Technical paper

Gender-questioning teenagers: puberty blockers and hormone treatment vs placebo

Matilda Gosling

December 2022

Sex Matters is a human rights organisation campaigning for clarity about sex in law, policy and language

sex-matters.org | info@sex-matters.org

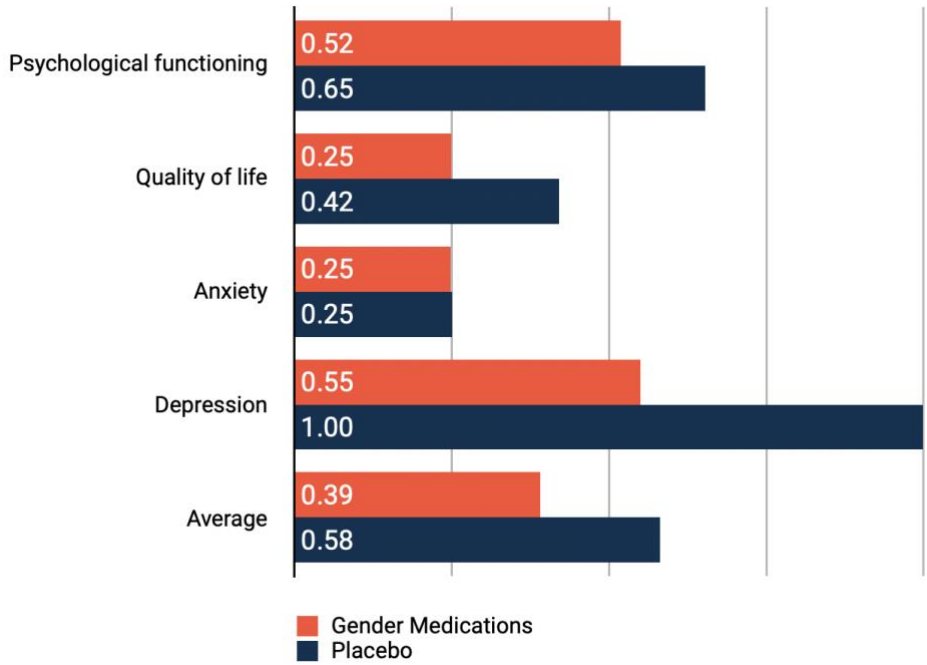
sexmatters

Contents

Abstract

The gender-affirmative medical model uses evidence from a small number of low-quality studies to show puberty blockers and hormone treatment have a positive effect on teenagers' mental health. Separate evidence from medical trials suggests the placebo effect on mental health outcomes is large and significant. This paper adds to the literature on the treatment of gender dysphoria by using data from comparable studies on teenagers' mental health to assess whether puberty blockers and hormone treatment are better than placebo at alleviating distress. It looks at trials of medication specific to mental health – for example, drugs aimed at treating depression, irritability, schizophrenia and generalised anxiety disorder – to isolate the placebo effect. The data shows that the average improvement in mental health over the course of treatment is no bigger for gender medications than it is for placebo. The headline data, in fact, suggests that placebo does more for teenagers' mental health than gender medication does, although we do not have the aggregate data on variance that would be needed to state this categorically.

Figure 1. Effect size of gender medication vs placebo on teenagers' mental health



Gender medication comprises puberty blockers and/or hormone treatment for gender-questioning teenagers. Uses all comparable, longitudinal data since 2010. Scales are reversed where appropriate to show positive change.

Introduction

The current evidence base on medication for gender-questioning teenagers is used to justify a model of "gender-affirming care", through which children are given puberty blockers, hormone treatment and, later, surgery. While the wisdom of this model is starting to be examined in England, it remains prevalent in several other countries, including Scotland, Australia and the United States. Research papers make seemingly evidence-based assertions like this one: "In short term studies gender affirming hormone treatment with both estradiol and testosterone has been found to be safe and improve mental health and quality of life outcomes." [1] But what if the observed, contested and relatively small differences in mental health outcomes over the course of treatment are nothing more, or even less, than a placebo effect?

In medicine, a placebo effect happens when people take an inert substance that has a positive effect on outcomes based on their expectations or beliefs. It can also happen in psychotherapy when a non-specific treatment is given[2], although that is outside of this paper's scope.

The placebo effect can be strong. It is not well understood or discussed in the case of gender medication[3], where clinical practice is underpinned by a weak evidence base. This area of medicine has a combination of features that make it particularly prone to placebo effects. These include a celebratory media and social media promotion of the treatment model by prestigious clinicians, medical organisations and celebrities to a rapidly growing number of vulnerable young people, many with neurodevelopmental and psychiatric disorders.[4]

> "In a randomised controlled trial (RCT), patients are allocated into two or more groups. One group of patients is given an active treatment and the other a placebo (in a well-designed study, neither group knows which they have been given, and neither do their clinicians; this is called a 'double-blind' study). Measurements at the start and end of the study test whether outcomes have improved more for the treatment group than they have for the placebo group. If they have, and these results are statistically significant, the treatment is determined a success."

There are no experimental studies looking at whether medical treatment for gender dysphoria – puberty blockers and hormone treatment – work better than a placebo in helping patients to ameliorate their feelings of distress. Instead, we can turn to RCT data looking at outcomes in adolescents for other medical treatment relating to adolescent mental health conditions. Many meta-analyses of mental health treatments look at placebo response, or a binary representation

[1] Salas-Humara et al (2019). (For full details, see **References** section.)

[2] American Psychological Association. Placebo effect. *APA Dictionary of Psychology.* (accessed 24 October 2022).

[3] Clayton, A. (2022a).

[4] Clayton, A. (2022b).

of whether individual cases responded to treatment.[5] The gender dysphoria studies instead tend to report mean changes[6], meaning that meta-analytical findings cannot directly be compared with them. The purpose of this research was therefore to compare the placebo effects of drug treatment on mental health outcomes with the effects on mental health of puberty blockers and hormone treatment for gender dysphoria.[7]

Trials on a specific area of mental health – for example, depression – are likely to have a larger placebo effect than those not specifically designed to address a particular mental health area. If I feel depressed, for example, and I take a pill that I believe will help with this, my depression levels will probably fall more as a result of placebo than if I am taking a pill I believe will treat anxiety. In order to get around this limitation, which might otherwise make the placebo effect in these studies bigger than if a comparable trial were run with gender medication, I have included all secondary outcomes data where the relevant scales are used – for example, measures of anxiety included in a study on depression. I have also included all relevant studies looking at trials in areas of mental health not relevant to the indicators in question, such as those looking at irritability and schizophrenia.

Study populations are not, and can never be, perfectly comparable. The mental health profile of teenagers with gender dysphoria is similar in profile, however, to those who are referred to health services for other mental health conditions.[8] These populations can therefore reasonably be compared, in the absence of better data. There is a potential for the placebo effect to confound the effects of taking a pill with the effects of having a relationship with a clinician, but this is not relevant to discussion here[9] – the point is to test whether gender medication has a positive effect on teenagers' mental health outcomes over and above types of treatment that have no or little potential to cause harm (whether that's a sugar pill or a positive relationship with a clinician).

Comparing mental health treatment effects of gender medication in teenagers with placebo effects in comparable populations is important. This is because of the risks of long-term physical harm from the gender-affirming treatment approach, including effects on bone density levels, infertility and loss of sexual function.[10] Evidence showing positive mental health impacts

[5] For example, Locher et al (2017).

[6] With the exception of Tordoff et al (2022), who report a binary change based on scale variables recoded into categorical variables.

[7] Surgical outcomes were excluded for two reasons. The first is that surgery on gender-dysphoric patients tends to happen (not always, but most of the time) after the age of 18, and the focus here is minor adolescents. The second is that, in the interests of making findings as comparable as possible, two types of drug treatment are more comparable than drug treatment with surgery.

[8] Zucker, K. J. (2019).

[9] Hróbjartsson et al (2011).

[10] For example Vlot et al (2017); Krishna et al (2019).

ought reasonably to be greater than placebo at worst, and overwhelming at best, to justify these physical impacts. Longitudinal studies of gender-questioning adolescents are used as the main evidence base to support gender-affirming medicine; despite many flaws in the design of individual studies,[11] they are the best-quality studies on medical treatment for gender-questioning teenagers, and the ones that are most comparable to RCTs. These studies are therefore the ones selected for a detailed analysis in this technical paper.

The sensitivities of the subject area and the heat of the debate surrounding it mean that this analysis is likely to be subject to an unusual degree of scrutiny. In order to allow results to be interrogated by those who wish to do so, full data tables and search terms have been provided.

Approach and results

Approach

General selection criteria

Studies were included if they were published from 2010 onwards and were specific to adolescents. Searches were run using the Google Scholar platform, with additional data collection run through JSTOR, Ebscohost and CORE. Studies were included that were longitudinal in nature, taking prospective data before and after, or at the end of, treatment; and that took quantitative measurements, not descriptive ones. The focus was mental health outcomes, not behavioural ones.

