AI Manuscript Review

Manuscript

Generated: 1/28/2026
Review Type: Premium Multi-Agent Review

Editorial Decision: PENDING

EXECUTIVE SUMMARY

This manuscript was reviewed by 7 AI agents. The majority recommendation is 'Revise and Resubmit' with an average score of 5/10. A total of 11 consensus issues and 10 unique concerns were identified. Please review the detailed feedback below for specific guidance on required revisions.

DECISION LETTER

Editorial Summary: Manuscript Under Review

Manuscript ID: vOpoAnkPj0PVU2QjS9Tr **Date:** 2026-01-29 **Editor:** AI Editor-in-Chief
Number of Reviews: 7

Overview of Reviews

This manuscript has been reviewed by our multi-agent panel. See the Decision Letter above for detailed reviewer summary data.

Points of Consensus

1. **MAJOR: No Power Analysis or Sample Size Justification** — Raised by multiple reviewers - Sample sizes vary substantially across outcomes (76,189 for wellbeing vs. 223,568 for virgin status). For non-significant findings, lack of power analysis prevents distinguishing between true null effects and statistical limitations.

2. **G: Abstract, Main outcome measures section (Paragraph 4)** — Raised by multiple reviewers - The Oxford comma before "and" improves clarity in this list. "Sub- sample" should be written as one word "subsample" without hyphenation. "Age of first having sex" is grammatically awkward.

3. **S: Key Messages (Paragraph 5)** — Raised by multiple reviewers - "Casual" and "causal" have completely different meanings. This is a critical error that changes the scientific meaning of the statement.

4. **S: Introduction (Paragraph 6)** — Raised by multiple reviewers - "Short-term" should be hyphenated when used as a compound adjective modifying "adverse effects."

5. **C: Introduction (Paragraph 7)** — Raised by multiple reviewers - Quotation marks around "over the counter" are unnecessary for this common phrase. "Such as" is preferred over "like" in formal scientific writing when providing examples.

Points of Divergence

Reviewers were largely in agreement on the key issues identified above.

Required Revisions

1. **MAJOR: No Power Analysis or Sample Size Justification** — Source: systematic - Action: Add methods subsection: "We calculated statistical power using an online MR power calculator [cite mRnd], assuming R^2 of instruments with exposure = [X], sample sizes as reported, and $\pm = 0.05$ (two-tailed)."

2. **G: Abstract, Main outcome measures section (Paragraph 4)** — Source: technical_reviewer - Action: Change to: Number of children, age at first sexual intercourse, number of sexual partners, odds of being a virgin, and self-reported wellbeing, all measured in the male subsample of the UKB.

3. **S: Key Messages (Paragraph 5)** — Source: technical_reviewer - Action: Change to: We find evidence for a causal association

4. **S: Introduction (Paragraph 6)** — Source: technical_reviewer - Action: Change to: Randomized trials have been effective in identifying various short-term adverse effects

5. **C: Introduction (Paragraph 7)** — Source: technical_reviewer - Action: Change to: Now that PDE5 inhibitors are available over the counter in countries such as the United Kingdom

6. **C: Methods, Study design (Paragraph 9)** — Source: technical_reviewer - Action: Change to: The *cis* analysis considered only variants within or near the gene region encoding the drug target as potential instruments.

7. **G: Methods (Paragraph 10)** — Source: technical_reviewer - Action: Change to: Variant-outcome information on number of children fathered (OpenGWAS ID: ukb-b-2227, 209,872 males) was extracted from The Elsworth UK Biobank GWAS in the OpenGWAS repository

8. **C: Figure 4 Caption** — Source: technical_reviewer - Action: Change to: H0 to H4 represent the posterior probabilities of these hypotheses in the Bayesian colocalisation.

9. **C: Abstract, Secondary outcomes (Paragraph 4)** — Source: technical_reviewer - Action: Change to: *cis*-MR

10. **[Statistical] Sample Size** — Source: Statistical Methods Agent - Action: 1) Conduct a sensitivity analysis using non-overlapping samples where possible. Consider using the ICBP data excluding UK Biobank participants for the exposure GWAS, or use an alternative blood pressure dataset.

Recommended Revisions

- Failure to Address Substantial Sample Overlap — Source: domain_expert
- Conceptual Mismatch Between Genetic Proxy and Pharmacological Intervention — Source: domain_expert
- Causal language overreaches the design/identification — Source: skeptic
- Two-sample MR claim conflicts with substantial sample overlap (bias risk) — Source: skeptic
- Exposure/proxy definition is ambiguous and potentially invalid (PDE5 inhibition vs BP vs sildenafil) — Source: skeptic

Optional Suggestions

Please review the detailed individual reviewer reports for additional suggestions that may strengthen the manuscript.

Editor's Comments to Authors

This manuscript has received thorough review from our multi-agent panel. Please carefully address each of the required revisions listed above. The individual reviewer reports contain detailed feedback with specific locations and recommendations.

Editorial Decision

Decision: Revise and Resubmit

Rationale: Based on 7 independent reviews, the consensus recommendation is 'Revise and Resubmit'. A total of 11 consensus issues were identified that require attention before the manuscript can proceed.

REQUIRED CHANGES

1. MAJOR: No Power Analysis or Sample Size Justification
2. G: Abstract, Main outcome measures section (Paragraph 4)
3. S: Key Messages (Paragraph 5)
4. S: Introduction (Paragraph 6)
5. C: Introduction (Paragraph 7)
6. C: Methods, Study design (Paragraph 9)
7. G: Methods (Paragraph 10)
8. C: Figure 4 Caption
9. C: Abstract, Secondary outcomes (Paragraph 4)
10. [Statistical] Sample Size: STAT-001

SUGGESTED IMPROVEMENTS

1. Failure to Address Substantial Sample Overlap

2. Conceptual Mismatch Between Genetic Proxy and Pharmacological Intervention

3. Causal language overreaches the design/identification

4. Two-sample MR claim conflicts with substantial sample overlap (bias risk)

5. Exposure/proxy definition is ambiguous and potentially invalid (PDE5 inhibition vs BP vs sildenafil)

Domain Expert

Model: gemini/gemini-2.5-pro

Journal Article Review: Manuscript Under Review

Summary Assessment

This manuscript presents a well-designed and highly relevant drug-target Mendelian Randomisation (MR) study investigating the causal effects of sildenafil on male fertility and wellbeing. The study's premise is strong, the introduction is exceptionally clear, and the planned use of sophisticated methods like colocalization demonstrates a methodologically robust approach. However, the study's central conclusions are currently undermined by a major, unaddressed methodological flaw concerning substantial sample overlap between the exposure and outcome datasets.

Major Concerns

Sample Overlap Issue

1. **Failure to Analytically Address Substantial Sample Overlap** — **Location**: Methods/Statistical Analysis Section

- **Problem**: The authors correctly identify a critical issue in their analysis, stating, *"Since around two-thirds of the participants in the Evangelou et al. GWAS were from the UKB we expect there to be substantial sample overlap between our exposure and outcome samples."* However, merely acknowledging this problem is insufficient. In two-sample MR, significant sample overlap, particularly when combined with weak instruments, biases the causal estimate towards the confounded observational association, potentially invalidating the results. This is a well-established limitation in the field, and current standards require it to be analytically managed.

- **Recommendation**: The authors must implement a statistical solution to account for this overlap. The following are standard approaches in the field:

1. Employ a method like GSMR (Generalized Summary-data-based Mendelian Randomization) that models the LD structure and can account for sample overlap.

2. Conduct a one-sample MR analysis entirely within the UK Biobank for both the exposure (blood pressure) and the outcomes. This would serve as a crucial sensitivity analysis to validate the two-sample findings.

3. Quantify the potential bias using established methods, such as those described by Burgess et al. (Stat Med, 2016), to demonstrate the likely magnitude and direction of the bias on the reported effect estimates.

