk/wp-content/uploads/2022/09/20220726_Evidence-review_Gender-affirming-hormones_For-upload_Final.pdf

National Institute for Health and Care Excellence. (2020b). *Evidence Review: Gonadotrophin Releasing Hormone Analogues for Children and Adolescents with Gender Dysphoria*. National Institute for Health and Care Excellence. https://webarchive.nationalarchives.gov.uk/ukgwa/20250310143633/https://cass.independent-review.uk/wp-content/uploads/2022/09/20220726_Evidence-review_GnRH-analogues_For-upload_Final.pdf

National Institutes of Health. (2024). The impact of early medical treatment in transgender youth: Description. *NIH RePorter*. https://web.archive.org/web/20251001160006/https://reporter.nih.gov/search/nb3hKc_s_0ildNVfWVYsmw/project-details/10854729

New Jersey. (n.d). NJ Administrative Code 10:74-1.4. https://bit.ly/4nD3pD1

Noone, C., Southgate, A., Ashman, A., Quinn, E., Comer, D., Shrewsbury, D., Ashley, F., Hartland, J., Paschedag, J., Gilmore, J., Kennedy, N., Woolley, T. E., Heath, R., Goulding, R., Simpson, V., Kiely, E., Coll, S., White, M., Grijseels, D. M., … McLamore, Q. (2025). Critically appraising the Cass Report: Methodological flaws and unsupported claims. *BMC Medical Research Methodology*, *25*(1), 128.

Nunes-Moreno, M., Furniss, A., Cortez, S., Davis, S. M., Dowshen, N., Kazak, A. E., ... & Nokoff, N. J. (2025). Mental health diagnoses and suicidality among transgender youth in hospital settings. *LGBT Health*, *12*(1), 20–28.

Olson-Kennedy, J., Chan Y., Garofalo, R., Spack, N., Chen, D., Clark, L., Ehrensaft, D., Hidalgo, M., Tishelman, A., Rosenthal, S. 2019. Impact of early medical treatment for transgender youth: Protocol for the longitudinal, observational trans youth care study. *JMIR Research Protocols*, *8*(7), e14434.

Olson-Kennedy, J., Durazo-Arvizu, R., Wang, L., Wong, C. F., Chen, D., Ehrensaft, D., ... & Rosenthal, S. M. (2025). Mental and emotional health of youth after 24

months of gender-affirming medical care initiated with pubertal suppression. https://doi.org/10.1101/2025.05.14.25327614

Olson-Kennedy, J., Wang, L., Wong, C. F., Chen, D., Ehrensaft, D., Hidalgo, M. A., Tishelman, A. C., Chan, Y.-M., Garofalo, R., Radix, A. E., & Rosenthal, S. M. (2025). Emotional health of transgender youth 24 months after initiating gender-affirming hormone therapy. *Journal of Adolescent Health, 77*(1), 41–50.

Olson, K. R., Raber, G. F., & Gallagher, N. M. (2024). Levels of satisfaction and regret with gender-affirming medical care in adolescence. *JAMA Pediatrics*, *178*(12), 1354–1361.

Pollock, M., Fernandes, R. M., Becker, L. A., Pieper, D., & Hartling, L. (2024). Chapter V: Overviews of reviews [last updated August 2023]. In *Cochrane Handbook for Systematic Reviews of Interventions*, Version 6.5. Cochrane. https://web.archive.org/web/20250501195623/www.training.cochrane.org/handbook

Pollock, M., Fernandes, R. M., Pieper, D., Tricco, A. C., Gates, M., Gates, A., & Hartling, L. (2019). Preferred Reporting Items for Overviews of Reviews (PRIOR): a protocol for development of a reporting guideline for overviews of reviews of healthcare interventions. *Systematic Reviews*, *8*(1), 335.

Regenstreif, L. (2023). Olson J. R01 HD082554-01A1, 2014: The impact of early medical treatment in transgender youth. http://bit.ly/48fuzLT

Restar, A. J. (2023). Gender-affirming care is preventative care. *The Lancet Regional Health–Americas*, *24*.

Rider, G. N., Weideman, B. C., Ehrensaft, D., Choudhary, K., Connor, J. J., Feldman, J., ... & Berg, D. (2025). Scientific integrity and pediatric gender healthcare: Disputing the HHS Review. *Sexuality Research and Social Policy*, 1–6. https//:10.1007/s13178-025-01221-5

Riedel, S. (2025, May 1). HHS report on trans youth relies on junk science and endorses conversion therapy. *Them*.

https://web.archive.org/web/20250512135459/https://www.them.us/story/hhs-report-trans-youth-gender-affirming-care-brian-christine-conversion-therapy

Ryan, B. (2024, September 6). Top plastic surgeon won't commit to opposing bans of gender-transition treatment for minors, plus he decries WPATH for lack of transparency. *Hazard Ratio*. https://web.archive.org/web/20250124012342/https://benryan.substack.com/p/top-plastic-surgeon-wont-commit-to

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *BMJ*, *312*(7023), 71–72.

Sinai, J., Kulatunga-Moruzi, C., & Jorgensen, S. (2025). Levels of satisfaction and regret are far from settled. *JAMA Pediatrics*, *179*(5), 579–580.