Gender medication study selection

Studies were identified since 2010 that took a quantitative, longitudinal design and measured the mental health effects of puberty blockers and hormone treatment given to gender-questioning teenagers. One study looking at surgical outcomes was excluded[12] because it did not report findings separately for the medication element, and because it included data already covered in another study.[13] The eight studies that met the inclusion/exclusion criteria are outlined in the **References** section.[14] Although comparison groups were outlined in three of

[11] See the full research paper that this technical paper informs, to be published separately, for a fuller discussion.

[12] De Vries et al (2014).

[13] De Vries et al (2011).

[14] Measures of body image and/or gender dysphoria were excluded as they are not assessed as outcomes in comparison studies. Data on suicidal ideation and self harm was categorical, so could not be included in a composite mean score relating to mental health. A more detailed assessment of both these areas is offered in the accompanying research paper.

these studies, they were not given a placebo, so cannot be used to test the question explored in this paper.

Comparator RCT study selection

Studies were included that looked at treatment for mental health or neurodevelopmental conditions, and that measured mental health outcomes using one or more of the scales used within the gender medication studies (Beck Depression Inventory – BDI, Child Behavior Checklist – CBCL, Centre for Epidemiologic Studies Depression Scale – CESD, Children's Global Assessment Scale – CGAS, General Well-Being Schedule – GWBS, Patient Health Questionnaire 9 – PHQ-9, Quality of Life Enjoyment and Satisfaction Questionnaire – Short Form – QLES-Q-SF, Quick Inventory of Depressive Symptomatology – QIDS, Screen for Child Anxiety Related Disorders – SCARED, Strengths and Difficulties Questionnaire – SDQ, State-Trait Anxiety Inventory – STAI, Youth Self Report – YSR). Studies were included when they used medication (excluding food supplements) and had a placebo group. They were excluded when they included a comorbidity with physical illness. They were also excluded when they used a preventative, as opposed to treatment, approach, or when they employed an open-label approach (one in which both clinicians and patients know which group they are in). Studies looking at teenage addictions were excluded; although addictions might reasonably be seen as a mental health condition, there is a strong physical component that warrants exclusion from this particular research.

The following search terms were employed to do the initial sift of studies prior to a final decision (based on the criteria above) about inclusion or exclusion:

"Beck Depression Inventory" OR BDI OR "Child Behavior Checklist" OR CBCL OR "Center for Epidemiologic Studies Depression Scale" OR CESD OR CES-D OR "Children's Global Assessment Scale" OR CGAS OR "General Well-Being Schedule" OR GWBS OR "Patient Health Questionnaire" OR PHQ-9 OR "Quality of Life Enjoyment and Satisfaction" OR P-Q-LES OR PQLES OR QLES OR Q-LES OR "Quick Inventory of Depressive Symptomatology" OR QIDS OR "Screen for Child Anxiety Related Disorders" OR SCARED OR "Strengths and Difficulties Questionnaire" OR SDQ OR "State-Trait Anxiety Inventory" OR STAI OR YSR intitle:adolescents OR intitle:youth OR intitle:"young people" OR intitle:adolescence intitle:RCT OR intitle:trial OR intitle:randomized OR intitle:randomised OR intitle:blind OR intitle:placebo placebo -intitle:protocol (2010 onwards)

Some of the strings were run separately due to search engine character count limits. All studies were included if they met the criteria set out above (academic posters were excluded as they do not report data in sufficient detail). Fifteen studies met the final inclusion criteria, with 22 separate measurements of relevant mental-health outcomes.

Data selection

Data from the gender medication studies was excluded in the main analysis when it used a scale that was unavailable in comparator studies,[15] although it was included in the alternative analysis. One scale was not named and therefore had to be excluded,[16] and data on suicidal ideation and suicides in a separate study was excluded as it failed to use comparable measurement periods.[17] Where different figures were offered in different parts of study papers on sample sizes, the ones offered in data tables on outcomes were given priority. Where different measurement timeframes were offered, the longest one available was selected. If baseline data was divided into all participants who started the study and those who completed it, the latter was selected.

Data was gathered for all relevant measures across all studies on: baseline and endline sample sizes; baseline and endline mean scores; mean change; and standard deviations for the baseline mean, endline mean and mean change. In many cases, not all this data was available, and so a range of additional data was collected to allow the key measures to be calculated: standard errors; confidence intervals; and T statistics/p values relating to mean changes. A note was kept separately of the scale maximum for each measure. This sometimes differed by study; for example, some reported raw numbers and some a percentage.

Tordoff et al (2022) was excluded from the main analysis, as data was not included in a form that could be compared with data from the other studies; the authors had recoded scale variables into categorical variables and did not report scale means. No raw data was available. It would be worth including in any future meta-analysis if original data tables can be sourced from the study authors.

Analysis

Main analysis

Cohen's d, which measures the difference between two means in a standardised way, was used to calculate effect sizes for the gender medication and placebo groups. While Cohen's d is more commonly used to show the effect size between groups (comparing endline means of treatment groups vs placebo groups), it was used here to test the effect size of gender medication vs placebo by assessing baseline to endline mean differences, rather than raw endline mean scores. These groups had different baseline scores, so the within-group mean difference is the object of interest, not the raw endline mean score. If effect sizes for a given

[15] The following scales were excluded on this basis: CBCL/YSR, GWBS, PHQ-9, SDQ and STAI.

[16] Suicidal ideation – Achille et al (2020).

[17] Kuper et al (2020).

mental health measure are similar for gender medications and placebo, it is likely that gender medication is no more effective than placebo; similarly, a larger effect size for one treatment approach is likely to show that it is more effective than the other one (this last point cannot be tested, though, in the absence of aggregate data on variance).

In many cases, especially for the gender medication studies, SDs were not available for the mean change, and alternative data was used to calculate them. This alternative data included standard errors (SE), confidence intervals relating to the mean change and p values of statistical tests relating to the mean change. Cochrane's guidance[18] was used to calculate mean change SDs where data was missing, with small adjustments where necessary to take account of the focus on within-group change, not between-group raw scores. Correlation coefficients were calculated and used to impute mean change SD in four cases within the gender medication studies. Five measures with missing data from comparator studies had to be excluded. Details of adjustments are available in the appendices, as are the key formulae used.

Data was separated into areas relating to the five comparable scales: psychological functioning (CGAS), quality of life (QLES), anxiety (SCARED) and depression (BDI/QIDS). The BDI and QIDS data were merged into a single indicator for depression to reduce potential bias due to small samples.

The baseline and endline mean averages for each area of well-being were calculated by weighting each mean by the sample size, and the pooled standard deviation of the mean change was calculated as follows:

$$SD_{pooled} = \frac{\sqrt{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2 + \ldots + (n_k - 1)SD_k^2}}{n_1 + n_2 + \ldots + n_k - k}$$

Finally, Cohen's d was used to calculate the within-group effect size for each of the four areas:

$$d = \frac{M_{pooled}^{EL} - M_{pooled}^{BL}}{SD_{pooled}}$$

An alternative analysis was conducted to verify if the findings held under different conditions, in which all gender medication measures were included (not just those that used scales available in comparator studies). Where more than one measure of mental health outcome was available within a single study, data was weighted accordingly to ensure teenagers who responded to

[18] Higgins et al (2022).

more than one mental health survey were not over-represented in the final analysis. Cohen's d was conducted on the weighted samples at an aggregate level within groups.

Results

Main analysis

The tables and diagrams underneath demonstrate that gender medication is no better than placebo for teenagers' mental health. The effect size of gender medication is the same as placebo on teenagers' anxiety, and it is less than that of placebo on psychological functioning, quality of life and depression.