Strengths

- The study employs a sophisticated and appropriate drug-target cis-MR design, using genetic variants in the *PDE5A* gene region as a proxy for sildenafil action, which is a significant strength.
- The introduction provides an exceptionally clear and compelling rationale for the study, effectively situating the work within the current literature.
- The planned use of multiple robust sensitivity analyses, including colocalization and two-step MR, indicates a high level of methodological rigor.

Recommendation

Major Revision

Justification: The study addresses an important clinical question with a strong and sophisticated study design. However, the failure to analytically address the substantial sample overlap between the exposure and outcome datasets is a major methodological flaw that undermines confidence in the reported causal estimates. The manuscript is potentially suitable for publication, but only after this critical issue has been thoroughly addressed through additional sensitivity analyses as recommended above.

Adversarial Skeptic

Model: openai/gpt-5.2

Journal Article Review: Manuscript Under Review

Summary Assessment

The manuscript addresses an interesting and potentially high-impact question—whether genetic proxies of PDE5 inhibition relate to male reproductive outcomes—using a cis-MR framework with pre-specified outcomes and a stated robustness plan (e.g., colocalisation, two-step cis-MR, replication). The overall rationale is clear.

However, the current framing repeatedly drifts from *genetic evidence about target perturbation* to *claims about sildenafil/PDE5 inhibitor treatment effects*, which exceeds what cis-MR can support unless several strong identification assumptions are explicitly demonstrated and upheld.

Major Concerns

1. **Treatment-effect (drug-use) claims are overstated relative to cis-MR identification** — *Abstract/Conclusions (Paragraphs 4–5)*:

- **Quote**: "**Conclusions: This study provides genetic support for PDE5 inhibitors increasing the number of children that men have.**" (Paragraph 4) - **Quote**: "**We find evidence for a casual [sic] association between genetically proxied sildenafil use and number of children fathered.**" (Paragraph 5) - **Challenge**: The authors implicitly assume that "genetically proxied PDE5 inhibition" is equivalent to "sildenafil use" and further equivalent to "PDE5 inhibitors increasing" the outcome—i.e., a drug treatment effect. But cis-MR, even when well executed, is evidence about lifelong genetic perturbation of a target region and only maps to drug effects under additional assumptions (e.g., instrument validity within the cis-region, absence of LD-confounding/pleiotropy, correct target and direction, meaningful scaling to pharmacologic inhibition, and no design features that inflate certainty). - **Why it matters**: Readers may interpret the work as supporting a clinical recommendation ("PDE5 inhibitors increase fertility/number of children"), when the design more cautiously supports "evidence consistent with PDE5A target perturbation being associated with…" Overstating the estimand and certainty risks misinterpretation, particularly in a medically actionable context. - **What would address it**: - Revise causal/treatment language throughout to **target-perturbation language** (e.g., "genetic evidence consistent with PDE5A perturbation being associated with number of children fathered") and avoid "sildenafil use" unless you explicitly justify that mapping and scaling. - Add explicit wording in the Abstract and Discussion that cis-MR does **not** directly estimate the effect of taking sildenafil (dose, timing, duration) and that interpretation as a drug effect requires additional assumptions. - Ensure the final claims are aligned with what your robustness checks can actually rule out (e.g., colocalisation supporting a shared causal variant helps, but does not by itself convert a cis-association into a clinical drug effect).

Minor Issues

- **Abstract (Paragraph 5)**: Typo/wording — "**casual** association" should be "**causal** association."

Strengths

- Clear motivation for a drug-target MR (cis-MR) approach applied to PDE5/PDE5A and reproductive outcomes.
- Outcomes appear pre-specified, and the manuscript signals an appropriate robustness strategy (e.g., colocalisation, two-step cis-MR, replication).

Questions for Authors

1. What is the precise estimand: **PDE5A target perturbation**, **PDE5 inhibition generally**, or **sildenafil exposure**? How do you justify the mapping from the genetic proxy to "sildenafil use" specifically?
2. What explicit assumptions are you making to translate lifelong genetic perturbation into an inference about **pharmacologic PDE5 inhibitor treatment**, and where are these stated (ideally in Abstract/Key Messages/Discussion)?
3. How will you phrase conclusions if robustness checks support target-level association but do not uniquely support "sildenafil/PDE5 inhibitors increase…" as a treatment claim?

Recommendation

Major Revision

Justification: The central scientific question is strong and the analytic framework is plausibly

appropriate, but the Abstract/Conclusions currently overstate what cis-MR can identify by implying a drug-use/treatment effect ("sildenafil use," "PDE5 inhibitors increasing…"). Correcting the estimand, causal language, and interpretive boundaries is necessary to prevent clinically consequential over-interpretation.

Systematic Reviewer

Model: anthropic/claude-sonnet-4-5-20250929

Journal Article Review: Manuscript Under Review

Summary Assessment

This Mendelian randomization study investigates the association between genetically proxied PDE5 inhibition (sildenafil) and fertility/sexual activity outcomes in men using ICBP exposure data (N=757,601) and UK Biobank outcomes (N"H 4 5 0 , 0 0 0) .

While the research question is interesting and potentially clinically relevant, the manuscript has a fundamental methodological flaw that must be addressed before publication: substantial sample overlap between exposure and outcome datasets (~60%) that violates two-sample MR assumptions and is acknowledged but not corrected. Additionally, the methods section is incomplete, cutting off mid-sentence, preventing full evaluation of the study's methodological rigor.

Major Concerns

1. Sample Overlap Violates Two-Sample MR Assumptions

- **Problem**: The authors acknowledge that approximately two-thirds of the ICBP GWAS participants ("H 4 5 4 ,d 6 7 7,601) overlap with the UK Biobank outcome sample ("H 4 5 0 , 0 0 0) . This represents near-complete overlap and fundamentally violates the independence assumption of two-sample MR. This overlap can: - Inflate Type I error rates and bias effect estimates - Introduce winner's curse bias - Invalidate reported standard errors and confidence intervals - Make the reported p-values unreliable Simply acknowledging this problem without correction is insufficient. The primary finding (0.21 additional children, 95% CI: 0.08-0.35, FDR p=0.01) cannot be interpreted as valid under standard two-sample MR assumptions.

- **Recommendation**: The authors must either: 1. Re-analyze using only non-overlapping ICBP cohorts for exposure data 2. Apply established correction methods for sample overlap (e.g., Burgess et al. 2016 adjustments to standard errors) 3. Convert to a one-sample MR framework with appropriate adjustments 4. Demonstrate through simulation that the degree of overlap does not materially bias results for this specific scenario Without addressing this, the statistical validity of all findings is compromised.

2. Incomplete Methods Section

- **Problem**: The methods section cuts off mid-sentence during the description of instrument selection: "...or (4) conduct one-samp". This prevents full evaluation of: - Complete instrument selection criteria - Details of sensitivity analyses - Full statistical methodology - How pleiotropy was assessed - What additional robustness checks were performed

- **Recommendation**: Provide the complete methods section. Ensure all standard MR methodological elements are included: instrument selection criteria (F-statistics, R^2 values), pleiotropy assessment (MR-Egger intercept, heterogeneity tests), sensitivity analyses (weighted median, MR-PRESSO), and handling of weak instruments.

Minor Issues

[Note: The summary indicates 7 minor issues exist, but detailed descriptions were not provided in the section reviews. These should be itemized once the complete manuscript is available for review.]

Strengths

- **Clinically relevant research question**: The potential fertility effects of PDE5 inhibitors are understudied and have public health implications
- **Appropriate use of MR framework**: Mendelian randomization is suitable for investigating potential causal effects of medications
- **Transparent acknowledgment**: Authors openly acknowledge the sample overlap issue, demonstrating scientific integrity
- **Large sample sizes**: Both exposure and outcome datasets are substantial, providing adequate statistical power

Questions for Authors

1. What proportion of the ICBP sample comes from UK Biobank versus other cohorts? Can you provide a breakdown?

2. Have you conducted sensitivity analyses restricted to non-overlapping samples to verify the robustness of findings?

3. What are the F-statistics and variance explained (R^2) for the PDE5A genetic instruments?

4. Were any pleiotropy assessment methods applied (MR-Egger, heterogeneity tests, MR-PRESSO)?

5. Why was the decision made to proceed with two-sample MR despite acknowledging substantial overlap rather than using one-sample methods from the outset?

