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 14 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: hJ9Wmt4wSNBss72jpUuP **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. **Paragraph 7: Incomplete Text Identification** - Paragraph 7: The text "using ITS linear regression to estimate immediate changes in quali" is incomplete and appears to be cut off mid-word. This needs to be completed.

2. **Patient Register Criteria Issue** - Issue: "*percentage of patients on each disease register who were not excluded by automatic criteria…*"

3. **MAJOR: Sample Size Reporting** - The number of practices in Scotland and England is never specified, despite this being essential information for interpreting the analysis and assessing statistical power. For STROBE compliance, the n

4. **MAJOR: Autocorrelation in Time Series** - Interrupted time series analyses with repeated measurements over time require adjustment for autocorrelation (serial correlation) in the error terms. Standard regression assumes independent errors, wh

5. **Study Design Specification** - Paragraph 7: The study design is described as "an interrupted time-series analysis". This is an underspecification. The design is a **controlled** interrupted time series (CITS), which is a much stron

Points of Divergence

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

Required Revisions

1. **Incomplete Paragraph Text** - Action: This needs to be completed
2. **Patient Register Criteria** - Action: Issue: "*percentage of patients on each disease register who were not excluded by automatic criteria…*"
3. **Sample Size Reporting** - Action: Add to Methods section: "There were [N] general practices in Scotland and [N] practices in England included in the analysis. Practice numbers remained stable across the study period [or describe any c
4. **Autocorrelation Handling** - Action: Specify in Statistical Analysis section: "We tested for first-order autocorrelation using the Durbin-Watson statistic and modeled autocorrelation using [specify method: Newey-West standard errors / Pr
5. **Study Design Specification** - Action: should be used consistently throughout the manuscript to accurately reflect the methodological rigor.
6. **Confidence Interval Correction** - Action: Change to: 95% CI -45.5 to -35.0
7. **Summary Box Revision** - Action: Consider revising to emphasize what's genuinely novel or actionable rather than restating findings.
8. **Recommendation Phrasing** - Action: Change to: These changes are consistent with the recommendations of many pay-for-performance programme designers
9. **Indicator Selection Clarification** - Action: Change to: stroke or transient ischaemic attack (TIA)
10. **Figure 1 Caption Update** - Action: Change to: Multiple group ITS analysis comparing indicators

Recommended Revisions

- [CRITICAL] Omission of Autocorrelation Handling in Statistical Methods
- [MAJOR] Insufficient Detail on the "Recording vs. Delivery" of Care Distinction
- ITS design may be too underdetermined (only 3 pre + 3 post annual points) to support trend-based counterfactual claims
- Potential differential measurement/recording change after Scotland abolished incentives (data generation may not be comp
- Indicator selection may induce selection-on-observables and limits inference ("subset" may not represent QOF as a whole)

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 14 consensus issues were identified that require attention before the manuscript can proceed.

REQUIRED CHANGES

1. Paragraph 7: The text "using ITS linear regression to estimate immediate changes in quali" is incomplete and appears to

2. Issue: “*percentage of patients on each disease register who were not excluded by automatic criteria…*”

3. MAJOR: Sample Size Not Reported

4. MAJOR: No Discussion of Autocorrelation in Time Series

5. Paragraph 7: The study design is described as "an interrupted time-series analysis". This is an underspecification. The

6. C: Paragraph 3, Abstract, Results section

7. Paragraph 4, Summary Box**: "What this study adds" largely repeats information from the abstract without adding new in

8. C: Paragraph 5, Introduction

9. C: Paragraph 7, Methods, Indicator Selection

10. T: Figure 1 Caption

SUGGESTED IMPROVEMENTS

1. [CRITICAL] Omission of Autocorrelation Handling in Statistical Methods

2. [MAJOR] Insufficient Detail on the "Recording vs. Delivery" of Care Distinction

3. ITS design may be too underdetermined (only 3 pre + 3 post annual points) to support trend-based counterfactual claims

4. Potential differential measurement/recording change after Scotland abolished incentives (data generation may not be comp

5. Indicator selection may induce selection-on-observables and limits inference (“subset” may not represent QOF as a whole)

Domain Expert

Model: gemini/gemini-2.5-pro

Journal Article Review: Manuscript Under Review

Summary Assessment

This manuscript presents a well-designed and highly policy-relevant national controlled interrupted time series (CITS) analysis on the impact of abolishing a major primary care pay-for-performance scheme in Scotland. The use of England as a control group within a natural experiment framework is a significant strength. The findings, which suggest a decline in recorded quality of care for most process indicators following the withdrawal of financial incentives, have important implications for international health policy.

Major Concerns

[MAJOR] Critical Omission of Autocorrelation Handling in Statistical Methods

- **Problem**: The description of the interrupted time series (ITS) analysis critically omits any mention of how autocorrelation was assessed or corrected. Time series data are almost invariably autocorrelated, and failing to account for this can lead to underestimated standard errors and an inflated Type I error rate, potentially rendering the findings of statistical significance spurious. This is a fundamental requirement for rigorous ITS analysis and a major gap in the current methodology.
- **Recommendation**: The authors must explicitly state the method used to test for autocorrelation (e.g., Durbin-Watson test, plotting residuals, correlograms) and the method used to adjust for it in the regression models (e.g., using Newey-West standard errors, the Cochrane-Orcutt procedure, or fitting an ARIMA model). Adherence to best-practice reporting guidelines for ITS studies is essential for the credibility of the results.

Minor Issues

(No minor issues were detailed in the provided section review for consolidation.)