Table 1. Effect sizes, main analysis

Area of mental health	Gender medications	Placebo
Psychological functioning	0.52	0.65
Quality of life	0.25	0.42
Anxiety	0.25	0.25
Depression	0.55	1.00
Average	**0.39**	**0.58**

Table 2. Psychological functioning

Measure	Gender medications	Placebo
Pooled standard deviation	10.8	12.6
Average baseline mean	64.7	46.2
Average endline mean	70.2	54.4
Effect size	**0.52**	**0.65**

Figure 2. Forest plot: psychological functioning

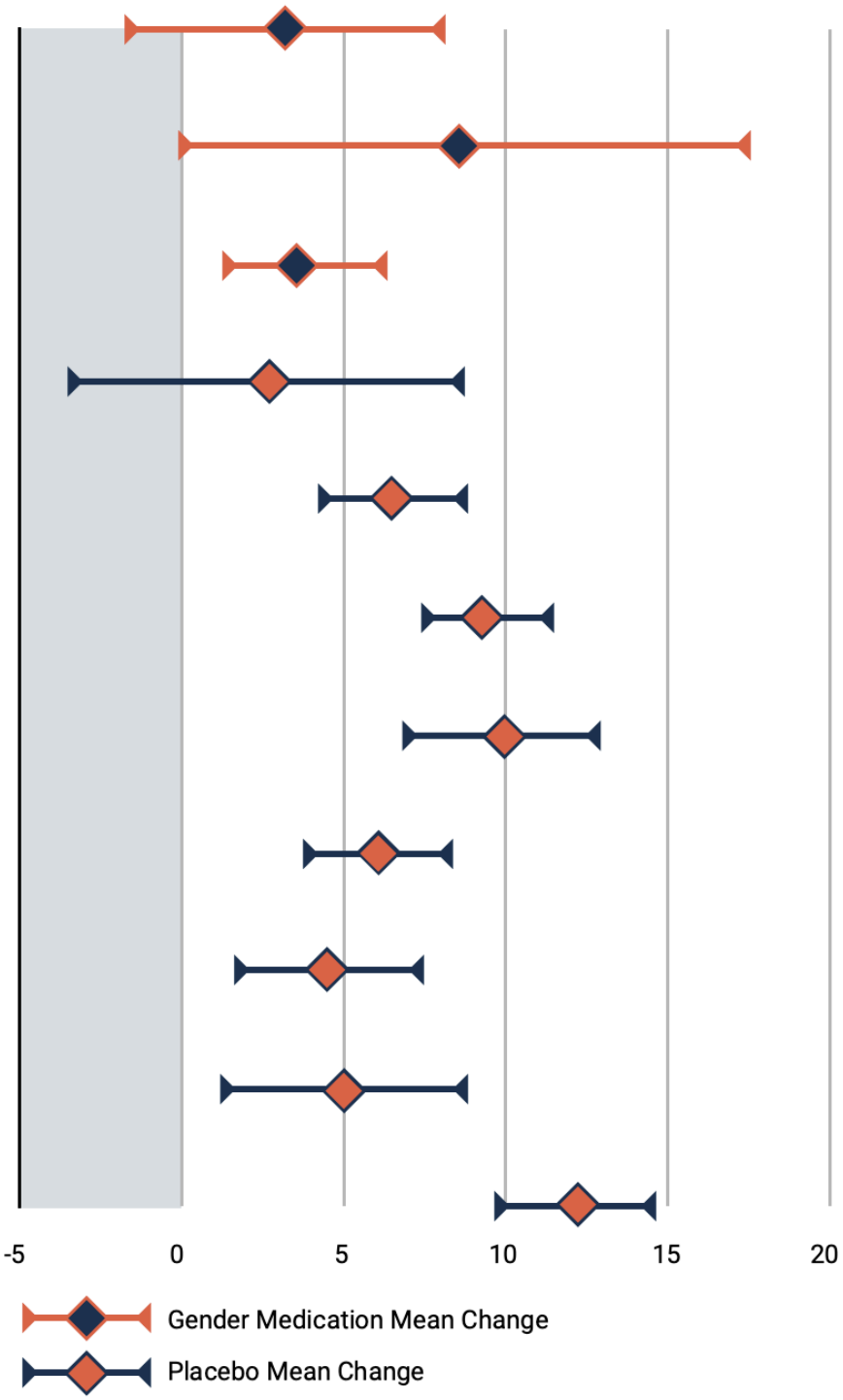


Table 3. Quality of life

Measure	Gender medications	Placebo
Pooled standard deviation	42.2	13.6
Average baseline mean	61.5	48.9
Average endline mean	72.0	54.7
Effect size	**0.25**	**0.42**

Figure 3. Forest plot: quality of life

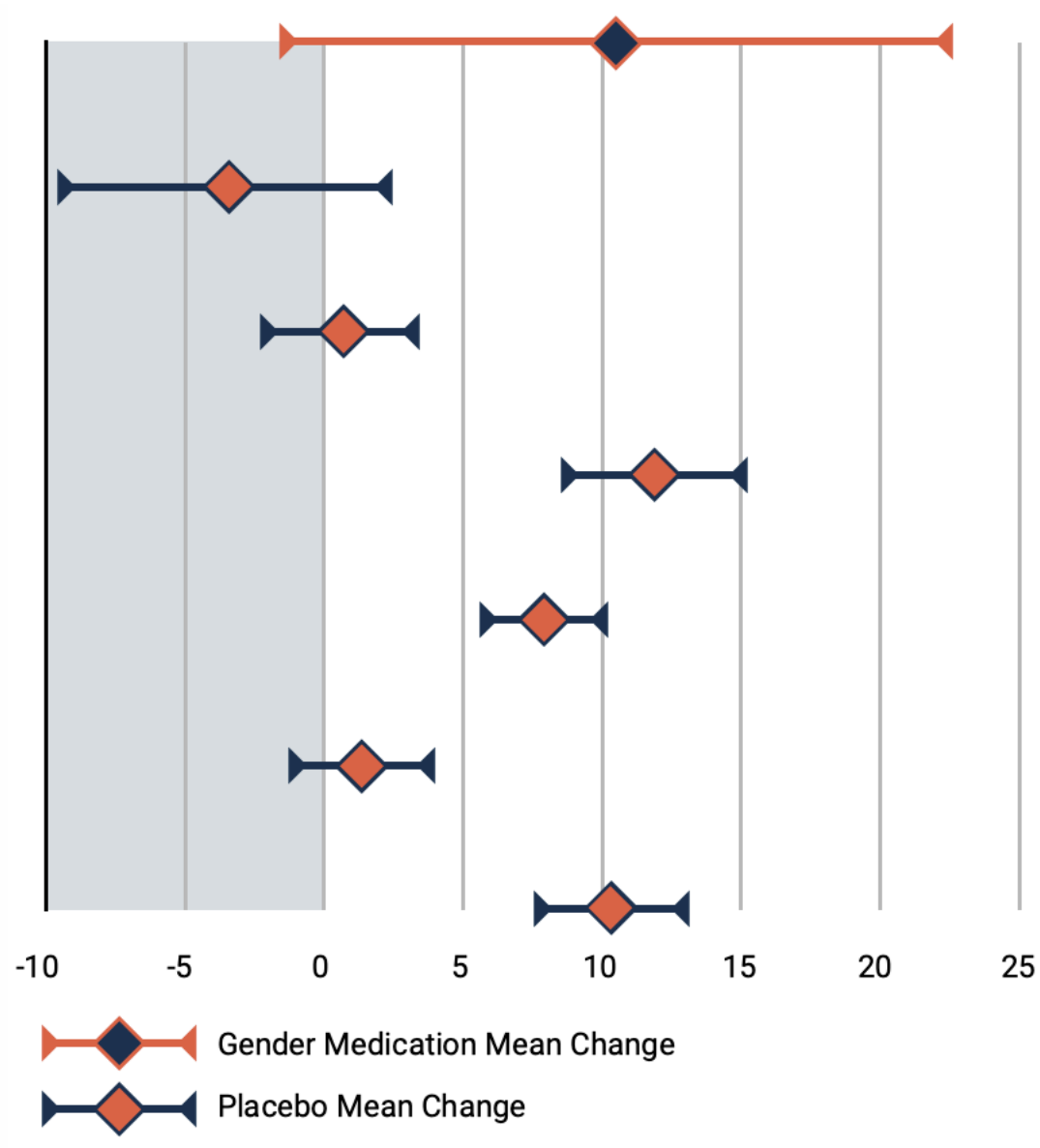


Table 4. Anxiety

Measure	Gender medications	Placebo
Pooled standard deviation	15.3	15.5
Average baseline mean	32.4	40.0
Average endline mean	28.6	36.1
Effect size	**0.25**	**0.25**