Recommendation

Major Revision

Justification: The manuscript addresses an interesting and clinically relevant question using appropriate genetic epidemiology methods. However, the fundamental methodological flaw of substantial sample overlap between exposure and outcome datasets invalidates the current statistical inference. This must be corrected before the work can be considered for publication. Additionally, the incomplete methods section prevents full evaluation of the study's rigor. Once these major concerns are addressed and the manuscript is complete, the work has potential merit for publication.

Pragmatic Reviewer

Model: anthropic/claude-sonnet-4-5-20250929

Journal Article Review: Manuscript Under Review

Summary Assessment

This Mendelian randomization study investigating sildenafil's effects on male fertility presents methodologically sound work with a potentially clinically meaningful finding (0.21 additional children fathered). However, the manuscript suffers from significant accessibility barriers that would limit its impact on clinical readers and policymakers. The core finding lacks practical context necessary for interpretation, and the rationale for using genetic proxies to study a widely-available, well-characterized drug is not adequately explained for non-specialist audiences. These communication gaps prevent readers from understanding both the "so what?" and "now what?" of this research.

Major Concerns

1. Primary Finding Presented Without Interpretable Context — Abstract, Results

Problem: The core finding—"0.21 more children (95% CI: 0.08–0.35)"—is reported without baseline reference, making it impossible for readers to assess practical magnitude. Is this a 2% increase or a 20% increase? Even with a narrow confidence interval demonstrating precision, the clinical significance remains opaque. A statistically significant effect of unknown practical importance fails to provide actionable information for clinicians or policymakers.

Recommendation: Provide immediate contextualization in the abstract: 1. State the baseline: "UK men in this cohort typically father 1.9 children on average" 2. Calculate relative effect: "representing approximately an 11% increase" 3. Interpret practical significance: "This effect size, if causal, suggests a clinically meaningful impact on male fertility that warrants further investigation in clinical trials."

Add similar contextualization when presenting this finding in Results and Discussion. Consider including a "back-of-envelope" calculation: "In a population of 1,000 men, this would translate to approximately 210 additional children conceived."

2. Insufficient Justification for Mendelian Randomization Approach — Introduction

Problem: The introduction states that "traditional observational studies are undermined by residual confounding" and that MR "strengthens causal inference," but fails to explain WHY genetic proxies are needed for a drug that is: (a) widely prescribed, (b) taken for acute episodes rather than chronically, (c) has well-characterized pharmacology, and (d) could be studied in randomized trials or straightforward observational designs. The "obvious question" for clinicians is never answered: Why use genetics when we could just compare men who take sildenafil to those who don't? What specific confounders are you avoiding?

Recommendation: Add 2-3 sentences explaining the specific inferential advantage: - "While sildenafil is prescribed for erectile dysfunction, men with ED differ systematically from men without ED in ways difficult to measure (relationship quality, comorbidities, lifestyle factors). These unmeasured differences confound traditional comparisons of users vs. non-users." - "Mendelian

randomization uses genetic variants associated with drug response to create 'nature's randomized trial,' providing estimates free from confounding by indication." - "This approach allows us to estimate the causal effect of sildenafil independent of the underlying conditions for which it is prescribed."

Without this explanation, sophisticated readers may question whether MR is methodological overkill, while clinical readers will be confused about the study design's purpose.

Minor Issues

- **Abstract accessibility**: The term "genetically proxied sildenafil" appears without definition. Consider: "Using genetic variants that mimic sildenafil's mechanism of action (genetically proxied sildenafil)..."
- **Statistical reporting precision**: Report confidence intervals with same decimal precision as point estimate for consistency
- **Target audience ambiguity**: Unclear whether this is written for geneticists, epidemiologists, or clinicians. Consider adding a sentence in the Introduction stating: "This study is designed to inform both reproductive medicine clinicians and researchers investigating pharmacological approaches to male fertility."

Strengths

- **Methodological rigor**: Mendelian randomization is an appropriate design for addressing confounding by indication in this context (once the rationale is clarified)
- **Statistical precision**: Narrow confidence interval (0.08–0.35) provides good precision around the effect estimate, ruling out both null effects and implausibly large effects
- **Novel research question**: Investigating sildenafil's fertility effects (beyond erectile function) addresses an under-studied area with potential clinical importance
- **Transparent reporting**: FDR correction for multiple testing is appropriate and clearly reported

Questions for Authors

1. What is the baseline fertility rate (children fathered) in the comparison group? This is essential for interpreting the 0.21 effect.
2. Have you considered including a "plain language summary" box that translates the genetic proxy approach into clinical terms?
3. Could you provide absolute effect estimates alongside relative effects to aid clinical interpretation?
4. What would be the clinical recommendation for practitioners reading this paper? Should they consider prescribing sildenafil for fertility concerns, or is this purely hypothesis-generating for future RCTs?

Recommendation

Minor Revision

Justification: This manuscript presents methodologically sound work with a potentially important finding, but communication barriers prevent non-specialist readers from understanding its significance and applicability. The two major concerns are entirely addressable through improved contextualization (Issue 1) and clearer explanation of the study rationale (Issue 2). Neither issue threatens the validity of the findings—they only limit comprehensibility. Once revised, this work could meaningfully inform clinical discussions about sildenafil's broader effects beyond erectile dysfunction. The narrow confidence intervals and novel research question represent valuable contributions that deserve publication after these accessibility improvements.

Statistical Methods Agent

Model: anthropic/claude-opus-4-5-20251101

Statistical Methods Review

Discipline: Genetic Epidemiology / Pharmacoepidemiology

Study Design: Two-sample cis-Mendelian Randomization (Drug-target MR)

Statistical Methods Identified: - Two-sample Mendelian randomization - cis-MR (drug-target MR) - Inverse-variance weighted meta-analysis with random effects - Bayesian colocalization (Coloc, HyPrColoc) - LD Check sensitivity analysis - Two-step cis-MR for pleiotropy adjustment - Benjamini-Hochberg FDR correction - F-statistic for instrument strength - BOLT-LMMfor GWAS - MR Same Population Test - Bootstrap standard errors

Overall Quality: Acceptable

Summary

This manuscript presents a two-sample cis-Mendelian randomization (MR) study investigating the association between genetically proxied PDE5 inhibition (sildenafil) and fertility, sexual activity, and wellbeing outcomes. The study leverages genetic variants in the PDE5A gene region to proxy drug target perturbation, using blood pressure as a biomarker for PDE5 inhibition.

Overall Assessment: The manuscript demonstrates methodological sophistication in applying drug-target MR methods, including appropriate instrument validation (HyPrColoc, LD Check), positive control analyses, multiple testing correction (Benjamini-Hochberg), and sensitivity analyses (Two-step cis-MR, colocalization). However, several statistical issues require attention, ranging from incomplete reporting of effect sizes and confidence intervals to concerns about sample overlap bias, power calculations, and the interpretation of null colocalization findings.

Key Strengths: - Well-designed cis-MR framework with appropriate instrument selection - Multiple sensitivity analyses addressing key MR assumptions - Application of FDR correction for multiple testing - Positive control analyses validating the instrument - Sex-stratified replication analysis

Key Concerns: 1. Substantial sample overlap (~60%) between exposure and outcome GWASs may bias results away from the null despite adequate F-statistics 2. Low power in colocalization analyses raises concerns about false positive findings 3. Incomplete reporting of heterogeneity statistics for the MR meta-analysis 4. Missing power calculations for primary and secondary outcomes 5. Interpretation of wide confidence intervals in female analyses requires more cautious language 6. The scaling of MR estimates to 100mg sildenafil requires additional justification regarding linearity assumptions

Statistical Issues (10 found)

STAT-001: Sample Size (Major)

Location: Methods section, Page 11-12; Results section, Page 14-15

The manuscript acknowledges substantial sample overlap (~60%) between the exposure GWAS

(ICBP, which includes UK Biobank) and the outcome GWASs (UK Biobank). While the authors correctly note that weak instrument bias typically biases toward the null with no overlap, sample overlap in two-sample MR can actually bias results AWAY from the null when the exposure and outcome are measured in the same individuals. With F-statistics of 25-32, the bias from sample overlap may not be negligible. The formula for bias from sample overlap is approximately: bias "H (overlap proportion) × (1/F-statistic) × (confounding bias). With 60% overlap and F~30, this represents a potential ~2% inflation of estimates, which could be meaningful for marginal findings.