Strengths

- **Strong Study Design**: The use of a national controlled interrupted time series (CITS) analysis is a robust and appropriate method for this research question.
- **High Policy Relevance**: The study leverages a natural experiment to provide crucial evidence on the effects of removing financial incentives in healthcare, a topic of major international interest.

- **Important Contribution**: The findings provide significant new insights into the sustainability of quality improvements after the withdrawal of pay-for-performance schemes.

Questions for Authors

(No questions were detailed in the provided section review for consolidation.)

Recommendation

Major Revision

Justification: The study addresses a critical policy question with a strong and appropriate quasi-experimental design. Its findings are timely and of high potential impact. However, the failure to report on the handling of autocorrelation in the time series analysis is a major methodological omission that undermines confidence in the reported statistical significance. This issue must be thoroughly addressed through re-analysis (if necessary) and transparent reporting before the manuscript can be considered for publication. The core study is strong, and these revisions are achievable.

Adversarial Skeptic

Model: openai/gpt-5.2

Journal Article Review: Manuscript Under Review

Summary Assessment

This manuscript addresses an important, policy-relevant question using a plausible natural experiment: Scotland's abolition of QOF incentives contrasted with England's continuation. The framing is clear and the controlled interrupted time series (ITS) approach is, in principle, appropriate.

However, based on the provided excerpt, the design appears underpowered/underdetermined for trend-based counterfactual inference (only 3 annual pre- and 3 annual post-intervention time points), which substantially weakens causal interpretation and makes results highly sensitive to single-year anomalies.

Major Concerns

1. ITS is underdetermined with only 3 pre- and 3 post-intervention annual observations

- **Quote**: "*the data consisted of three annual measurements before and three annual measurements after the year of removal financial incentives in Scotland.*"
- **Problem (assumption stress-test)**: The analysis implicitly assumes that a stable, approximately linear pre-intervention trend can be estimated from only three observations and used to generate a credible counterfactual. With *n*=3 pre points, the slope is extremely sensitive to any idiosyncratic year (e.g., unusual influenza season, coding change, guideline update), and standard ITS diagnostics (autocorrelation, structural breaks, influential points, functional-form checks) are largely uninformative. This raises a serious risk that apparent "effects" reflect regression-to-the-mean or one-off shocks rather than the policy change.
- **Recommendation (what would address it)**: - Explicitly acknowledge this identification limitation in the Methods/Discussion (and temper any causal language accordingly). - Add robustness checks that do **not** hinge on estimating a linear pre-trend from three points (e.g., pre-period mean anchoring; leave-one-year-out analyses; alternative functional forms; placebo interruption years / permutation-based inference). - If feasible, increase the number of time points (e.g., quarterly/monthly) or extend the series further back using harmonized indicator definitions.

2. Controlled ITS validity hinges on comparator comparability and parallel pre-trends, which are difficult to assess with only three pre points

- **Problem (what if…)**: Even if England is a reasonable control, the controlled ITS effectively assumes that, absent Scotland's abolition, Scotland would have followed England (or the Scotland pre-trend) after the intervention. With only three annual pre points, any "parallel trends" assessment is weak, and differential contemporaneous shocks (other policy changes, guideline updates, data-system transitions) could drive divergences that are then misattributed to QOF removal.

- **Why it matters**: If the England–Scotland relationship is unstable or affected by concurrent changes, the estimated "abolition effect" can be biased in either direction.

- **Fix needed**: Pre-specify and report diagnostic evidence on pre-intervention comparability; consider additional controls/negative control outcomes; and include sensitivity analyses that allow for non-parallel pre-trends (e.g., differential slopes, alternative counterfactual constructions such as synthetic control—if feasible given the data structure).

3. Potential differential measurement/recording changes post-abolition could masquerade as true care changes

- **Problem (authors assume…)**: The design appears to assume that indicator measurement is consistent over time and across countries. Yet a central concern in pay-for-performance repeal is that recording/documentation incentives change even if clinical care does not—creating a "measurement effect" (coding/recording drop) rather than a "care effect."

- **Why it matters**: If observed declines are partly (or mainly) reduced recording, conclusions about worsening quality of care (or loss of performance) would be overstated.

- **Fix needed**: Provide evidence/argument that recording practices remained comparable post-abolition; where possible, triangulate with outcomes less susceptible to coding incentives (hard endpoints), auditing data, or external validation; and present the interpretation explicitly as "care and/or recording" unless disentangled.

4. Multiplicity/multiple testing risk (many indicators, subgroup contrasts, or model variants) may inflate false positives

- **Problem (what if…)**: If numerous indicators and contrasts are tested without correction or a clearly pre-specified primary outcome, statistically "significant" findings may reflect chance—especially with a fragile time-series design.

- **Why it matters**: Overstates evidentiary strength and can yield a selective narrative around nominal $p<0.05$ results.

- **Fix needed**: Clearly identify a primary outcome (or limited family of primary outcomes), apply multiplicity control (e.g., FDR), and/or emphasize effect sizes with uncertainty intervals and a hierarchy of inference (confirmatory vs exploratory).

5. Inference with very few time points: standard errors/model assumptions may be unreliable

- **Problem (assumption checking)**: With six annual points total, segmented regression standard errors, autocorrelation corrections, and model-based inference can behave poorly; results can hinge on distributional/independence assumptions that are not testable here.

- **Why it matters**: Can lead to overstated precision and overconfident conclusions.

- **Fix needed**: Use small-sample-appropriate approaches (e.g., permutation/randomization inference around interruption timing; bootstrap with caution and transparency), present sensitivity of estimates to single observations, and foreground uncertainty.