Figure 4. Forest plot: anxiety

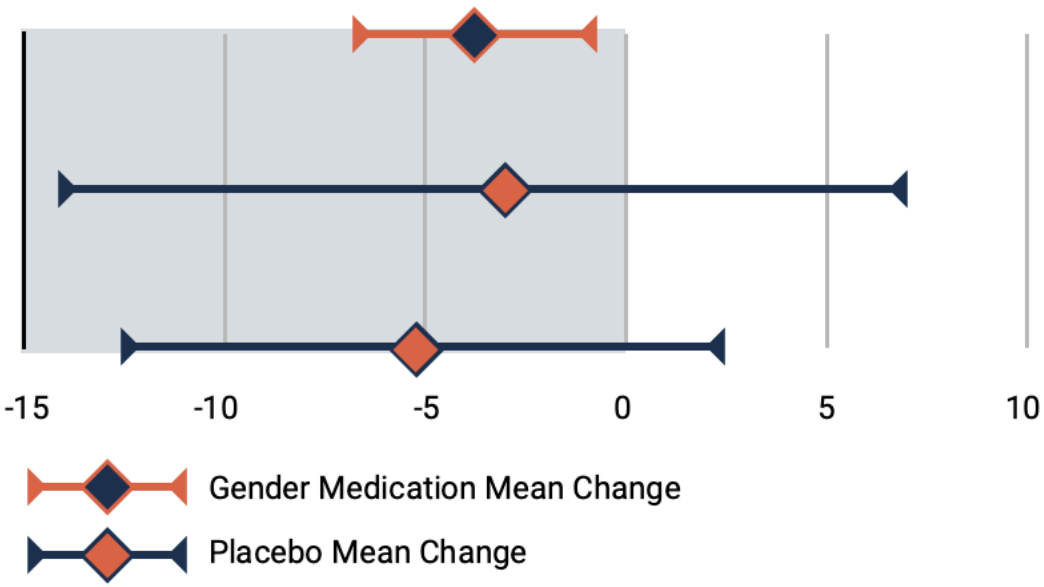
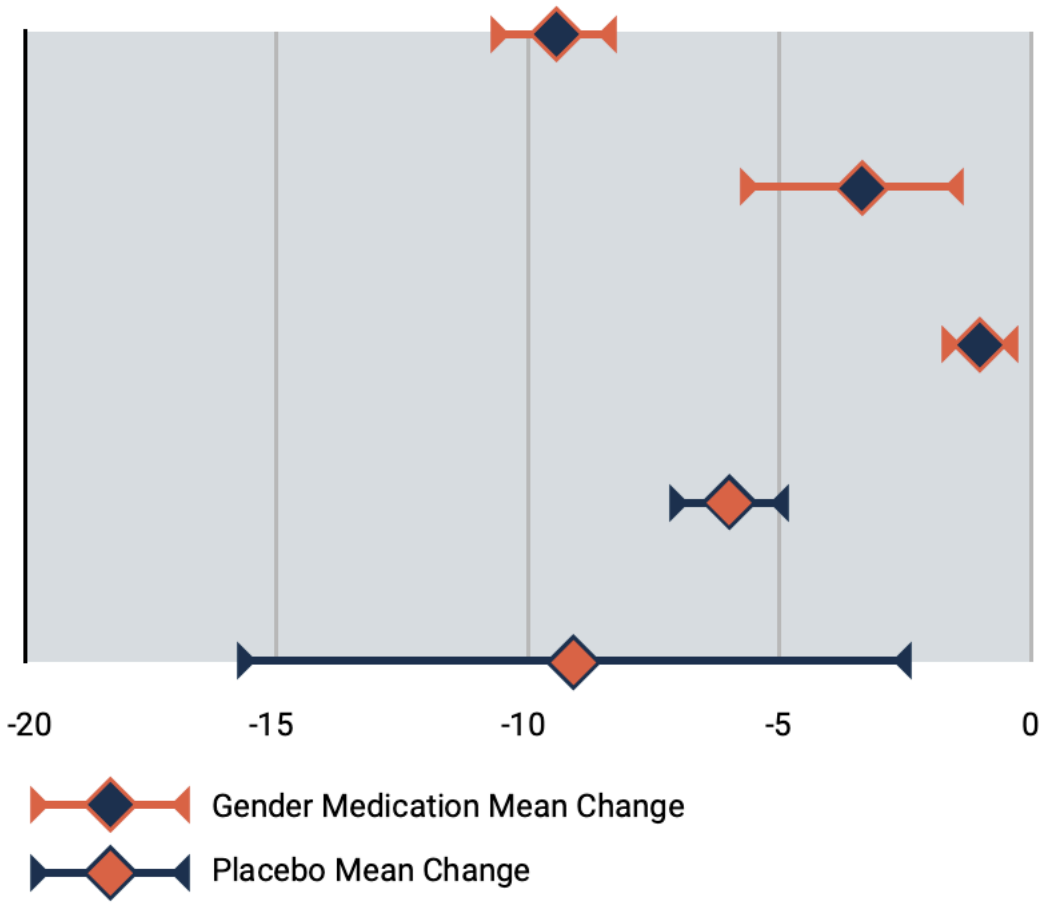


Table 5. Depression

Measure	Gender medications	Placebo
Pooled standard deviation	4.7	6.5
Average baseline mean	9.2	19.2
Average endline mean	6.6	12.7
Effect size	**0.55**	**1.00**

Figure 5. Forest plot: depression



Alternative analysis

This table shows the effect size when all gender measures are included, including those on non-comparable scales. Data has been weighted within each group (gender medication vs placebo) by sample size, and scales have been reversed where needed for comparability. As with the scale-matched data in the main analysis, the overall effect size of gender medications is less than that of placebo.

Table 6. Alternative analysis

Measure	Gender medications	Placebo
Pooled standard deviation	14.5	12.8
Average baseline mean	42.0	45.5
Average endline mean	47.2	53.0
Effect size	**0.36**	**0.59**

Limitations

- The comparator studies include RCTs that are not double blinded. When clinicians and/or patients know that an active drug has been prescribed, the patients are more responsive to treatment. The placebo effect is therefore bigger in studies that are truly blind. There was not enough data to limit the selection of comparison studies here to only those that were truly double blinded, but the placebo effect in comparison studies might be even larger if there were enough studies to limit the inclusion criteria in this way.[19]
- Selection of search terms may have excluded some relevant studies; others may have been excluded due to the search platform selected. The search was comprehensive, but not exhaustive, as it only used one platform. Google Scholar was selected due to its superior performance compared to many alternatives in finding citations (88% of all those available).[20] It is recommended that an academic team conducts a full meta-analysis in this area, using a range of alternative platforms for full coverage.
- The effects of psychotherapy were not always accounted for. They were confounded in four of the seven main gender medication studies, and there was an element of confounding in some of the comparator RCTs too. However, as this was a limitation across both groups, it is unlikely to have a strong effect on the comparisons between them.
- In some cases, mean scores were converted into mean T scores,[21] not displayed as raw data. This will have affected comparability.
- There was insufficient data to calculate the standard deviation relating to the mean change in all cases, so correlation coefficients based on studies with more data had to be used for some calculations. The correlation coefficient average was 0.63; if it is higher in reality, the effect size of gender medication will be higher than stated here, and if it is lower, the effect size will be lower.
- The usual caveats relating to correlation and causality apply. We cannot say that changes in mental health are a direct result of a particular intervention, although the experimental design of RCTs means that we can have more confidence in the causal relationship than we can in the non-experimental longitudinal studies that are used to interrogate gender medication.
- Teenage-onset gender dysphoria is a relatively recent phenomenon. Even using a cut-off of 2010 for study selection, many recent studies include only children whose dysphoria started earlier in childhood. This data may not therefore be directly relevant to those whose dysphoria began later.

[19] Kirsch, I. (2019).

[20] Martín-Martín et al (2021).

[21] For example, de Vries et al (2011).

- Scale maximums and minimums were often not given within the original research papers, and had to be sourced from elsewhere. It is possible that researchers may have deviated from the sources used in their scoring approach. This limitation affects scales that were reversed for the alternative analysis.
- Populations may not all be independent; for example, it is likely that there is some overlap between the data from Costa et al (2015) and Carmichael et al (2021), but it was not made clear, so a decision could not be made on whether to include or exclude it.
- In some cases, a revised version of a particular scale was used in different studies. A decision was made to include all instances of the same scale in order to keep the pool of included studies at a reasonable size, but certain questions may have been slightly different.
- The quality of the evidence in the gender medication studies is poor. There are small sample sizes, high levels of attrition, meaningless comparison groups and a degree of obfuscation in reporting of results. Any related analysis, including this one, is limited by the shortcomings of the primary studies. More details of these issues are set out in the accompanying paper. Some of the comparator studies also suffer from small sample sizes.

Potential critiques

Some of the mental health outcomes under review aren't intended as primary outcomes of gender-affirming medicine.
In order to deal with this potential limitation, comparison studies – in addition to including primary mental health outcome measurement – include all those for which mental health outcomes have been measured as secondary outcomes. In other words, the comparison data includes that for which particular mental health outcomes might not have been anticipated – looking at anxiety outcomes in a study on depression medication, for example.[22]

The evidence base is insufficient to be able to draw firm conclusions from it.
The evidence base is certainly limited and peppered with caveats. Those who make such a critique may wish to apply it to the evidence base for medical treatment for gender-questioning teenagers, and decide whether a treatment that causes physical harm is justifiable on such a limited evidence base.

If you re-ran the analysis using a different technique, you might get different findings.
Full transparency in the approach and accompanying data tables has been given here to allow people to re-run the analysis or to amend the approach as appropriate. I have re-run the data in four different ways to test whether the findings would hold under different conditions. Under

[22] Li et al (2022).

each scenario, the placebo effect is greater than the effect of gender medication on mental health outcomes.