Singal, J. (2018, July/August). When children say they're trans. *The Atlantic*. https://web.archive.org/web/20250501184530/https://www.theatlantic.com/magazine/archive/2018/07/when-a-child-says-shes-trans/561749/

Society for Evidence-Based Gender Medicine. (2025, June). Notable publications in gender medicine. *SEGM Digest*, *2*. https://web.archive.org/web/20250829125350/https://segm.org/SEGM-Digest-Issue2-2025

Spencer, B. (2025, April 2). NHS swaps gender drugs for 'holistic' care. *The Sunday Times*. https://web.archive.org/web/20250420045726/https://www.thetimes.com/uk/healthcare/article/nhs-swaps-gender-drugs-for-holistic-care-jxhm3b6vk

Substance Abuse and Mental Health Services Administration. (2023). *Moving Beyond Change Efforts: Evidence and Action to Support and Affirm LGBTQI+ Youth.* Rockville, MD: Center for Substance Abuse Prevention.

Sullivan, A. (2025). *Review of Data, Statistics and Research on Sex and Gender*. Social Research Institute, UCL.

Taylor, J., Hall, R., Heathcote, C., Hewitt, C. E., Langton, T., & Fraser, L. (2024a). Clinical guidelines for children and adolescents experiencing gender dysphoria or incongruence: A systematic review of guideline quality (part 1). *Archives of Disease in Childhood*, *109*(Suppl 2), s65–s72.

Taylor, J., Hall, R., Heathcote, C., Hewitt, C. E., Langton, T., & Fraser, L. (2024b). Clinical guidelines for children and adolescents experiencing gender dysphoria or incongruence: A systematic review of recommendations (part 2). *Archives of Disease in Childhood*, *109*(Suppl 2), s73–s82.

Taylor, J., Mitchell, A., Hall, R., Heathcote, C., Langton, T., Fraser, L., & Hewitt, C. E. (2024). Interventions to suppress puberty in adolescents experiencing gender dysphoria or incongruence: A systematic review. *Archives of Disease in Childhood*, *109*(Suppl 2), s33–s47.

Taylor, J., Mitchell, A., Hall, R., Langton, T., Fraser, L., & Hewitt, C. E. (2024). Masculinising and feminising hormone interventions for adolescents experiencing gender dysphoria or incongruence: A systematic review. *Archives of Disease in Childhood*, *109*(Suppl 2), s48–s56.

Tennessee. (2024). TN Code §71-5-144. https://web.archive.org/web/20251001161948/https://law.justia.com/codes/tennessee/title-71/chapter-5/part-1/section-71-5-144/

The White House. (2025, January 28). Protecting children from chemical and surgical mutilation. https://web.archive.org/web/20250128222356/https://www.whitehouse.gov/presidential-actions/2025/01/protecting-children-from-chemical-and-surgical-mutilation/

Tordoff, D. M., Wanta, J. W., Collin, A., Stepney, C., Inwards-Breland, D. J., & Ahrens, K. (2022). Mental health outcomes in transgender and nonbinary youths receiving gender-affirming care. *JAMA Network Open*, *5*(2), e220978.

Twenge, J. M., Wells, B. E., Le, J., & Rider, G. N. (2025). Increases in self-identifying as transgender among US adults, 2014–2022. *Sexuality Research and Social Policy*, *22*(2), 755–773.

U.S. v. Skrmetti (No. 23-477), *Expert Researchers and Physicians*, Amicus curiae brief (Supreme Court of the United States, September 3, 2024). https://web.archive.org/web/20250408155518/https://www.supremecourt.gov/DocketPDF/23/23-477/323851/20240904161709482_23-477%20Amicus%20Brief.pdf

United States Joint Statement. (2023, August 23). *United States joint statement against conversion efforts*. https://web.archive.org/web/20250812091103/https://usjs.org/usjs-final-version/

University of Minnesota (2025). World-class leadership in sexual and gender health. https://web.archive.org/web/20250912092717/https://med.umn.edu/sexualhealth

University of Utah College of Pharmacy, Drug Regimen Review Center. (2025). *Gender-Affirming Medical Treatments for Pediatric Patients with Gender Dysphoria*. Salt Lake City, UT: University of Utah. https://web.archive.org/web/20250601064306/https://le.utah.gov/AgencyRP/reportingDetail.jsp?rid=636

van der Loos, M. A., Klink, D. T., Hannema, S. E., Bruinsma, S., Steensma, T. D., Kreukels, B. P., ... & Wiepjes, C. M. (2023). Children and adolescents in the Amsterdam cohort of gender dysphoria: Trends in diagnostic-and treatment trajectories during the first 20 years of the Dutch protocol. *Journal of Sexual Medicine*, *20*(3), 398–409.

van der Loos, M. A., Vlot, M. C., Klink, D. T., Hannema, S. E., den Heijer, M., & Wiepjes, C. M. (2023). Bone mineral density in transgender adolescents treated with puberty suppression and subsequent gender-affirming hormones. *JAMA Pediatrics*, *177*(12), 1332–1341.

Whiting, P., Savović, J., Higgins, J. P., Caldwell, D. M., Reeves, B. C., Shea, B., Davies, P., Kleijnen, J., Churchill, R., & ROBIS group (2016). ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *Journal of Clinical Epidemiology*, *69*, 225–234.

Woozle effect. (2025, October 31). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Woozle_effect&oldid=1315094638

World Professional Association for Transgender Health & United States Professional Association for Transgender Health. (2025, May 2). *WPATH and USPATH Response to the HHS Report on Gender Dysphoria.* https://web.archive.org/web/20250506104135/https://wpath.org/wp-content/uploads/2025/05/WPATH-USPATH-Response-to-HHS-Report-02May2025-1.pdf

Yuhas, A. (2021, March 31). It's time to revisit the satanic panic. *New York Times*. https://web.archive.org/web/20210331110917/https://www.nytimes.com/2021/03/31/us/satanic-panic.html