Minor Issues

The prompt indicates numerous minor issues across sections, but only a limited excerpt of the section-by-section reviews is visible here. I cannot responsibly consolidate the full set of minor comments without the missing content. If you share the remaining section reviews (or the manuscript text for those parts), I can merge all minor edits into a single, non-redundant list as requested.

Strengths

- High policy relevance with a credible natural experiment setup (Scotland vs England).
- Clear research question and generally appropriate choice of a controlled ITS framework in principle.
- National-level perspective likely enhances real-world applicability (contingent on measurement comparability).

Questions for Authors

1. Are higher-frequency data (quarterly/monthly) available for these indicators to increase time points and strengthen ITS identification? If not, why not?
2. What evidence can you provide that indicator definitions and recording/coding practices remained stable across the intervention, particularly after Scotland's incentive removal?
3. How many outcomes/indicators and subgroup analyses were tested, and what was pre-specified as primary vs exploratory? Will you apply multiplicity adjustments or a clearly articulated hierarchy?
4. What do results look like under leave-one-year-out analyses (both pre and post years), and under placebo interruption years?
5. How do you justify the England comparator's suitability given contemporaneous policies/guidelines—especially with limited pre-period information?

Recommendation

Major Revision

Justification: The manuscript's question and natural experiment are compelling, but the current ITS setup (as described) appears too weakly identified to support strong counterfactual/causal claims, and it is highly vulnerable to measurement changes and analytic fragility. Substantial additional sensitivity analyses, clearer limitation statements, and (ideally) more time points or alternative inference strategies are needed for robust, publishable conclusions.

Systematic Reviewer

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

Journal Article Review: Manuscript Under Review

Summary Assessment

This manuscript presents a policy-relevant controlled interrupted time series analysis examining the impact of removing financial incentives for quality-of-care indicators in Scottish general practices, using England as a control. The natural experiment design is appropriate and timely. However, the manuscript requires substantial revision before publication. Most critically, the statistical methods section is incomplete (ending mid-sentence), and there is a mathematically impossible confidence interval in the abstract. Additionally, essential methodological details are missing, including sample sizes, power calculations, and specifications for handling time series autocorrelation and practice-level clustering.

Major Concerns

1. **Incomplete Manuscript Submission** — Statistical Analysis section (Paragraph 7): - Problem: The text terminates abruptly mid-sentence: "Initial analysis used single-group analysis of trends in Scotland before and after incentive withdrawal at the end of financial year 2015-2016, using ITS li—" This suggests the manuscript was not fully prepared for submission. - Recommendation: Complete the Statistical Analysis section, ensuring full description of the controlled interrupted time series methodology, model specifications, and all analytical approaches. Verify that all sections of the manuscript are complete before resubmission.

2. **Confidence Interval Mathematical Error** — Abstract (Paragraph 3): - Problem: The reported confidence interval is impossible: "(-5.0 percentage-points, 95%CI 8.4 to -1.5 for HbA1c less than 75mmol/l)". The lower bound (8.4) cannot exceed the upper bound (-1.5), and the point estimate (-5.0) lies outside the stated interval. This should be "95%CI -8.4 to -1.5". - Recommendation: Correct to "-5.0 percentage-points (95%CI -8.4 to -1.5)" and systematically verify all confidence intervals throughout the manuscript (abstract, results tables, figures) for similar transposition errors.

3. **Missing Sample Size and Power Analysis** — Methods: - Problem: The manuscript does not report the total number of practices included in Scotland and England, the number of measurements per practice, or any sample size justification. For a natural experiment, a power/sensitivity analysis is essential to establish whether the design can detect policy-relevant effect sizes. - Recommendation: Report: (1) Number of practices in Scotland and England at baseline and each follow-up; (2) attrition/dropout patterns if any; (3) a sensitivity analysis demonstrating minimum detectable effect sizes given your sample and study design. Present this as a table or flow diagram showing practice inclusion.

4. **Inadequate Specification of Time Series Autocorrelation Methods** — Statistical Analysis: - Problem: Interrupted time series data inherently exhibit autocorrelation (serial correlation over time). The manuscript does not specify how autocorrelation was assessed or addressed. Failing to account for autocorrelation inflates Type I error rates and produces incorrect confidence intervals. - Recommendation: Specify: (1) Tests used to detect autocorrelation (e.g., Durbin-Watson, ACF/PACF plots); (2) Correlation structure assumed in models (e.g., AR(1), MA(1)); (3) How model residuals were checked for remaining autocorrelation. If using generalized least squares or ARIMA models, state this explicitly with model orders.

5. **Inadequate Specification of Practice-Level Clustering Methods** — Statistical Analysis: - Problem: Data are nested (repeated measurements within practices, practices potentially within regions). The manuscript does not specify how practice-level clustering was addressed. Ignoring clustering produces underestimated standard errors and false-positive findings. - Recommendation: Specify: (1) Whether practice-level random effects or fixed effects were included; (2) Clustering approach for standard errors (e.g., cluster-robust standard errors, multilevel modeling); (3) Intraclass correlation coefficients (ICCs) for practice clustering; (4) Sensitivity analyses showing results are robust to clustering assumptions.

6. **Parallel Trends Assumption Not Verified** — Methods/Results: - Problem: The controlled ITS design (difference-in-differences framework) assumes Scotland and England had parallel trends in quality indicators before the intervention. This critical assumption is mentioned but not formally tested or demonstrated. - Recommendation: Present pre-intervention trend comparisons for all 16 indicators graphically (supplementary figure) and statistically (test for differential pre-intervention slopes between Scotland and England). If parallel trends don't hold for some indicators, acknowledge this limitation and consider alternative designs (e.g., synthetic control methods) or restrict conclusions to indicators meeting the assumption.