Evidence: "Since around two-thirds of the participants in the Evangelou et al. GWAS were from the UKB we expect there to be substantial sample overlap between our exposure and outcome samples... However, given the F-statistics of 23 for SBP and 18 for DBP, we would expect the risk of weak instrument bias to be approximately 4% and 6%, respectively."

Recommendation: 1) Conduct a sensitivity analysis using non-overlapping samples where possible. Consider using the ICBP data excluding UK Biobank participants for the exposure GWAS, or use an alternative blood pressure GWAS without UK Biobank. 2) Apply the Burgess et al. (2016) formula to quantify the maximum potential bias from sample overlap. 3) Report the exact percentage of sample overlap and discuss its implications more thoroughly. 4) Consider implementing the MR-RAPS method which is more robust to weak instruments and sample overlap.

Code Examples:

R (packages: mr.raps):

```
# Quantify potential bias from sample overlap
# Burgess et al. 2016 Genetic Epidemiology formula

overlap_proportion <- 0.60
F_statistic_DBP <- 31.58
F_statistic_SBP <- 25.53

# Maximum relative bias (as proportion of true effect)
max_relative_bias_DBP <- overlap_proportion / F_statistic_DBP
max_relative_bias_SBP <- overlap_proportion / F_statistic_SBP

cat("Maximum relative bias (DBP instrument):",
    round(max_relative_bias_DBP * 100, 2), "%\n")
cat("Maximum relative bias (SBP instrument):",
    round(max_relative_bias_SBP * 100, 2), "%\n")

# For more robust analysis with sample overlap, consider MR-RAPS
library(mr.raps)

# Example implementation
# mr_raps_result <- mr.raps(
#   b_exp = beta_exposure,
#   b_out = beta_outcome,
#   se_exp = se_exposure,
#   se_out = se_outcome,
#   over.dispersion = TRUE,
#   loss.function = "huber"
# )
```

MR-RAPS provides more robust inference under weak instruments and potential sample overlap

Example Write-up: > We acknowledge substantial sample overlap (~60%) between the exposure GWAS (ICBP) and outcome GWASs (UK Biobank). To quantify potential bias, we applied the Burgess et al. (2016) formula: maximum bias "Hoverlap × (1/F) × Ã _ Y / Ã W̄X. With F-statistics of 25.5-31.6 and 60% overlap, the maximum bias inflation is approximately 2-2.4%. As a sensitivity analysis, we repeated the primary analysis using [non-overlapping sample/alternative GWAS],

which yielded consistent results (²= X.XX, 95% CI: X.XX-X.XX).

Literature Support: Burgess S, Davies NM, Thompson SG. Bias due to participant overlap in two-sample Mendelian randomization. Genetic Epidemiology. 2016;40(7):597-608. Zhao Q et al. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. Annals of Statistics. 2020;48(3):1742-1769.

STAT-002: Statistical Reporting (Major)

Location: Results section, Page 14-16; Table 2

The manuscript does not report heterogeneity statistics (I^2, Q-statistic, tau^2) for the random-effects meta-analysis of Wald ratios across multiple genetic variants. When using multiple correlated variants in MR, heterogeneity in causal estimates across variants can indicate pleiotropy or other violations of MR assumptions. The absence of these statistics makes it difficult to assess the consistency of causal estimates and the validity of the MR assumptions.

Evidence: "These Wald estimates were then meta-analysed with a multiplicative random effects model while using the LD matrix to account for the correlation between variants."

Recommendation: Report heterogeneity statistics for all MR analyses: 1) Cochran's Q statistic with p-value, 2) I^2 statistic with interpretation (low <25%, moderate 25-75%, high >75%), 3) Between-variant variance (tau^2). If heterogeneity is high, discuss potential sources and consider additional sensitivity analyses (e.g., leave-one-out analysis, MR-PRESSO for outlier detection).

Code Examples:

R (packages: TwoSampleMR, MendelianRandomization):

```
library(TwoSampleMR)
library(MendelianRandomization)

# After running MR analysis, extract heterogeneity statistics
# Using MendelianRandomization package for correlated variants

# Example with correlated variants
mr_result <- mr_ivw_correl(
  mr_input(
    bx = beta_exposure,
    bxse = se_exposure,
    by = beta_outcome,
    byse = se_outcome,
    correlation = ld_matrix
  )
)

# Extract heterogeneity
cat("Q statistic:", mr_result@Heter.Stat[1], "\n")
cat("Q p-value:", mr_result@Heter.Stat[2], "\n")

# Calculate I-squared
Q <- mr_result@Heter.Stat[1]
df <- length(beta_exposure) - 1
I_squared <- max(0, (Q - df) / Q * 100)
cat("I-squared:", round(I_squared, 1), "%\n")

# For leave-one-out analysis
loo_results <- mr_leaveoneout(dat)
plot(loo_results)
```

The MendelianRandomization package handles correlated variants and provides heterogeneity statistics

Example Write-up: > Wald ratio estimates were pooled using inverse-variance weighted random-effects meta-analysis, accounting for LD between variants. For the association with number of children fathered, there was low heterogeneity across the five genetic variants (I^2 = 15.2%, Q = 4.71, p = 0.32, Ä ≐ 0.002), supporting consistency of causal estimates. Heterogeneity statistics for all outcomes are presented in Supplementary Table X.

Literature Support: Bowden J, et al. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression. International Journal of Epidemiology. 2015;44(6):1961-1974. Greco MF, et al. Detecting pleiotropy in Mendelian randomisation studies with summary data. Statistics in Medicine. 2015;34(21):2926-2940.

STAT-003: P-Value Interpretation (Major)

Location: Results section, Page 15-16; Discussion section, Page 19-21

The colocalization analysis for the primary finding (number of children fathered) shows H1 as the most likely hypothesis (91% posterior probability), indicating evidence only for a causal variant affecting blood pressure but not the outcome. The posterior probability for H4 (shared causal variant) is only 0.7%, which is far below the conventional threshold of 80% for supporting colocalization. While the authors acknowledge this and attribute it to low power, they proceed to interpret the MR finding as causal evidence. This represents a significant interpretive concern - if colocalization fails, the MR estimate may be confounded by LD.

Evidence: "Although the posterior probability of H4 was 3.5 times the posterior probability of H3,

H1 was the most likely hypothesis (91%) in our colocalization analysis (Table 2), implying suboptimal power in the colocalisation... Despite the absence of supportive evidence from coloc, we believe it is unlikely that our results are attributable to confounding by LD."

Recommendation: 1) Acknowledge more prominently that the primary finding lacks colocalization support and should be interpreted with caution. 2) Consider using colocalization methods designed for low-power settings (e.g., coloc with adjusted priors, SuSiE-coloc for fine-mapping). 3) Report power calculations for the colocalization analysis. 4) Frame the finding as 'suggestive evidence requiring replication' rather than 'genetic support for a causal effect'. 5) Consider whether the outcome GWAS has sufficient power to detect the expected effect size in the gene region.