As it isn't a meta-analysis, key comparator studies may have been missed.
The review is certainly not exhaustive; it represents a first attempt to define the state of the evidence base, and I hope that follow-up studies will add a meta-analytical component. It is my view, though, that any bias will lie in the same direction in both gender medicine and comparator studies, as I have used the same search platform for both, and ensured the same scales are used and that data is weighted to support comparability. Adding other search platforms may correct for small biases, but I don't anticipate they will change the headline findings.

It hasn't been peer reviewed.
Formal peer review happens as part of the journal publication process. It is not something that can be undertaken by an independent researcher for both financial and logistical reasons. My hope is that this technical paper, representing a first pass at the data, will be picked up by an academic team and taken through this process, with my input. In the meantime, it has been informally peer reviewed by academic colleagues.

Conclusions

Analysis of current research demonstrates that gender medications, in the form of puberty blockers and hormone treatment, are not any better than taking a placebo in positively affecting teenagers' mental health. The limited available data suggests they may be worse.

Follow-up research in this area is urgently needed, including a full meta-analysis. Sourcing original data tables from study authors would be worthwhile. This would allow tests to be run for statistically significant differences in mental health outcomes between groups.

References

Placebo effect

Chiffi, D., & Zanotti, R. (2016, December). Knowledge and belief in placebo effect. *The Journal of Medicine and Philosophy*, 42(1): 70-85).

Koechlin, H., Kossowsky, J., Gaab, J., & Locher, C. (2018). How to address the placebo response in the prescription SSRIs and SNRIs in children and adolescents. *Expert Opinion on Drug Safety*, *17*(6), 537-540.

Hróbjartsson, A., Kaptchuk, T. J., & Miller, F. G. (2011). Placebo effect studies are susceptible to response bias and to other types of biases. *Journal of Clinical Epidemiology*, *64*(11), 1223-1229.

Meissner, K., Bingel, U., Colloca, L., Wager, T. D., Watson, A., & Flaten, M. A. (2011). The placebo effect: advances from different methodological approaches. *Journal of Neuroscience*, *31*(45), 16117-16124.

Kirsch, I. (2019). Placebo effect in the treatment of depression and anxiety. *Frontiers in Psychiatry*, *10*, 407.

Placebo effect relating to treatment for gender dysphoria[23]

Clayton, A. (2022a). Commentary on Levine: A Tale of Two Informed Consent Processes. *Journal of Sex & Marital Therapy*, 1-8.

Clayton, A. (2022b). Gender-Affirming Treatment of Gender Dysphoria in Youth: A Perfect Storm Environment for the Placebo Effect – The Implications for Research and Clinical Practice. *Archives of Sexual Behavior*, 1-12.

Biggs, M. (2020). Gender dysphoria and psychological functioning in adolescents treated with GnRHa: Comparing Dutch and English prospective studies. *Archives of Sexual Behavior*, 49(7), 2231-2236.

Biggs, M. (2020). *The Tavistock's Experimentation with Puberty Blockers.* Department of Sociology, University of Oxford.

Biggs, M. (2022). The Dutch Protocol for Juvenile Transsexuals: Origins and Evidence. *Journal of Sex & Marital Therapy*, 1-21.

Gosselin, J. T. (2012). Sexual dysfunctions and disorders. In *Psychopathology* (pp. 319-358). Routledge.

Mahfouda, S., Moore, J. K., Siafarikas, A., Zepf, F. D., & Lin, A. (2017). Puberty suppression in transgender children and adolescents. *The Lancet Diabetes & Endocrinology*, *5*(10), 816-826.

Moxon, S. P. (2022). Sex is not non-binary (or mutable), and neither is sexual identity or orientation. *New Male Studies*, *11*(1).

Nolan, B. J., Frydman, A. S., Leemaqz, S. Y., Carroll, M., Grossmann, M., Zajac, J. D., & Cheung, A. S. (2022). Effects of low-dose oral micronised progesterone on sleep, psychological distress,

[23] Search terms: *"placebo effect" OR "placebo response" "gender dysphoria" "puberty blockers" OR GnRHAs OR "cross-sex hormones" OR "hormone treatment" OR "gender-affirming hormones"*. Irrelevant studies were excluded from this list. The search was widened beyond adolescents for this, given a lack of specific literature on them.

and breast development in transgender individuals undergoing feminising hormone therapy: a prospective controlled study. *Endocrine Connections*, *11*(5).

Wright, C. (2020). A spoonful of sugar: Medication and the psychoanalytic body. *Psychoanalysis, Culture & Society*, *25*(2), 135-154.

Longitudinal studies on gender medication relating to adolescents

The following studies meet the selection criteria set out in this paper; the rationale for excluded studies (e.g. de Vries et al, 2014) is set out in section 3.

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*(1), 1-5.

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.

Carmichael, P., Butler, G., Masic, U., Cole, T. J., De Stavola, B. L., Davidson, 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.

Costa, R., Dunsford, M., Skagerberg, E., Holt, V., Carmichael, P., & Colizzi, M. (2015). Psychological support, puberty suppression, and psychosocial functioning in adolescents with gender dysphoria. *The Journal of Sexual Medicine*, *12*(11), 2206-2214.

de Lara, D. L., Rodríguez, O. P., Flores, I. C., Masa, J. L. P., Campos-Muñoz, L., Hernández, M. C., & Amador, J. T. R. (2020). Psychosocial assessment in transgender adolescents. Anales de Pediatría (English Edition), 93(1), 41-48.

De Vries, A. L., Steensma, T. D., Doreleijers, T. A., & Cohen-Kettenis, P. T. (2011). Puberty suppression in adolescents with gender identity disorder: A prospective follow-up study. The Journal of Sexual Medicine, 8(8), 2276-2283.

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).

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-e220978.

Comparator RCTs

Included

Berk, M., Mohebbi, M., Dean, O. M., Cotton, S. M., Chanen, A. M., Dodd, S., … & Davey, C. G. (2020). Youth Depression Alleviation with Anti-inflammatory Agents (YoDA-A): a randomised clinical trial of rosuvastatin and aspirin. *BMC Medicine*, 18(1), 1-12.

Correll, C. U., Tocco, M., Hsu, J., Goldman, R., & Pikalov, A. (2022). Short-term efficacy and safety of lurasidone versus placebo in antipsychotic-naïve versus previously treated adolescents with an acute exacerbation of schizophrenia. *European Psychiatry*, 65(1).

Davey, C. G., Chanen, A. M., Hetrick, S. E., Cotton, S. M., Ratheesh, A., Amminger, G. P., … & Berk, M. (2019). The addition of fluoxetine to cognitive behavioural therapy for youth depression (YoDA-C): a randomised, double-blind, placebo-controlled, multicentre clinical trial. *The Lancet Psychiatry*, 6(9), 735-744.

de la Loge, C., Hunter, S. J., Schiemann, J., & Yang, H. (2010). Assessment of behavioral and emotional functioning using standardized instruments in children and adolescents with partial-onset seizures treated with adjunctive levetiracetam in a randomized, placebo-controlled trial. *Epilepsy & Behavior,* 18(3), 291-298.

DelBello, M. P., Goldman, R., Phillips, D., Deng, L., Cucchiaro, J., & Loebel, A. (2017). Efficacy and safety of lurasidone in children and adolescents with bipolar I depression: a double-blind, placebo-controlled study. *Journal of the American Academy of Child & Adolescent Psychiatry*, 56(12), 1015-1025.

Findling, R. L., McKenna, K., Earley, W. R., Stankowski, J., & Pathak, S. (2012). Efficacy and safety of quetiapine in adolescents with schizophrenia investigated in a 6-week, double-blind, placebo-controlled trial. *Journal of Child and Adolescent Psychopharmacology*, 22(5), 327-342.

Goldman, R., Loebel, A., Cucchiaro, J., Deng, L., & Findling, R. L. (2017). Efficacy and safety of lurasidone in adolescents with schizophrenia: a 6-week, randomized placebo-controlled study. *Journal of Child and Adolescent Psychopharmacology*, 27(6), 516-525.

Ichikawa, H., Mikami, K., Okada, T., Yamashita, Y., Ishizaki, Y., Tomoda, A., … & Tadori, Y. (2017). Aripiprazole in the treatment of irritability in children and adolescents with autism spectrum disorder in Japan: A randomized, double-blind, placebo-controlled study. *Child Psychiatry & Human Development*, 48(5), 796-806.

Li, X., Mo, X., Liu, T., Shao, R., Teopiz, K., McIntyre, R. S., … & Lin, K. (2022). Efficacy of Lycium barbarum polysaccharide in adolescents with subthreshold depression: interim analysis of a randomized controlled study. *Neural Regeneration Research*, 17(7), 1582.