7. **Multiple Comparison Adjustment Not Specified** — Statistical Analysis: - Problem: The study examines 16 separate quality indicators, creating 16 hypothesis tests. With no adjustment for multiple comparisons, the expected number of false positives at $\alpha = 0.05$ is 0.8 tests. The manuscript does not state whether/how multiple testing was addressed. - Recommendation: Specify your approach: (1) If no adjustment: explicitly justify (e.g., "indicators were pre-specified policy targets, each representing distinct clinical outcomes; adjustment would be overly conservative"); (2) If adjustment used: specify method (Bonferroni, Benjamini-Hochberg FDR) and report both adjusted and unadjusted p-values. Consider designating primary vs. exploratory outcomes.

8. **Missing Subgroup Analyses Specification** — Methods: - Problem: The abstract mentions "pre-specified subgroup analyses" but the methods section does not describe what subgroups were examined, why they were chosen, or how they were pre-specified. - Recommendation: Add a subsection describing: (1) All pre-specified subgroups (e.g., by practice size, deprivation, urban/rural, baseline performance); (2) Rationale for each subgroup; (3) Evidence of pre-specification (registered protocol, SAP); (4) Methods for testing subgroup interactions; (5) Multiple testing adjustment for subgroup analyses.

Minor Issues

- **Abstract formatting**: Confidence intervals should follow a consistent format. Recommend "(estimate, 95%CI lower to upper)" throughout.

- **Time period description**: Clarify whether "financial year 2015-2016" refers to April 2015-March 2016 (UK convention) or calendar year. Specify exact intervention date (stated as April 2016 in summary but should be explicit in methods).

- **Data source description**: Specify the names of the data systems used (e.g., QOF database, NHS Digital) and confirm data availability/access procedures.
- **Outcome measurement timing**: Clarify whether measurements are annual snapshots (single point) or aggregated over the year, and specify measurement months/quarters.
- **Missing data handling**: Describe how missing practice data or incomplete follow-up was handled.
- **Statistical software**: Report the statistical software and package versions used for analysis (important for reproducibility of time series models).

Strengths

- **Policy-relevant natural experiment**: Exploits a well-defined policy change in Scotland with England as a contemporaneous control group, providing stronger causal inference than single-group before-after comparisons.
- **Comprehensive outcome coverage**: Examining 16 quality indicators provides a broad assessment of incentive removal impact across multiple clinical domains.
- **Appropriate study design**: Controlled interrupted time series is the optimal design for this quasi-experimental evaluation.
- **Population-level data**: Using all general practices (rather than a sample) maximizes statistical power and generalizability.

Questions for Authors

1. Were any practices excluded from analysis? If so, on what criteria and how many?
2. Were the 16 quality indicators pre-registered before analysis, or were they selected post-hoc? If pre-registered, please reference the protocol.
3. Did Scotland and England have identical indicator definitions and measurement procedures during the study period?
4. Were there any contemporaneous policy changes or events in Scotland or England (2013-2019) that might confound the incentive withdrawal effect?
5. How was the three-year follow-up period chosen? Was this determined by data availability or based on hypothesized time to effect?
6. Were sensitivity analyses conducted using different modeling approaches (e.g., segmented regression, ARIMA with transfer function)?

Recommendation

Major Revision

Justification: The research question is important and the study design is sound, but the manuscript cannot be evaluated in its current state. The incomplete methods section and mathematical error in the abstract require immediate correction. More fundamentally, critical methodological details are missing (sample sizes, power analysis, autocorrelation handling,

clustering approach, parallel trends verification, and multiple testing adjustment). These are not optional reporting elements for interrupted time series studies—they are essential for assessing validity. The authors should consult STROBE guidelines for observational studies and specific guidance for ITS studies (e.g., Bernal et al. 2017, J Epidemiol Community Health). With thorough revision addressing these methodological reporting gaps, this study could make a valuable contribution to the health policy literature.

Pragmatic Reviewer

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

Journal Article Review: Manuscript Under Review

Summary Assessment

This natural experiment examining Scotland's withdrawal from the Quality and Outcomes Framework provides valuable evidence on the impact of financial incentives on primary care quality. The interrupted time-series design is methodologically sound and the scope comprehensive. However, the manuscript requires substantial revision to translate statistical findings into clinically meaningful terms and to clarify whether observed changes reflect actual care delivery versus documentation practices. Currently, the paper reports numerous percentage-point changes without helping readers understand patient-level health consequences.

Major Concerns

1. **Statistical Findings Lack Clinical Translation** — Abstract and Results: - **Problem**: The abstract reports effect sizes (e.g., "-18.5 percentage-points for blood pressure control in PAD, -16.6 for stroke, -10.4 for diabetes") but never translates these to patient impact. Non-specialist readers cannot determine clinical significance. What do these changes mean for cardiovascular events, hospitalizations, or mortality? How many patients are affected in absolute terms? - **Quote**: "blood pressure control in patients with peripheral arterial disease (-18.5 percentage-points, 95%CI -22.1 to -14.9), stroke (-16.6 percentage-points, 95%CI -20.6 to -12.7), diabetes (-10.4 percentage-points, 95%CI -13.0 to -7.8)" - **Recommendation**: (1) Add to Abstract after listing key findings: "In absolute terms, these changes represent approximately [X thousand] fewer patients with controlled blood pressure across Scotland, potentially leading to [estimated number] additional cardiovascular events over [timeframe] based on established risk relationships." (2) In Discussion, interpret at least 2-3 key findings in terms of downstream health outcomes using published risk equations. For example, translate BP control changes to stroke/MI risk using Framingham or similar validated models.