Code Examples:

R (packages: coloc, susieR):

```
library(coloc)
library(susieR)  # For SuSiE-coloc

# Standard coloc analysis
coloc_result <- coloc.abf(
  dataset1 = list(beta = beta_bp, varbeta = se_bp^2,
                  N = n_bp, type = "quant"),
  dataset2 = list(beta = beta_outcome, varbeta = se_outcome^2,
                  N = n_outcome, type = "quant"),
  MAF = maf
)

# Report all posterior probabilities
cat("PP.H0 (no association):", round(coloc_result$summary["PP.H0.abf"], 3), "\n")
cat("PP.H1 (trait 1 only):", round(coloc_result$summary["PP.H1.abf"], 3), "\n")
cat("PP.H2 (trait 2 only):", round(coloc_result$summary["PP.H2.abf"], 3), "\n")
cat("PP.H3 (both, different variants):", round(coloc_result$summary["PP.H3.abf"], 3),
"\n")
cat("PP.H4 (shared variant):", round(coloc_result$summary["PP.H4.abf"], 3), "\n")

# For low-power settings, consider adjusting priors
# or using SuSiE-coloc for better fine-mapping
# coloc_susie <- coloc.susie(susie_result1, susie_result2)

# Power calculation for coloc
# Approximate power depends on sample size and effect size
power_coloc <- function(n1, n2, h2_1, h2_2, n_snps) {
  # Simplified power approximation
  # Full implementation in Wallace (2020)
  ncp1 <- n1 * h2_1 / n_snps
  ncp2 <- n2 * h2_2 / n_snps
  power <- pchisq(qchisq(0.05, 1, lower.tail = FALSE),
                  1, ncp = min(ncp1, ncp2), lower.tail = FALSE)
  return(power)
}
```

SuSiE-coloc provides better power for fine-mapping in low-power settings

Example Write-up: > Bayesian colocalization analysis did not support a shared causal variant between blood pressure and number of children fathered (PP_H4 = 0.7%, PP_H1 = 91.0%). This likely reflects insufficient power in the outcome GWAS rather than confounding by LD, as evidenced by: (1) PP_H4/PP_H3 ratio of 3.5, (2) supportive LD Check analysis (68% of top SNPs in LD with lead variant), and (3) similar peak patterns in locus plots. However, given the lack of formal colocalization support, we interpret this finding as suggestive evidence that requires replication in larger outcome GWASs or alternative study designs.

Literature Support: Wallace C. A more accurate method for colocalisation analysis allowing for

multiple causal variants. PLoS Genetics. 2021;17(9):e1009440. Giambartolomei C, et al. Bayesian test for colocalisation between pairs of genetic association studies. PLoS Genetics. 2014;10(5):e1004383.

STAT-004: Sample Size (Major)

Location: Methods section; entire manuscript

The manuscript does not report formal power calculations for either the MR analyses or the colocalization analyses. Given that several outcomes show null results (sexual partners, virginity, wellbeing) and the primary finding has weak colocalization support, it is important to understand whether these are true null effects or reflect insufficient power. Power in MR depends on the proportion of variance explained by the instruments (R^2), sample size, and true effect size.

Evidence: "No power calculations are presented in the manuscript. The authors state 'Since these GWASs had a similar sample size to the male-only GWASs, they should have similar statistical power' without formal justification."

Recommendation: 1) Report power calculations for the primary MR analyses using the mRnd tool or analytical formulas. 2) For null findings, report the minimum detectable effect size (MDES) at 80% power. 3) For colocalization, discuss the expected power given outcome GWAS sample sizes. 4) Consider using the Brion et al. (2013) or Burgess (2014) power formulas for MR.

Code Examples:

R (packages: base):

```
# Power calculation for two-sample MR
# Using analytical formula from Burgess (2014)

mr_power <- function(n_outcome, r2_instrument, alpha = 0.05,
                     true_effect, residual_var = 1) {
  # n_outcome: sample size of outcome GWAS
  # r2_instrument: variance explained by instruments in exposure
  # true_effect: hypothesized causal effect
  # residual_var: residual variance of outcome

  # Non-centrality parameter
  ncp <- (true_effect^2 * r2_instrument * n_outcome) / residual_var

  # Critical value
  crit <- qchisq(1 - alpha, df = 1)

  # Power
  power <- pchisq(crit, df = 1, ncp = ncp, lower.tail = FALSE)

  return(power)
}

# Calculate power for primary outcome
n_children <- 209872
 r 2 _ D B P   < -   0 . 0 0 0 2    #   A p p r o x i m a t e   R ²   f r o m   F - s t a t :   R

# Power curve
effect_sizes <- seq(0.05, 0.5, by = 0.05)
powers <- sapply(effect_sizes, function(e) {
  mr_power(n_children, r2_DBP, true_effect = e)
})

plot(effect_sizes, powers, type = "b",
     xlab = "True Effect Size (children per mmHg DBP)",
     ylab = "Power",
     main = "MR Power Analysis")
abline(h = 0.8, lty = 2, col = "red")

# Minimum detectable effect at 80% power
mdes <- effect_sizes[which.min(abs(powers - 0.8))]
cat("Minimum detectable effect at 80% power:", mdes, "\n")

# Alternative: use mRnd online tool
# https://shiny.cnsgenomics.com/mRnd/
```

The mRnd web tool (https://shiny.cnsgenomics.com/mRnd/) provides a user-friendly interface for MR power calculations

Example Write-up: > Power calculations were performed using the mRnd web tool (Brion et al., 2013). For the primary outcome (number of children fathered, N = 209,872), with R^2 = 0.02% explained by our instruments and ±= 0.05, we had 80% power to detect an effect of ²"e0.15 children per unit decrease in DBP. For outcomes showing null results, the minimum detectable effect sizes at 80% power were: sexual partners (²= 2.1), wellbeing (SMD = 0.08). The observed null findings are therefore consistent with either no true effect or effects smaller than our detectable threshold.

Literature Support: Brion MJ, Shakhbazov K, Visscher PM. Calculating statistical power in Mendelian randomization studies. International Journal of Epidemiology. 2013;42(5):1497-1501. Burgess S. Sample size and power calculations in Mendelian randomization with a single instrumental variable and a binary outcome. International Journal of Epidemiology. 2014;43(3):922-929.

STAT-005: Effect Size (Major)

Location: Methods section, Page 12; Results section, Page 14-15

The MR estimates are scaled to represent the effect of 100mg daily sildenafil, based on the assumption that sildenafil reduces DBP by 5.5 mmHg and SBP by 8.4 mmHg. However, this scaling assumes a linear dose-response relationship and that genetic proxies of PDE5 inhibition have equivalent effects to pharmacological inhibition. The manuscript does not adequately justify these assumptions or provide sensitivity analyses around the scaling factor.

Evidence: "At a dose of 100mg, sildenafil results in up to an 5.5 mmHg and 8.4 mmHg decrease in DBP and SBP respectively. To facilitate the interpretation of our results, we scaled our MR estimates to represent the blood pressure-lowering effect of a 100mg daily dose of sildenafil."

Recommendation: 1) Report both the unscaled (per mmHg) and scaled (per 100mg sildenafil) estimates. 2) Acknowledge the linearity assumption explicitly and discuss its plausibility. 3) Provide sensitivity analyses using different scaling factors (e.g., based on different doses or different studies of sildenafil's BP effects). 4) Discuss the difference between lifetime genetic exposure and intermittent pharmacological exposure.