McDougle, C. J., Thom, R. P., Ravichandran, C. T., Palumbo, M. L., Politte, L. C., Mullett, J. E., … & Posey, D. J. (2022). A randomized double-blind, placebo-controlled pilot trial of mirtazapine for anxiety in children and adolescents with autism spectrum disorder. *Neuropsychopharmacology*, 47(6), 1263-1270.

Murphy, T. K., Brennan, E. M., Johnco, C., Parker-Athill, E. C., Miladinovic, B., Storch, E. A., & Lewin, A. B. (2017). A double-blind randomized placebo-controlled pilot study of azithromycin in youth with acute-onset obsessive–compulsive disorder. *Journal of Child and Adolescent Psychopharmacology*, 27(7), 640-651.

Robb, A. S., Cueva, J. E., Sporn, J., Yang, R., & Vanderburg, D. G. (2010). Sertraline treatment of children and adolescents with posttraumatic stress disorder: a double-blind, placebo-controlled trial. *Journal of Child and Adolescent Psychopharmacology*, 20(6), 463-471.

Singh, J., Robb, A., Vijapurkar, U., Nuamah, I., & Hough, D. (2011). A randomized, double-blind study of paliperidone extended-release in treatment of acute schizophrenia in adolescents. *Biological Psychiatry*, 70(12), 1179-1187.

Towbin, K., Vidal-Ribas, P., Brotman, M. A., Pickles, A., Miller, K. V., Kaiser, A., … & Stringaris, A. (2020). A double-blind randomized placebo-controlled trial of citalopram adjunctive to stimulant medication in youth with chronic severe irritability. *Journal of the American Academy of Child & Adolescent Psychiatry*, 59(3), 350-361.

Missing data

These are studies that are likely to have collected relevant data but did not report mean scores at both baseline and endline research stages. It would be worth requesting relevant data tables from study authors in any future meta analysis.[24]

Arango, C., Buitelaar, J. K., Fegert, J. M., Olivier, V., Pénélaud, P. F., Marx, U., … & Wolañczyk, T. (2022). Safety and efficacy of agomelatine in children and adolescents with major depressive disorder receiving psychosocial counselling: a double-blind, randomised, controlled, phase 3 trial in nine countries. *The Lancet Psychiatry*, 9(2), 113-124.

Findling, R. L., Atkinson, S., Bachinsky, M., Raiter, Y., Abreu, P., Ianos, C., & Chappell, P. (2022). Efficacy, Safety, and Tolerability of Flexibly Dosed Ziprasidone in Children and Adolescents with

[24] There are also two poster presentations for which the authors may have relevant data:

Davari-Ashtiani, R., Parvaresh, N., & Akhondzadeh, S. (2011). Gabapentin as a combination treatment with lithium in adolescents with bipolar disorder: a dou-ble-blind, randomized, placebo-controlled clinical trial. *International Clinical Psychopharmacology*, 26, e40.

Strawn, J. R., Moldauer, L., Hahn, R. D., Wise, A., Bertzos, K., Eisenberg, B., ... & Knutson, J. A. (2022). 2.9 A Multicenter Double-Blind, Placebo-Controlled Trial of Escitalopram in Children and Adolescents With Generalized Anxiety Disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, 61(10), S185.

Mania in Bipolar I Disorder: A Randomized Placebo-Controlled Replication Study. *Journal of Child and Adolescent Psychopharmacology*, *32*(3), 143-152.

Raison, C. L., Siu, C., Pikalov, A., Tocco, M., & Loebel, A. (2020). C-reactive protein and response to lurasidone treatment in children and adolescents with bipolar I depression: Results from a placebo-controlled trial. *Brain, Behavior, and Immunity*, 84, 269-274.

Strawn, J. R., Mills, J. A., Schroeder, H., Mossman, S. A., Varney, S. T., Ramsey, L. B., … & DelBello, M. P. (2020). Escitalopram in adolescents with generalized anxiety disorder: a double-blind, randomized, placebo-controlled study. *The Journal of Clinical Psychiatry*, 81(5), 6584. [CDRS-R]

Wehmeier, P. M., Schacht, A., Dittmann, R. W., Helsberg, K., Schneider-Fresenius, C., Lehmann, M., … & Ravens-Sieberer, U. (2011). Effect of atomoxetine on quality of life and family burden: results from a randomized, placebo-controlled, double-blind study in children and adolescents with ADHD and comorbid oppositional defiant or conduct disorder. *Quality of Life Research*, 20(5), 691-702.

Other

Boaden, K., Tomlinson, A., Cortese, S., & Cipriani, A. (2020). Antidepressants in children and adolescents: meta-review of efficacy, tolerability and suicidality in acute treatment. *Frontiers in Psychiatry*, 717.

De Vries, A. L., McGuire, J. K., Steensma, T. D., Wagenaar, E. C., Doreleijers, T. A., & Cohen-Kettenis, P. T. (2014). Young adult psychological outcome after puberty suppression and gender reassignment. *Pediatrics*, *134*(4), 696-704.

Higgins, J.P.T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J., & Welch, V.A. (Eds). (2022). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.3. Available from www.training.cochrane.org/handbook

Krishna, K. B., Fuqua, J. S., Rogol, A. D., Klein, K. O., Popovic, J., Houk, C. P., … & Lee, P. A. (2019). Use of gonadotropin-releasing hormone analogs in children: update by an international consortium. *Hormone Research in Paediatrics*, *91*(6), 357-372.

Li, Y., Huang, J., He, Y., Yang, J., Lv, Y., Liu, H., … & Li, L. (2019). The impact of placebo response rates on clinical trial outcome: A systematic review and meta-analysis of antidepressants in children and adolescents with major depressive disorder. *Journal of Child and Adolescent Psychopharmacology*, *29*(9), 712-720.

Locher, C., Koechlin, H., Zion, S. R., Werner, C., Pine, D. S., Kirsch, I., … & Kossowsky, J. (2017). Efficacy and safety of selective serotonin reuptake inhibitors, serotonin-norepinephrine

reuptake inhibitors, and placebo for common psychiatric disorders among children and adolescents: a systematic review and meta-analysis. *JAMA Psychiatry*, *74*(10), 1011-1020.

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. *Scientometrics*, *126*(1), 871-906.

McPhate, L., Williams, K., Vance, A., Winther, J., Pang, K., & May, T. (2021). Gender variance in children and adolescents with neurodevelopmental and psychiatric conditions from Australia. *Archives of Sexual Behavior*, 50(3), 863-871.

Meister, R., Abbas, M., Antel, J., Peters, T., Pan, Y., Bingel, U., … & Hebebrand, J. (2020). Placebo response rates and potential modifiers in double-blind randomized controlled trials of second and newer generation antidepressants for major depressive disorder in children and adolescents: a systematic review and meta-regression analysis. *European Child & Adolescent Psychiatry*, *29*(3), 253-273.

Rojas-Mirquez, J. C., Rodriguez-Zuñiga, M. J. M., Bonilla-Escobar, F. J., Garcia-Perdomo, H. A., Petkov, M., Becerra, L., … & Linnman, C. (2014). Nocebo effect in randomized clinical trials of antidepressants in children and adolescents: systematic review and meta-analysis. *Frontiers in Behavioral Neuroscience*, *8*, 375.

Salas-Humara, C., Sequeira, G. M., Rossi, W., & Dhar, C. P. (2019). Gender affirming medical care of transgender youth. *Current Problems in Pediatric and Adolescent Health Care*, *49*(9), 100683.

van der Miesen, A. I., de Vries, A. L., Steensma, T. D., & Hartman, C. A. (2018). Autistic symptoms in children and adolescents with gender dysphoria. *Journal of Autism and Developmental Disorders*, 48(5), 1537-1548.

Vlot, M. C., Klink, D. T., den Heijer, M., Blankenstein, M. A., Rotteveel, J., & Heijboer, A. C. (2017). Effect of pubertal suppression and cross-sex hormone therapy on bone turnover markers and bone mineral apparent density (BMAD) in transgender adolescents. *Bone*, *95*, 11-19

Zucker, K. J. (2019). Adolescents with gender dysphoria: Reflections on some contemporary clinical and research issues. *Archives of Sexual Behavior,* 48(7), 1983-1992.