2. **Critical Conceptual Ambiguity: Care Delivery vs. Documentation** — Throughout: - **Problem**: The manuscript does not adequately distinguish between three distinct phenomena: (a) actual deterioration in care quality, (b) reduced documentation of unchanged care, or (c) gaming that inflated baseline measures. This is a fundamental interpretive issue that affects all recommendations. The paper measures what gets recorded, not necessarily what care is delivered. - **Recommendation**: (1) Add a dedicated subsection in Discussion titled "Documentation vs. Care Delivery" that explicitly addresses this limitation. (2) Discuss which indicators are most susceptible to documentation changes (e.g., blood pressure recording) versus those reflecting actual care processes (e.g., medication prescribing). (3) In Abstract and Conclusion, qualify main findings: "reductions in recorded quality" not "reductions in quality" throughout. (4) Consider analyzing indicators less susceptible to documentation bias as sensitivity analyses.

3. **Effect Sizes Without Precision Interpretation** — Results: - **Problem**: Confidence intervals are reported but not interpreted for actionability. Some CIs are wide enough to span both negligible and substantial effects. Readers need guidance on what the range of plausible effects means for policy decisions. - **Recommendation**: For key findings, add interpretive statements about CI width. For example: "The 95% CI for diabetes BP control spans 7.8 to 13.0 percentage-points, suggesting the true effect is robustly moderate-to-large in magnitude" or conversely "The CI for [indicator X] crosses both clinically meaningful and negligible effect thresholds, limiting actionable conclusions."

4. **Missing Practical Implications for Policymakers** — Discussion: - **Problem**: The paper stops at "incentives matter" without guiding next steps. Should England reverse course? Should Scotland reinstate modified incentives? What features of incentive programs preserve quality while reducing gaming? Policymakers reading this need actionable guidance. - **Recommendation**: Add a subsection "Implications for Policy Design" that addresses: (1) Under what conditions might incentive removal be appropriate versus harmful? (2) What alternative quality assurance mechanisms might replace withdrawn incentives? (3) Which clinical areas appear most vulnerable to incentive withdrawal and might require targeted intervention? (4) What does this imply for ongoing QOF reforms in England?

5. **Abstract Lacks Stand-Alone Clarity** — Abstract: - **Problem**: The abstract assumes familiarity with QOF, percentage-point changes, and ITS methodology. A general medical reader in JAMA or BMJ would struggle to extract the core message. What changed, why does it matter, and what should happen as a result? - **Recommendation**: Restructure abstract to prioritize clarity: (1) First sentence: explain what QOF is in plain terms. (2) Results: lead with the most important finding in clinical terms, then provide statistical details. (3) Conclusion: add one concrete "take-home" sentence suitable for a press release. Example: "Withdrawal of financial incentives was associated with clinically meaningful reductions in recorded care quality, particularly for cardiovascular risk management, suggesting incentive programs—despite their limitations—play a role in maintaining documented quality standards."

Minor Issues

- **Abstract, Methods**: "16 indicators" - specify upfront that these are process measures (e.g., BP recording) vs. outcome measures (e.g., cardiovascular events) to set appropriate expectations.
- **Introduction**: The gap between Scotland and England's QOF policies should be introduced earlier for readers unfamiliar with UK healthcare systems.
- **Results presentation**: Consider leading with the most clinically impactful finding rather than listing all 16 indicators in descending order. This improves narrative flow.
- **Figure legends**: Ensure all figures are interpretable without reading the main text. Current legends may assume too much knowledge about QOF timing and indicator definitions.
- **Statistical methods**: Confirm whether seasonal adjustment was performed for indicators with known seasonal variation (e.g., flu vaccination, diabetes screening).
- **Sensitivity analyses**: Consider mentioning whether results were robust to different control group definitions or alternative time windows.
- **Limitations**: Acknowledge inability to observe care quality not captured in electronic health records (e.g., patient counseling, shared decision-making).
- **Limitations**: Discuss potential contamination if Scottish practitioners altered behavior in anticipation of QOF withdrawal before official implementation.

- **Discussion**: Address why some indicators showed minimal change—what distinguished robust from vulnerable quality measures?
- **Tables**: Add a column showing baseline achievement rates to contextualize the magnitude of changes.
- **References**: Ensure inclusion of recent systematic reviews on pay-for-performance to position findings within broader literature.
- **Language precision**: Replace "quality deteriorated" with "recorded quality declined" throughout to maintain conceptual accuracy.
- **Methods**: Clarify whether indicator definitions remained constant between Scotland and England during the study period.
- **Results**: For non-significant findings, report the CI width to distinguish "no effect" from "underpowered."
- **Conclusion**: Avoid overstating causality—observed associations are consistent with incentive effects but alternative explanations remain possible.
- **Acknowledgments**: Consider thanking patient advisors if any were consulted on interpretation of clinical significance.
- **Data availability**: Confirm whether code for ITS analysis will be shared to support reproducibility.

Strengths

- **Strong natural experiment design**: Scotland vs. England comparison provides robust quasi-experimental evidence on a policy question of international relevance.
- **Comprehensive scope**: Analysis of 16 indicators across multiple chronic conditions provides broader picture than single-disease studies.
- **Appropriate methodology**: Interrupted time-series with control group addresses key confounders and temporal trends.
- **Policy relevance**: Addresses timely question as many countries reconsider pay-for-performance programs.
- **Transparent reporting**: Effect sizes and confidence intervals reported consistently; methods appear replicable.
- **Long follow-up**: Sufficient post-intervention period to observe sustained effects rather than transient responses.