Code Examples:

R (packages: base):

```
# Sensitivity analysis for different scaling factors

# Original estimate (per mmHg DBP)
beta_per_mmHg <- 0.050  # example value
se_per_mmHg <- 0.007

# Different scaling scenarios based on sildenafil dose
scaling_factors <- data.frame(
  dose = c("25mg", "50mg", "100mg", "100mg (upper)"),
  dbp_reduction = c(1.5, 3.0, 5.5, 8.0)  # mmHg
)

# Calculate scaled effects
scaling_factors$beta_scaled <- beta_per_mmHg * scaling_factors$dbp_reduction
scaling_factors$se_scaled <- se_per_mmHg * scaling_factors$dbp_reduction
scaling_factors$ci_lower <- scaling_factors$beta_scaled - 1.96 *
scaling_factors$se_scaled
scaling_factors$ci_upper <- scaling_factors$beta_scaled + 1.96 *
scaling_factors$se_scaled

print(scaling_factors)

# Report both scaled and unscaled
cat("\nUnscaled estimate (per 1 mmHg DBP reduction):\n")
 c a t ( s p r i n t f ( "     ²   =   % . 3 f   ( 9 5 % %   C I :   % . 3 f   t o   % . 3 f ) \ n
            beta_per_mmHg,
            beta_per_mmHg - 1.96*se_per_mmHg,
            beta_per_mmHg + 1.96*se_per_mmHg))

cat("\nScaled estimate (per 100mg sildenafil, 5.5 mmHg DBP):\n")
 c a t ( s p r i n t f ( "     ²   =   % . 3 f   ( 9 5 % %   C I :   % . 3 f   t o   % . 3 f ) \ n
            scaling_factors$beta_scaled[3],
            scaling_factors$ci_lower[3],
            scaling_factors$ci_upper[3]))
```

Always report unscaled estimates alongside scaled estimates for transparency

Example Write-up: > MR estimates were scaled to approximate the effect of 100mg daily sildenafil, assuming linear dose-response. Per 1 mmHg decrease in DBP, genetically proxied PDE5 inhibition was associated with 0.050 (95% CI: 0.036-0.065) additional children. Scaling to the 5.5 mmHg DBP reduction observed with 100mg sildenafil yields 0.28 (95% CI: 0.20-0.36) additional children. We note that this scaling assumes: (1) linearity between genetic and pharmacological effects, (2) equivalent biological mechanisms, and (3) that acute BP effects translate to chronic exposure effects. Sensitivity analyses using alternative scaling factors (50mg: 3.0 mmHg; 25mg: 1.5 mmHg) yielded effects of 0.15 (95% CI: 0.11-0.19) and 0.08 (95% CI: 0.05-0.10) children, respectively.

Literature Support: Gill D, et al. Mendelian randomization for studying the effects of perturbing drug targets. Wellcome Open Research. 2021;6:16. Schmidt AF, et al. Genetic drug target validation using Mendelian randomisation. Nature Communications. 2020;11:3255.

STAT-006: Confidence Intervals (Minor)

Location: Results section, Page 16-17; Discussion section, Page 18-19

The female analysis for number of live births shows a confidence interval that spans from negative to positive values (²= -0.14, 95% CI: -0.28 to 0.01), yet the authors state this 'suggests that our results linking PDE5 inhibition to increased fertility are specific to men.' The wide CI and near-null result in women does not provide strong evidence of sex-specificity; rather, it indicates uncertainty. A formal test for sex interaction would be needed to claim sex-specificity.

Evidence: "We did not find evidence of an association between genetically proxied sildenafil and the number of live births in women (²= 0.14, 95% CI: -0.28 to 0.01), although the estimate was in the opposite direction to that in men... Our sensitivity cis-MR analyses did not find an association between genetically proxied PDE5 inhibition and the number of live births for UKB participants, suggesting that our results linking PDE5 inhibition to increased fertility are specific to men."

Recommendation: 1) Conduct a formal test for sex interaction by comparing the male and female effect estimates (z-test for difference in coefficients). 2) Report the p-value for the sex interaction. 3) If the interaction is not statistically significant, revise the language to acknowledge that the evidence for sex-specificity is suggestive rather than conclusive. 4) Consider meta-analyzing male and female estimates to obtain an overall effect.

Code Examples:

R (packages: base):

```
# Test for sex interaction (difference in coefficients)

# Male estimates
beta_male <- 0.276
se_male <- (0.355 - 0.196) / (2 * 1.96)  # Derive SE from CI

# Female estimates
beta_female <- -0.136
se_female <- (0.005 - (-0.277)) / (2 * 1.96)  # Derive SE from CI

# Z-test for difference
z_diff <- (beta_male - beta_female) / sqrt(se_male^2 + se_female^2)
p_interaction <- 2 * pnorm(-abs(z_diff))

cat("Male effect:", round(beta_male, 3), "(SE:", round(se_male, 3), ")\n")
cat("Female effect:", round(beta_female, 3), "(SE:", round(se_female, 3), ")\n")
cat("Z-statistic for difference:", round(z_diff, 2), "\n")
cat("P-value for sex interaction:", format.pval(p_interaction, digits = 3), "\n")

# Interpretation
if (p_interaction < 0.05) {
  cat("\nConclusion: Significant sex difference (p < 0.05)\n")
} else {
  cat("\nConclusion: No significant sex difference; ",
      "sex-specificity claim not statistically supported\n")
}
```

A significant interaction test is required to claim sex-specific effects

Example Write-up: > To assess sex-specificity, we compared male and female effect estimates using a z-test for the difference in coefficients. The effect in men (²= 0.28, SE = 0.04) differed from that in women (²= -0.14, SE = 0.07), with a z-statistic of 5.2 (p < 0.001), providing evidence for sex-specific effects. However, given the wide confidence interval in women and the observational nature of MR, we interpret this as suggestive evidence of sex-specificity requiring replication.

Literature Support: Altman DG, Bland JM. Interaction revisited: the difference between two estimates. BMJ. 2003;326(7382):219. Richardson TG, et al. Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis. PLoS Medicine. 2020;17(3):e1003062.

STAT-007: Multiple Testing (Minor)

Location: Methods section, Page 12; Results section, Page 14-17

While the authors appropriately apply Benjamini-Hochberg FDR correction for the five primary outcomes, the total number of tests conducted is larger when including: (1) primary analyses with DBP, (2) secondary analyses with SBP, (3) positive control analyses, (4) female replication analyses, and (5) two-step cis-MR sensitivity analyses. The multiplicity correction strategy across all these analyses is not clearly specified.

Evidence: "The Benjamini and Hochberg correction was used to account for multiple testing... Our results suggest that PDE5 inhibition is associated with fathering 0.28 more children [95% CI: 0.20–0.36, pfdr < 0.001]"

Recommendation: 1) Clarify the scope of the FDR correction (which analyses were included in the correction). 2) Consider a hierarchical testing approach where primary analyses are tested first, and secondary/sensitivity analyses are interpreted descriptively. 3) Report both unadjusted and FDR-adjusted p-values for transparency. 4) Explicitly state that sensitivity analyses (two-step

cis-MR, female replication) were not included in the multiplicity correction and should be interpreted as exploratory.

Code Examples:

R (packages: base):

```
# Clarify FDR correction scope

# Primary outcomes (5 tests)
primary_p_values <- c(
  children = 0.001,
  age_first_sex = 0.015,
  sexual_partners = 0.955,
  virgin = 0.837,
  wellbeing = 0.837
)

# Apply BH correction to primary outcomes only
fdr_primary <- p.adjust(primary_p_values, method = "BH")

results_table <- data.frame(
  Outcome = names(primary_p_values),
  P_unadjusted = primary_p_values,
  P_FDR = fdr_primary,
  Significant_FDR = fdr_primary < 0.05
)

print(results_table)

# Note: Secondary/sensitivity analyses interpreted descriptively
cat("\nNote: Secondary analyses (SBP instrument, female replication,\n",
    "two-step cis-MR) are exploratory and not included in FDR correction.\n")
```

Clearly document which analyses are included in multiplicity correction

Example Write-up: > We applied Benjamini-Hochberg FDR correction across the five primary outcomes (number of children, age first sex, sexual partners, virginity, wellbeing) using the DBP instrument, controlling FDR at 5%. Secondary analyses using SBP instruments and sensitivity analyses (two-step cis-MR, female replication) were not included in the multiplicity correction and are interpreted as exploratory. Both unadjusted and FDR-adjusted p-values are reported in Table 2.

Literature Support: Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B. 1995;57(1):289-300.

STAT-008: Statistical Reporting (Minor)

Location: Results section, Page 14-17; Tables

Several statistical results are incompletely reported. Specifically: (1) Exact p-values are not consistently reported (some show 'p < 0.001' rather than exact values); (2) Degrees of freedom are not reported for the meta-analysis; (3) The number of variants contributing to each analysis is not always clear in the results text; (4) Standard errors are not reported alongside confidence intervals in all tables.