Additional tables

Table 7. Main analysis

Study	GM/P	Scale	Sample size BL	Sample size EL	BL mean	EL mean	Change	Change SD	Calcs* used for SD?
Achille et al (2020)	GM	QLES-Q	50	50	61.5	72	10.5	42.2	*
Carmichael et al (2021)	GM	CGAS	42	12	62.9	66	3.1	7.8	**
Costa et al (2015)	GM	CGAS	101	35	58.72	67.4	8.68	14.0	*
De Lara et al (2020)	GM	BDI	23	23	19.3	9.7	9.5	2.9	
De Vries et al (2011)	GM	BDI	41	41	8.31	4.95	3.36	7.0	*
De Vries et al (2011)	GM	CGAS	41	41	70.24	73.90	3.66	7.9	*
Kuper et al (2020)	GM	SCARED	102	102	32.4	28.6	3.8	15.8	**
Kuper et al (2020)	GM	QIDS	118	118	9.4	7.3	2.1	4.4	**
Kuper et al (2020)	GM	QIDS	125	125	5.8	5.9	-0.1	3.6	**
Berk et al (2020)	P	QIDS	42	37	17.7	12.3	5.4	N/A	
Berk et al (2020)	P	QLES-Q	42	37	34.6	50	15.4	N/A	
Correll et al (2022)	P	CGAS	18	18	47.4	50.1	2.7	12.3	
Correll et al (2022)	P	CGAS	94	94	43.3	49.8	6.5	11.6	
Correll et al (2022)	P	QLES-Q	18	18	52.9	49.4	-3.5	11.9	
Correll et al (2022)	P	QLES-Q	94	94	52.4	53.1	0.7	13.6	
Davey et al (2019)	P	QIDS	77	77	17.0	11.0	6	5.1	*
Davey et al (2019)	P	QLES-Q	77	77	36.5	48.4	11.9	14.1	*
Delbello et al (2017)	P	CGAS	170	170	49.5	58.8	9.3	12.9	
Delbello et al (2017)	P	QLES-Q	170	170	49.7	57.6	7.9	14.3	
Findling et al (2012)	P	CGAS	73	73	41.8	51.78	9.98	13.1	
Goldman et al (2017)	P	CGAS	112	112	43.9	50	6.1	11.6	
Goldman et al (2017)	P	QLES-Q	112	112	52.5	53.9	1.4	13.8	
Ichikawa et al (2017)	P	CGAS	45	45	42.3	46.8	4.5	9.4	
Li et al (2022)	P	BDI	14	14	31.29	22.219	9.071	11.6	*
Li et al (2022)	P	SCARED	14	14	48.43	45.43	3	18.2	*
McDougle et al (2022)	P	SCARED	10	10	28.1	23	5.1	10.5	*
Murphy et al (2017)	P	CGAS	14	14	49.07	52.68	3.61	N/A	
Murphy et al (2017)	P	SCARED	14	14	20.43	21.97	-1.54	N/A	
Robb et al (2010)	P	QLES-Q	62	61	49.5	59.8	10.3	10.9	
Singh et al (2011)	P	CGAS	51	51	48.8	53.8	5	13.82	
Strawn et al (2015)	P	CGAS	137	133	48.6	60.8	12.2	13.8	
Towbin et al (2020)	P	CGAS	26	26	42.9	47.2	4.3	N/A	

Data has been rounded in the data tables, but full, unrounded figures were used in the data calculations; there may be minor variations in results if using the data tables rather than the original data.

Sample sizes are taken from endline measurement points; baseline figures are higher due to attrition. GM: gender medication studies (puberty blockers and/or hormone treatment) | PBO: placebo (from comparator studies) | BL: baseline | EL: endline | SD: standard deviation |

*A calculation was necessary to calculate the SD of the mean change that was more than just using the standard error/sample size.

**A correlation coefficient was necessary to calculate the SD.

Note that repeated measures for Correll et al (2022) show different groups within the same study, and should be treated as independent when looking at the same scale.

Measures that had to be left out of the analysis due to absence of mean change SD data are shown in grey.

Table 8. Alternative analysis

Study	GM/P	Scale	Scale reversed?	Sample size EL	Adjusted EL sample	BL mean	EL mean	Change	Change SD
Achille et al (2020)	GM	QLES-Q		50	16.7	61.5	72	10.5	42.2
Achille et al (2020)	GM	CESD-R	Y	48	16.0	39.6	47.1	7.5	13.0
Achille et al (2020)	GM	PHQ-9	Y	49	16.3	19.2	22.75	3.6	6.6
Allen et al (2019)	GM	GWBS		47	47	61.7	70.23	8.5	13.6
Carmichael et al (2021)	GM	CGAS		12	4.0	62.9	66	3.1	8.4
Carmichael et al (2021)	GM	CBCL*	Y	11	3.7	20.6	18.5	-2.1	8.6
Carmichael et al (2021)	GM	YSR*	Y	15	5.0	27.1	25.1	-2.0	5.8
Costa et al (2015)	GM	CGAS		35	35	58.72	67.4	8.7	14.0
De Lara et al (2020)	GM	BDI	Y	23	5.75	44.7	54.3	9.6	2.9
De Lara et al (2020)	GM	SDQ**	Y	23	5.75	5.8	7.6	1.8	2.3
De Lara et al (2020)	GM	STAI-S	Y	23	5.75	47.7	64.2	16.5	20.9
De Lara et al (2020)	GM	STAI-T	Y	23	5.8	48	62.5	14.5	18.3
De Vries et al (2011)	GM	CGAS		41	8.2	70.24	73.90	3.7	7.9
De Vries et al (2011)	GM	BDI	Y	41	8.2	92.7	96.05	3.4	6.4
De Vries et al (2011)	GM	CBCL*	Y	54	10.8	20.0	26.54	6.5	13.8
De Vries et al (2011)	GM	STAI	Y	41	8.2	41.57	43.05	1.5	8.6
De Vries et al (2011)	GM	YSR*	Y	54	10.8	25.0	31.22	6.3	13.2
Kuper et al (2020)	GM	QIDS	Y	118	39.3	18.6	20.7	2.1	4.2
Kuper et al (2020)	GM	QIDS	Y	125	41.7	22.2	22.1	-0.1	3.5
Kuper et al (2020)	GM	SCARED	Y	102	34.0	50.6	54.4	3.8	15.3
Correll et al (2022)	P	CGAS		18	9	47.4	50.1	2.7	12.3
Correll et al (2022)	P	CGAS		94	47	43.3	49.8	6.5	11.6
Correll et al (2022)	P	QLES-Q		18	9	52.9	49.4	-3.5	11.9
Correll et al (2022)	P	QLES-Q		94	47	52.4	53.1	0.7	13.6
Davey et al (2019)	P	QLES-Q		77	38.5	36.5	48.4	11.9	14.1
Davey et al (2019)	P	QIDS	Y	77	38.5	11.0	17	6.0	5.1
Delbello et al (2017)	P	CGAS		170	85	49.5	58.8	9.3	12.9
Delbello et al (2017)	P	QLES-Q		170	85	49.7	57.6	7.9	14.3
Findling et al (2012)	P	CGAS		73	73	41.8	51.78	10.0	13.1
Goldman et al (2017)	P	CGAS		112	56.0	43.9	50	6.1	11.6

Goldman et al (2017)	P	QLES-Q		112	56.0	52.5	53.9	1.4	13.8
Ichikawa et al (2017)	P	CGAS		45	45.0	42.3	46.8	4.5	9.4
Li et al (2022)	P	BDI	Y	14	7.0	32.7	41.781	9.1	11.6
Li et al (2022)	P	SCARED	Y	14	7.0	34.57	37.57	3.0	18.2
McDougle et al (2022)	P	SCARED	Y	10	10.0	54.9	60	5.1	10.5
Rob et al (2010)	P	QLES-Q		61	61	49.5	59.8	10.3	10.9
Singh et al (2011)	P	CGAS		51	51	48.8	53.8	5.0	13.82
Strawn et al (2015)	P	CGAS		133	133	48.6	60.8	12.2	13.8

Sample sizes are taken from endline measurement points; baseline figures are higher due to attrition. GM: gender medication studies (puberty blockers and/or hormone treatment) | PBO: placebo (from comparator studies) | BL: baseline | EL: endline | SD: standard deviation | *internalising sub-scale; **emotional sub-scale.

Note that repeated measures for Correll et al (2022) show different groups within the same study, and should be treated as independent when looking at the same scale.

Measures that had to be left out of the analysis due to absence of mean change SD data can be found in the first table in this section, under the main analysis.

Appendices

Additional analysis notes

Assumptions

Normal distributions were assumed for all variables. This was checked by looking at confidence intervals, where available, and where they sat in relation to the mean. The average of the confidence intervals of the mean change in relation to the mean change itself did not vary by more than 3% of a single SD, making the assumption hold true for the variables for which data was available. It should be noted, though, that confidence intervals (as original study data, as opposed to ones calculated using standard error/t distribution data) were not available for the majority of studies.

In several cases in the alternative analysis, the p value of the mean change was reported to be >0.001. The p value was necessary in calculating the SD of the mean change, as this was not

reported in any of the additional measures that formed part of this analysis. In the absence of a precise figure, 0.001 was used.