Questions for Authors

1. Have you conducted or considered a patient-level analysis linking the observed changes in process measures to actual health outcomes (cardiovascular events, mortality)? Even modeled estimates would substantially strengthen clinical interpretation.
2. Can you provide evidence from qualitative interviews or surveys with Scottish GPs on whether the changes reflect actual care reduction versus documentation reduction?

3. Were there any compensatory quality assurance mechanisms introduced in Scotland after QOF withdrawal that might confound interpretation?

4. For indicators showing minimal change, what distinguishes them from those showing large declines? Can this inform which aspects of care are robust without incentives?

5. Did you analyze whether effects varied by practice characteristics (e.g., list size, deprivation, baseline quality) to identify which settings are most vulnerable to incentive withdrawal?

Recommendation

Major Revision

Justification: This study addresses an important policy question with sound methodology and provides valuable evidence. However, it currently fails the pragmatic standard of helping non-specialists understand "so what?" and "now what?" The statistical findings must be translated into clinical terms, the conceptual ambiguity between documentation and care delivery must be addressed explicitly, and practical implications for policy must be developed. These are significant presentation issues that affect the paper's interpretability and impact, but they do not undermine the core scientific contribution. With focused revision on clinical interpretation and practical significance, this paper could make an important contribution to health policy literature.

The core finding—that withdrawing incentives was associated with reduced recorded quality—is valuable and appears robust. The authors need to help readers understand what this means for patient health and what policymakers should do with this information.

Statistical Methods Agent

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

Statistical Methods Review

Discipline: Health Services Research / Health Policy / Epidemiology

Study Design: Controlled Interrupted Time Series (Natural Experiment)

Statistical Methods Identified: - Interrupted time series (ITS) regression - Single-group ITS analysis - Multiple-group ITS analysis (controlled ITS) - Linear regression with step change and trend change parameters - Stata itsa command

Overall Quality: Acceptable

Summary

This manuscript examines the impact of abolishing primary care financial incentives (QOF) in Scotland on quality-of-care indicators using a controlled interrupted time series (ITS) design comparing Scotland to England. The study design is appropriate for this natural experiment, and the use of England as a control group strengthens causal inference. However, several statistical issues require attention.

Critical Issues: 1. Multiple testing across 16 indicators without correction inflates Type I error substantially 2. Limited time points (3 pre, 3 post) constrain ITS validity and preclude autocorrelation assessment 3. Practice-level clustering not addressed in the analysis

Major Issues: 1. No sensitivity analyses for parallel trends assumption 2. Absence of effect size measures beyond percentage point differences 3. Missing model diagnostics and assumption checks 4. No formal power analysis or sample size justification for detecting clinically meaningful differences

Minor Issues: 1. Inconsistent confidence interval reporting in abstract 2. Missing degrees of freedom 3. No explicit statement about statistical significance threshold

Overall, the study addresses an important policy question with a reasonable design, but the statistical analysis requires strengthening to meet top-tier journal standards. The lack of multiple testing correction is particularly concerning given the 16 simultaneous comparisons, and the clustering of patients within practices should be formally addressed.

Statistical Issues (14 found)

STAT-001: Multiple Testing (Critical)

Location: Methods section, Page 6; Results section, Pages 7-9; Tables 2 and 3

The study conducts 16 separate ITS analyses (one for each quality indicator) without any adjustment for multiple testing. This substantially inflates the familywise Type I error rate. With 16 independent tests at $\alpha = 0.05$, the probability of at least one false positive is approximately 1-(0.95)^16 = 56%. The problem is compounded because each indicator analysis includes multiple parameters (step change, trend change, 3-year difference), effectively multiplying the number of comparisons. The authors report 12 of 16 indicators showing significant reductions at 1-year and 10

of 16 at 3-years, but without correction, some of these could be false positives. This is particularly important for policy-relevant conclusions.

Evidence: "Results: In Scotland, performance reduced significantly compared to England on 12 of the 16 quality-of-care indicators 1-year after QOF was abolished, and on 10 of 16 indicators 3-years after abolition."

Recommendation: Apply a multiple testing correction appropriate for this context. Given that the 16 indicators represent a pre-specified family of tests examining the same policy intervention, and that some indicators are correlated (e.g., multiple BP indicators, multiple HbA1c thresholds), I recommend: (1) Primary approach: Apply Benjamini-Hochberg FDR correction at 5% to control the expected proportion of false discoveries. (2) Sensitivity analysis: Also report Holm-Bonferroni adjusted p-values for readers preferring FWER control. (3) Consider grouping indicators by type (complex process, intermediate outcome, treatment) and applying corrections within groups if different error rates are acceptable for different indicator types.