Evidence: "Our results suggest that PDE5 inhibition is associated with fathering 0.28 more children [95% CI: 0.20–0.36, pfdr < 0.001]... Our positive control analyses yielded an MR association in the

expected direction between PDE5 inhibition and erectile dysfunction (p < 0.001 and p = 0.049 for PDE5A-driven DBP and PDE5A-driven SBP respectively)"

Recommendation: 1) Report exact p-values to at least 3 significant figures (e.g., p = 2.3 × 10⁻⁵ rather than p < 0.001). 2) Include standard errors in all tables alongside confidence intervals. 3) Report the number of genetic variants (k) used in each analysis. 4) For meta-analyses, report degrees of freedom. 5) Consider using a standardized reporting format such as the STROBE-MR checklist.

Code Examples:

R (packages: base):

```
# Function for complete statistical reporting
report_mr_result <- function(beta, se, ci_lower, ci_upper, p_value,
                             p_fdr, n_variants, f_stat) {

  # Format p-value properly
  format_p <- function(p) {
    if (p < 0.001) {
      return(sprintf("%.2e", p))
    } else {
      return(sprintf("%.3f", p))
    }
  }

  result <- sprintf(
        "β = %.3f (SE = %.3f, 95%% CI: %.3f to %.3f,
    beta, se, ci_lower, ci_upper, format_p(p_value), format_p(p_fdr)
  )

  result <- paste0(result, sprintf(
    "Based on k = %d genetic variants, mean F-statistic = %.1f\n",
    n_variants, f_stat
  ))

  return(cat(result))
}

# Example usage
report_mr_result(
  beta = 0.276,
  se = 0.041,
  ci_lower = 0.196,
  ci_upper = 0.355,
  p_value = 0.00001,
  p_fdr = 0.0001,
  n_variants = 5,
  f_stat = 31.58
)
```

Use consistent formatting for all statistical results

Example Write-up: > Using 5 genetic variants (k = 5), genetically proxied PDE5 inhibition (instrumented by DBP) was associated with fathering 0.28 additional children (β = 0.28, SE = 0.04, 95% CI: 0.20-0.36, p = 2.1 × 10⁻⁵, FDR = 1.0 × 10⁻⁴). The analysis had an average F-statistic of 31.6, indicating adequate instrument strength.

Literature Support: Skrivankova VW, et al. Strengthening the Reporting of Observational Studies in Epidemiology Using Mendelian Randomization: The STROBE-MR Statement. JAMA. 2021;326(16):1614-1621.

STAT-009: Model Assumptions (Minor)

Location: Methods section, Page 11-12; Supplementary Methods

The manuscript uses an LD threshold of $r^2 < 0.35$ for clumping variants, which is relatively liberal compared to the more common threshold of $r^2 < 0.1$ or $r^2 < 0.01$ used in MR studies. While the authors account for LD in the meta-analysis using the LD matrix, residual correlation between variants can still inflate precision and affect heterogeneity estimates. The choice of $r^2 < 0.35$ is not justified.

Evidence: "Variants were then clumped with a LD threshold of $r^2 < 0.35$ and a distance threshold of 10,000 kilobases (kB)."

Recommendation: 1) Justify the choice of $r^2 < 0.35$ threshold (e.g., to retain more variants in a gene region with limited independent signals). 2) Conduct a sensitivity analysis using more stringent LD thresholds ($r^2 < 0.1$, $r^2 < 0.01$) to assess robustness. 3) Report the effective number of independent variants after accounting for LD. 4) Verify that the LD matrix used in the analysis accurately reflects the correlation structure.

Code Examples:

R (packages: TwoSampleMR, ieugwasr):

```
library(TwoSampleMR)
library(ieugwasr)

# Sensitivity analysis with different LD thresholds
ld_thresholds <- c(0.01, 0.1, 0.35, 0.5)

results_by_ld <- lapply(ld_thresholds, function(r2_thresh) {

  # Clump at this threshold
  clumped <- ld_clump(
    dat = exposure_dat,
    clump_r2 = r2_thresh,
    clump_kb = 10000,
    pop = "EUR"
  )

  # Run MR with remaining variants
  # (would need to harmonize and run MR)

  return(list(
    r2_threshold = r2_thresh,
    n_variants = nrow(clumped)
  ))
})

# Compare results
cat("Sensitivity analysis across LD thresholds:\n")
for (res in results_by_ld) {
  cat(sprintf("r² < %.2f: %d variants retained\n",
              res$r2_threshold, res$n_variants))
}
```

More stringent LD thresholds may reduce power but improve independence assumptions

Example Write-up: > Variants were clumped at $r^2 < 0.35$ to retain multiple signals within the PDE5A gene region while accounting for residual LD in the meta-analysis using the 1000 Genomes

European reference panel. This threshold was chosen to balance instrument strength with independence, given the limited number of genome-wide significant variants in the region. Sensitivity analyses using more stringent thresholds ($r^2 < 0.1$) yielded consistent results (²= 0.25, 95% CI: 0.15-0.35, k = 3 variants).

Literature Support: Burgess S, Zuber V, et al. Mendelian randomization with fine-mapped genetic data: Choosing from large numbers of correlated instrumental variables. Genetic Epidemiology. 2017;41(8):714-725.

STAT-010: Causality Claims (Minor)

Location: Abstract, Key Messages, Discussion, Conclusions

The manuscript uses causal language ('causal association', 'causal effects', 'genetic support for a beneficial effect') despite several limitations: (1) failed formal colocalization for the primary finding, (2) potential for horizontal pleiotropy not fully ruled out, (3) substantial sample overlap, and (4) the inherent limitations of MR as a quasi-experimental method. While MR provides stronger causal evidence than observational studies, the language should be appropriately hedged.

Evidence: "We find evidence for a casual association between genetically proxied sildenafil use and number of children fathered... This study provides genetic support for PDE5 inhibitors increasing the number of children that men have... We find genetic evidence to support an effect of PDE5 inhibition on men having more children."

Recommendation: 1) Replace 'causal association' with 'genetically predicted association' or 'MR-estimated effect'. 2) Add appropriate caveats when discussing causality (e.g., 'consistent with a causal effect, although residual confounding by LD cannot be fully excluded'). 3) Emphasize that MR provides 'genetic evidence consistent with' rather than 'proof of' causality. 4) Acknowledge the failed colocalization more prominently in conclusions.

Example Write-up: > Our MR analysis provides genetic evidence consistent with an effect of PDE5 inhibition on male fertility. However, given the absence of formal colocalization support and the inherent limitations of MR methodology, these findings should be interpreted as hypothesis-generating rather than confirmatory. Replication in independent samples and triangulation with other study designs are needed before causal conclusions can be drawn.

Literature Support: Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Human Molecular Genetics. 2014;23(R1):R89-R98. Lawlor DA, et al. Triangulation in aetiological epidemiology. International Journal of Epidemiology. 2016;45(6):1866-1886.

Results Accuracy Verification

Discipline: Molecular Biology / Genetic Epidemiology (Mendelian Randomization)

Tables Reviewed: 3 **Figures Reviewed:** 4

Overall Assessment: Poor

Summary

This Mendelian randomization study investigating PDE5 inhibition effects on fertility and sexual outcomes contains several significant accuracy issues requiring attention. The most critical finding is an impossible confidence interval (168 to 1.000) for age of first sex. Multiple major inconsistencies exist between the Abstract and Results sections, including discrepant point estimates (0.21 vs 0.28) and p-values (0.01 vs <0.001) for the primary outcome. Several outcomes (sexual partners, odds of being virgin, wellbeing) have different estimates reported in different locations without clear labeling of which instrument (DBP vs SBP) each corresponds to. Table content was unavailable for verification, limiting the ability to confirm text-table concordance. Presentation issues include inconsistent decimal precision for p-values and variable notation. The manuscript requires careful revision to ensure internal consistency before publication.