Calculating SDs with missing data

Where a standard error (SE) of the mean change was available, the following formula was used to calculate the SD:

$$SD = \frac{M_{endline} - M_{baseline}}{SE}$$

Where a confidence interval (CI) of the mean change was available, the formula was as follows, where UL is the upper limit of the confidence interval, and LL is the lower limit (T was calculated using the TINV function [=TINV, 0.05, DF], where DF = degrees of freedom, as sample sizes were too small to use the standard 3.92 denominator):

$$SD = \frac{\sqrt{N}X(CI_{UL} - CI_{LL})}{2t}$$

Where the p value of the mean change was available, the T statistic was calculated as follows: [=TINV, p, DF], and the SE was calculated as shown below, then used to calculate the SD as shown previously:

$$SE = |\frac{MD}{t}|$$

Confidence intervals of the mean change, where they were not available from the original studies, were calculated using standard errors as below (for the purposes of later creation of forest plots). As with the earlier note on creation of a t statistic, the TINV function was used – with the 95% confidence interval and relevant degrees of freedom – in place of the standard 1.96 figure, as sample sizes were too small to use the standard figure.

$$CI = \pm SE \times t$$

There were four measures in the gender medication group for which data on mean difference standard deviations were missing and for which no alternative data was available that would have allowed these to be calculated (95% CIs of the mean difference, p values relating to the mean change, SE of the mean change, etc). In these cases, correlation coefficients were calculated for the comparable studies for which data was available, using the approach outlined in Cochrane's guidelines. There were four GM measures that had the necessary data to

calculate this, or for which the necessary data could be calculated: standard deviations for the baseline mean, the endline mean and the mean difference. The formula used was as follows:

$$Corr = \frac{SD^2_{baseline} + SD^2_{endline} + SD^2_{change}}{2 \times SD_{baseline} \times SD_{endline}}$$

The average of the four correlation coefficients (Corra) was used to impute the SD of the baseline to endline change for the four measures in the gender medication group with missing data using the following approach:

$$SD_{change} = \sqrt{SD^2_{baseline} + SD^2_{endline} - (2 \times Corr_a \times SD_{baseline} \times SD_{endline}}$$

There were no comparator studies that had all the information needed to calculate a correlation coefficient; most were missing the endline standard deviation. The five measures with missing data from comparator studies therefore had to be excluded.

In the alternative analysis, three out of the 11 additional measures required a correlation coefficient to calculate the SD of the mean change. Five additional measures had the data needed to calculate a correlation coefficient. My original intention was to add these into the original four correlation coefficients, but two of the additional five measures had a high, negative coefficient. This made the average correlation coefficient much smaller than in the main analysis (0.10, as opposed to 0.63). A decision was made to keep to the original average of 0.63 – selecting the smaller coefficient would make it more likely that a smaller effect size was found for the gender medication, and I wanted to make sure that all close calls favoured the gender medication effect (making the findings more robust in the face of potential criticism/scrutiny by proponents of gender-affirming medical care).

Transformations

In the alternative analysis, some scales showed a positive effect through an increase between baseline and endline scores (CGAS, GWBS, QLES) and others showed a positive effect through a decrease between baseline and endline scores (BDI, CBCL, CESD-R, PHQ-9, QIDS, SCARED, STAI). The latter set of scales were reversed in order to make them comparable with the former set in the pooled analysis.

Where more than one measure of mental health outcome was available within a single study in the alternative analysis, data was weighted accordingly to ensure that the same individual who responded to more than one mental health measure was not over-represented in the final analysis. For example, Kuper et al (2011) used three different measures: QIDS (self-report), QIDS (clinician report) and SCARED, with an original endline sample size of 118, 125 and 102

respectively. Each sample size was divided by the total of the three figures to get the proportion of the total study sample it represented, then multiplied by the average of the three figures to get an adjusted sample size. The final adjusted sample sizes in this example were 39.3, 41.7 and 34.0 (total: 115, which is the average sample size of the three original figures).

Rationale for merging BDI and QIDS

There was only a single study with 14 participants for the BDI comparator group, and a single study with 77 participants in the QIDS comparator group. Having only one comparator study in each group increases the risk of bias, especially for the BDI group as so few people took part. All other scales had at least two studies included in the comparator analysis. Merging them had the additional benefit of making reporting more straightforward, as depression (as with the other areas of general psychological functioning, quality of life and anxiety) could be reported at an aggregate level, instead of by individual scale, making it more accessible to a lay readership. The downside of this approach is that the scales have a different maximum score. (This lack of a decent-sized sample is, of course, a consistent limitation of the gender medication studies too, and one that lends weight to the premise of this paper's analysis.)

Data availability and study notes

Gender medication studies

		Key data	Notes
Achille et al (2020)	QLES, CESD, PHQ	BLM, p value of mean change, ELM	
Allen et al (2019)	GWBS	BLM, BLM SE, ELM, ELM SE	
Carmichael et al (2021)	CGAS, CBCL, YSR	BLM, BLM CIs, ELM, ELM CIs	
Costa et al (2015)	CGAS	BLM, BLM SD, t distribution, ELM, ELM SD	
de Lara et al (2020)	BDI, SDQ, STAI-S, STAI-T	BLM, BLM SD/SE, p value of mean change, SE of mean change (BDI/STAI-T only), ELM, ELM SD/SE	When de Lara et al (2020) reported on SDQ, they stated the figures they gave were SD. However, these figures were assumed in the analysis to be SE, given their size and how they compared with other reported SDs in the same paper. STAI-T was calculated in the same way as STAI-S for comparability (SE of mean change was available only for STAI-T). An assumption was also made

			that 0.6 was the standard error of the mean change for BDI. This was not made clear in the reporting, but it is too small to be a confidence interval or a standard deviation. The data tables gave a mean change of 9.6, but a mean change of 9.5 was reported in the text. The latter has been used in the analysis, on the assumption that the difference was due to rounding in the data tables.
de Vries et al (2014)	BDI, CGAS, CBCL, YSR, STAI	BLM, BLM SD, f statistic, p value of mean change, ELM, ELM SD	
Kuper et al (2020)	SCARED, QIDS x 2	BLM, BLM SD, ELM, ELM SD	There were two measures of QIDS (self-report and clinician report). These were combined, as the measures covered the same individuals. The sample size varied slightly, so weighted baseline and endline means were calculated, and a measure of pooled standard deviation of the mean change was derived.

BLM: baseline mean | ELM: endline mean | SD: standard deviation | SE: standard error | CI: confidence interval

Comparator studies

		Key data	Notes
Berk et al (2020)	QIDS, QLES	BLM, BLM SD, ELM, ELM SD	In the absence of correlation coefficients for comparator studies, there was insufficient data to calculate the mean change SD, so this study was excluded from the final analysis.
Correll et al (2022)	CGAS, QLES	BLM, BLM SD, mean change, mean change SE	There were two separate groups of participants within the same study. This has been accounted for in the analysis.
Davey et al (2019)	QIDS, QLES	BLM, BLM SD, mean change, CIs of mean change	

Delbello et al (2017)	CGAS, QLES	BLM, BLM SD, mean change, mean change SE	
Findling et al (2012)	CGAS	BLM, BLM SE, mean change, mean change SE, CIs of mean change	
Goldman et al (2017)	CGAS, QLES	BLM, BLM SD, mean change, mean change SE	
Ichikawa et al (2017)	CGAS	BLM, BLM SE, mean change, mean change SE	
Li et al (2022)	BDI, SCARED	BLM, BLM SD, mean change, CIs of mean change	Mean changes and related confidence intervals were reported graphically, not numerically. The linked numbers were therefore estimated using the graphs, so may not be 100% accurate.
McDougle et al (2022)	SCARED	BLM, BLM SD, mean change, CIs of mean change	
Murphy et al (2017)	CGAS, SCARED	BLM, BLM SE, ELM, ELM SE	As with Berk et al, there was insufficient data to calculate the mean change SD, so this study was excluded from the final analysis.
Robb et al (2010)	QLES	BLM, BLM SD, mean change, mean change SE	
Singh et al (2011)	CGAS	BLM, BLM SD, mean change, mean change SD	
Strawn et al (2015)	CGAS	BLM, BLM SD, mean change, mean change SE	
Tobin et al (2020)	CGAS	BLM, BLM SE, ELM, ELM SE	As with Berk et al, there was insufficient data to calculate the mean change SD, so this study was excluded from the final analysis.

BLM: baseline mean | ELM: endline mean | SD: standard deviation | SE: standard error | CI: confidence interval

About the author

I am an independent social research consultant. This technical paper was prepared as part of work to clarify the evidence base for parents who may have a gender-questioning teenager, which is published separately by Sex Matters and will eventually inform a book chapter. No funding was received for this paper or for the accompanying research series. There are no financial conflicts of interest to declare.

Many thanks to Dr Amanda Gosling and Dr Alison Clayton for their comments and advice. Any errors are my own.

Sex Matters is a not-for-profit company registered by guarantee.
Company number: 12974690
Registered office: 63/66 Hatton Garden, Fifth Floor Suite 23, London, EC1N 8LE

Published 15th December 2022