Code Examples:

Stata (packages: multproc, qqvalue):

```
* After running itsa for all 16 indicators, collect p-values
* Example: Store p-values from step change at 1-year
matrix pvals = (0.001, 0.002, 0.015, ...) // 16 p-values

* Install multproc package if needed
* ssc install multproc

* Apply Benjamini-Hochberg FDR correction
multproc, pval(pvals) method(hochberg)

* Or use qqvalue for FDR
* ssc install qqvalue
qqvalue pvals, method(bh) qvalue(qvals)
```

Stata implementation using multproc or qqvalue packages for FDR correction

R (packages: stats):

```
# Collect p-values from all 16 ITS analyses
p_values <- c(0.001, 0.002, 0.015, 0.023, 0.034, 0.041,
              0.048, 0.052, 0.067, 0.089, 0.12, 0.15,
              0.21, 0.34, 0.45, 0.67)  # Example values

indicator_names <- c("MH02", "DM12", "PAD02", "STIA03", "HYP06",
                     "CHD02", "DM03", "DM02", "DM09", "DM08",
                     "DM07", "STIA09", "COPD07", "CHD07", "DM18", "CHD05")

# Benjamini-Hochberg FDR correction
p_adjusted_bh <- p.adjust(p_values, method = "BH")

# Holm-Bonferroni for sensitivity analysis
p_adjusted_holm <- p.adjust(p_values, method = "holm")

# Create results table
results <- data.frame(
  Indicator = indicator_names,
  P_unadjusted = p_values,
  P_BH_adjusted = p_adjusted_bh,
  P_Holm_adjusted = p_adjusted_holm,
  Sig_BH = p_adjusted_bh < 0.05,
  Sig_Holm = p_adjusted_holm < 0.05
)

print(results)
cat("\nSignificant after BH correction:", sum(results$Sig_BH), "of 16")
cat("\nSignificant after Holm correction:", sum(results$Sig_Holm), "of 16")
```

Base R implementation using p.adjust() function

Example Write-up: > To account for multiple testing across 16 quality indicators, we applied the Benjamini-Hochberg procedure to control the false discovery rate (FDR) at 5%. We report both unadjusted and FDR-adjusted p-values. After FDR correction, 9 of 16 indicators remained statistically significant at 3-years post-intervention (adjusted $p < 0.05$). As a sensitivity analysis, we also applied Holm-Bonferroni correction for familywise error rate control, with 7 indicators remaining significant.

Literature Support: Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B. 1995;57:289-300. Also see: Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology. 1990;1:43-46 for alternative view, though this requires explicit justification.

STAT-002: Model Assumptions (Critical)

Location: Methods section, Page 6; Discussion, Page 10

The ITS analysis uses only 3 time points before and 3 time points after the intervention. This is the absolute minimum for ITS analysis and creates several problems: (1) It precludes formal testing for autocorrelation, which is a key assumption violation in time series data; (2) It provides insufficient data to reliably estimate pre-intervention trends; (3) It reduces power to detect changes in trend; (4) It makes the analysis highly sensitive to outliers at any single time point. The authors acknowledge this limitation but do not provide sensitivity analyses or alternative approaches.

Evidence: "Limitations include that the time series have relatively few time-points (three before the intervention and three after) which is the minimum required for the method, and precludes examining for the presence of auto-correlation."

Recommendation: Given the data constraints, several approaches should be considered: (1) Report Newey-West standard errors which are robust to autocorrelation and heteroskedasticity even with limited time points; (2) Conduct sensitivity analyses using different functional forms (e.g., assuming no pre-trend vs. allowing pre-trend); (3) Consider bootstrapped confidence intervals; (4) If quarterly or monthly data are available for any indicators, use these for sensitivity analysis; (5) Explicitly test the parallel trends assumption between Scotland and England in the pre-intervention period.

Code Examples:

Stata (packages: base Stata):

```
* Standard ITS with itsa command
itsa outcome, single trperiod(2016) fig

* Alternative: Manual regression with Newey-West standard errors
* Create time variables
gen time = _n
gen post = (year >= 2016)
gen time_post = (year - 2016) * post

* Newey-West robust standard errors (lag = 1)
newey outcome time post time_post, lag(1)

* For multiple-group analysis
gen scotland = (country == "Scotland")
gen scotland_post = scotland * post
gen scotland_time = scotland * time
gen scotland_time_post = scotland * time_post

newey outcome time post time_post scotland scotland_post ///
       scotland_time scotland_time_post, lag(1)

* Bootstrap confidence intervals as sensitivity
bootstrap _b, reps(1000) seed(12345): regress outcome time post time_post
```

Newey-West standard errors provide robustness to autocorrelation. Bootstrap provides alternative inference.

R (packages: sandwich, lmtest, boot):

```
library(sandwich)
library(lmtest)
library(boot)

# Fit ITS model
model <- lm(outcome ~ time + post + time_post + scotland +
            scotland_post + scotland_time + scotland_time_post,
            data = df)

# Newey-West HAC standard errors
coeftest(model, vcov = NeweyWest(model, lag = 1))

# Bootstrap confidence intervals
boot_its <- function(data, indices) {
  d <- data[indices, ]
  fit <- lm(outcome ~ time + post + time_post + scotland +
            scotland_post + scotland_time + scotland_time_post,
            data = d)
  return(coef(fit))
}

set.seed(12345)
boot_results <- boot(data = df, statistic = boot_its, R = 1000)
boot.ci(boot_results, type = "bca", index = 6)  # For scotland_post coefficient
```

R implementation with HAC standard errors and bootstrap inference

Example Write-up: > We acknowledge that our time series has only 6 time points (3 pre-intervention, 3 post-intervention), which limits our ability to formally test for autocorrelation. To address this, we estimated Newey-West heteroskedasticity and autocorrelation consistent (HAC) standard errors with a lag of 1. As a sensitivity analysis, we also estimated models assuming no pre-intervention trend (step-change only) and compared results. The parallel trends assumption was assessed by testing for differential pre-intervention trends between Scotland and England (all $p > 0.10$).

Literature Support: Kontopantelis E, et al. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. BMJ. 2015;350:h2750. Recommends minimum of 8-10 time points. Also: Bernal JL, et al. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol. 2017;46:348-355.