Critical Issues

ACC-003 [statistical_plausibility] - Location: Results section, 'Age of first having sex' paragraph - The confidence interval for age of first having sex appears implausible. The reported CI (168 to 1.000) has a lower bound of 168 years which is impossible for age of first sex. This appears to be a typographical error. - Recommendation: Correct the CI lower bound - likely should be 0.168 rather than 168. Verify against original analysis output.

Major Issues

ACC-001 [text_table_mismatch] - Location: Results section, 'Number of children fathered' paragraph vs Abstract - The effect estimate for number of children fathered differs between the Results section and the Abstract. The Results section reports 0.28 more children while the Abstract reports 0.21 more children for the same outcome. - Recommendation: Verify which estimate is correct and ensure consistency between Abstract and Results sections. The confidence intervals also differ substantially, suggesting these may be from different analyses.

ACC-002 [internal_consistency] - Location: Results section, 'Number of children fathered' paragraph - The p-value reported in the Results section (p_fdr < 0.001) differs from the p-value in the Abstract (p = 0.01) for the same primary finding regarding number of children fathered. - Recommendation: Clarify whether these represent different analyses (e.g., DBP vs SBP instrumented trait) and ensure the primary result is consistently reported throughout.

ACC-004 [internal_consistency] - Location: Results section, p-value precision inconsistency - P-values are reported with inconsistent decimal precision throughout the Results section. Some are reported to 3 decimal places (p_fdr = 0.015, 0.933, 0.955, 0.837) while others use 2 decimal places (p_fdr = 0.01, 0.93, 0.87, 0.50). - Recommendation: Standardize p-value reporting to

consistent decimal places (recommend 3 decimal places or use < 0.001 notation).

ACC-005 [internal_consistency] - Location: Results section, 'Age of first having sex' paragraph - The same outcome (age of first having sex with SBP instrument) appears to have different p-values reported: p_fdr = 0.933 in one location and p_fdr = 0.93 in another, suggesting rounding inconsistency or different analyses. - Recommendation: Verify these refer to the same analysis and report consistently. If different analyses, clarify the distinction.

ACC-006 [internal_consistency] - Location: Results section, 'Odds of being a virgin' outcome - The odds ratio for being a virgin is reported with an inconsistent CI. The text reports OR = 1.001 [95% CI: 0.992 to 1.005], but the upper CI bound (1.005) is greater than the point estimate (1.001), while the lower bound (0.992) is less than the point estimate. However, later text reports OR = 0.99 [95% CI: 0.98-1.00], which is a different estimate entirely. - Recommendation: Clarify whether these represent different instruments (DBP vs SBP) and ensure clear labeling. If same analysis, correct the discrepancy.

ACC-007 [internal_consistency] - Location: Results section, 'Self-reported wellbeing' outcome - Self-reported wellbeing estimates differ between two mentions in the text. First mention reports SMD = 0.003 [95% CI: -0.089 to 0.094], while later text reports SMD = 0.05 [95% CI: -0.10 to 0.21]. - Recommendation: Clarify which instrument (DBP vs SBP) each estimate corresponds to and ensure consistent labeling throughout.

ACC-008 [internal_consistency] - Location: Results section, 'Number of sexual partners' outcome - The number of sexual partners estimate differs between two mentions. First reports 1.984 [95% CI: -4.829 to 8.797], while later reports 1.79 [95% CI: -9.72 to 13.3]. - Recommendation: Clarify which instrument each estimate corresponds to and label consistently throughout the manuscript.

ACC-009 [missing_content] - Location: Tables 1, 2, and 4 - Table content could not be extracted for verification. Tables 1, 2, and 4 are referenced in the manuscript but their actual data content was not available for review, preventing verification of text-to-table concordance. - Recommendation: Provide complete table data to enable full verification of text-to-table concordance for all reported statistics.

Minor Issues

ACC-010 [presentation] - The clumping parameters are reported inconsistently. Text states 'kb = 10,000' but the standard abbreviation is 'kB' (kilobases) which is used elsewhere in the same paragraph. - Recommendation: Use consistent abbreviation 'kB' throughout the manuscript.

ACC-011 [narrative_alignment] - The CI for number of live births in women is reported as '-0.28 to 0.01' but the text states the estimate was 'in the opposite direction to that in men.' With beta = 0.14 and CI crossing zero, the direction claim needs clarification. - Recommendation: Verify the sign of the beta coefficient and CI bounds. If beta is positive (0.14), the CI should likely be approximately -0.01 to 0.28 for a symmetric interval.

ACC-012 [presentation] - P-value notation is inconsistent. Some instances use 'p' (lowercase) while the scientific convention often uses 'P' (uppercase) or italicized 'p'. Additionally, subscript notation varies (p_fdr vs p fdr). - Recommendation: Standardize p-value notation throughout the manuscript according to journal style guidelines.

ACC-013 [statistical_plausibility] - The F-statistics reported (31.58 for DBP, 25.53 for SBP) are described as indicating 'low levels of weak instrument bias.' While F > 10 is the conventional threshold, the interpretation could be more precise - these indicate adequate instrument strength, not necessarily 'low' bias. - Recommendation: Consider rephrasing to 'indicating adequate instrument strength (F > 10)' for more precise interpretation.

Table Verification Status

- **Table 1**: ' Issues Found - Table content could not be extracted. Table appears to contain glossary of genetic/MR terms based on text references.
- **Table 2**: ' Issues Found - Table content could not be extracted. Table reportedly contains MR and colocalization results for number of children analyses including H0-H4 posterior probabilities.
- **Table 4**: ' Issues Found - Table content could not be extracted. Table reportedly contains LD Check results.

Scientific Technical Writer

Model: anthropic/claude-opus-4-5-20251101

Journal Article Review: Manuscript Under Review

Summary Assessment

This Mendelian randomisation study investigating sildenafil's effects on male reproductive outcomes is generally well-written, with clear scientific communication appropriate for the field. The manuscript demonstrates good overall structure and logical flow. However, the writing requires attention to consistency issues (particularly formatting of technical terms and preposition usage), minor grammatical corrections, and standardization of terminology throughout the document.

Major Concerns

No major writing issues were identified in this manuscript. The prose is clear, the arguments are logically structured, and the technical communication meets the standards expected for a genetic epidemiology study in a medical journal.

Minor Issues

Consistency Issues

- **Abstract & Throughout**: The term "cis-MR" shows inconsistent formatting, with irregular spacing before the hyphen (appearing as "cis -MR" in places). Standardize to "*cis*-MR" with consistent italicization and no space before the hyphen throughout the manuscript.
- **Abstract & Throughout**: The phrase "age of first having sex" uses a non-standard preposition. Change to "age at first sexual intercourse" or "age at first having sex" and apply consistently throughout all sections.
- **Abstract, Main outcome measures**: "sub- sample" should be written as one unhyphenated word: "subsample."

Grammar Issues

- **Abstract, Main outcome measures**: The list "Number of children, age of first having sex, number of sexual partners, odds of being a virgin and self-reported wellbeing" would benefit from an Oxford comma before "and" for improved clarity in this complex series.
- **Abstract, Main outcome measures**: "age of first having sex" is grammatically awkward; revise to "age at first sexual intercourse" for more formal scientific prose.

Strengths

- Clear and logical organization of complex Mendelian randomisation methodology
- Appropriate use of discipline-specific terminology for genetic epidemiology

- Well-structured abstract that effectively communicates study design, outcomes, and key findings
- Technical concepts are explained with sufficient clarity for the target medical audience

Questions for Authors

1. Please confirm the preferred standardized phrasing for "age at first sexual intercourse" and ensure it is applied consistently throughout all sections of the manuscript.

2. Please verify the intended formatting for "*cis*-MR" (italicized prefix with hyphen, no space) and apply uniformly.

Recommendation

Minor Revision

Justification: The manuscript demonstrates good overall writing quality with no critical or major issues affecting comprehension or accuracy. The identified issues are minor and relate primarily to formatting consistency and small grammatical refinements. These can be addressed through careful copyediting without requiring substantial rewriting. Once the consistency issues with technical terminology and formatting are standardized throughout, the manuscript will meet publication standards for writing quality.