STAT-003: Model Assumptions (Major)

Location: Methods section, Page 6

The analysis does not account for the hierarchical structure of the data, where patients are nested within practices. The data consists of practice-level aggregates (979 practices in Scotland, 7921 in England), but the ITS analysis appears to use country-level aggregated data without accounting for between-practice variation. This ignores the clustering of outcomes within practices and may underestimate standard errors, leading to inflated Type I error rates. The intraclass correlation (ICC) at the practice level could be substantial for quality indicators.

Evidence: "The analysis included data from 979 practices with 5,599,171 registered patients in Scotland and 7921 practices with 56,270,628 registered patients in England in 2013-14... We used this analysis to calculate absolute differences in documented quality-of-care in Scotland compared to England at 3-years after the removal of financial incentives. Analyses were conducted in Stata v14."

Recommendation: The authors should: (1) Clarify whether analyses used practice-level data or country-level aggregates; (2) If practice-level data are available, use multilevel/mixed-effects ITS models that account for practice-level clustering, or cluster-robust standard errors; (3) Report the intraclass correlation coefficient (ICC) for each indicator to quantify between-practice variation; (4) If only country-level aggregates were used, acknowledge this limitation and consider weighting by practice size or patient numbers.

Code Examples:

Stata (packages: base Stata):

```
* If practice-level data available:
* Mixed-effects ITS with random intercepts for practices
mixed outcome time post time_post scotland scotland_post ///
      scotland_time scotland_time_post || practice:, mle

* Calculate ICC
estat icc

* Alternative: Cluster-robust standard errors
regress outcome time post time_post scotland scotland_post ///
        scotland_time scotland_time_post, vce(cluster practice)

* If only aggregate data, weight by number of patients
itsa outcome [aweight=n_patients], single trperiod(2016)
```

Mixed-effects model or cluster-robust SEs to account for practice-level clustering

R (packages: lme4, lmerTest, performance, estimatr, clubSandwich):

```
library(lme4)
library(lmerTest)
library(clubSandwich)

# Mixed-effects ITS model with random intercepts for practices
model_mixed <- lmer(outcome ~ time + post + time_post + scotland +
                    scotland_post + scotland_time + scotland_time_post +
                    (1 | practice), data = df)

summary(model_mixed)

# Calculate ICC
performance::icc(model_mixed)

# Alternative: Cluster-robust standard errors
library(estimatr)
model_cr <- lm_robust(outcome ~ time + post + time_post + scotland +
                      scotland_post + scotland_time + scotland_time_post,
                      data = df, clusters = practice, se_type = "CR2")
summary(model_cr)
```

R implementation with mixed-effects models or cluster-robust standard errors

Example Write-up: > To account for clustering of outcomes within practices, we used multilevel interrupted time series regression with random intercepts for practices. The intraclass correlation coefficient (ICC) ranged from 0.05 to 0.15 across indicators, indicating moderate clustering. Standard errors were adjusted for this clustering, resulting in wider confidence intervals than naive analyses. As a sensitivity analysis, we also used cluster-robust standard errors clustered at the practice level.

Literature Support: Rabe-Hesketh S, Skrondal A. Multilevel and Longitudinal Modeling Using Stata. 3rd ed. Stata Press; 2012. Also: Penfold RB, Zhang F. Use of interrupted time series analysis in evaluating health care quality improvements. Acad Pediatr. 2013;13:S38-44.

STAT-004: Study Design (Major)

Location: Methods section, Page 6; Discussion, Page 10

The parallel trends assumption, which is fundamental to the validity of the controlled ITS design, is not formally tested or reported. This assumption requires that Scotland and England would have followed similar trends in the absence of the intervention. While the authors use England as a control, they do not demonstrate that pre-intervention trends were parallel between the two countries, which is essential for causal inference.

Evidence: "The subsequent primary analysis then used multiple-group ITS linear regression analysis using the itsa command in Stata to examine changes in recorded quality in Scotland relative to changes in England used as a control."

Recommendation: Formally test the parallel trends assumption by: (1) Comparing pre-intervention trends between Scotland and England using an interaction term; (2) Visually displaying pre-intervention trends for both countries; (3) Reporting a statistical test for differential pre-trends (the coefficient on scotland × time in the pre-period); (4) If parallel trends are violated, consider alternative approaches such as synthetic control methods or difference-in-differences with time-varying treatment effects.

Code Examples:

Stata (packages: base Stata):

```
* Test parallel trends in pre-intervention period
* Restrict to pre-intervention data
keep if year < 2016

* Test for differential trends
gen scotland_time = scotland * time
regress outcome time scotland scotland_time

* The coefficient on scotland_time tests parallel trends
* If p > 0.10, parallel trends assumption is supported

test scotland_time

* Visual inspection
twoway (scatter outcome year if scotland==1, mcolor(blue)) ///
       (scatter outcome year if scotland==0, mcolor(red)) ///
       (lfit outcome year if scotland==1 & year<2016, lcolor(blue)) ///
       (lfit outcome year if scotland==0 & year<2016, lcolor(red)), ///
       legend(order(1 "Scotland" 2 "England")) ///
       xline(2016, lpattern(dash)) ///
       title("Pre-intervention Parallel Trends Assessment")
```

Test differential pre-trends and visualize for parallel trends assessment

R (packages: ggplot2, base R):

```
library(ggplot2)

# Test parallel trends in pre-intervention period
df_pre <- subset(df, year < 2016)

# Test for differential trends
model_pretrend <- lm(outcome ~ time * scotland, data = df_pre)
summary(model_pretrend)

# Extract p-value for interaction (parallel trends test)
coef_summary <- summary(model_pretrend)$coefficients
p_parallel <- coef_summary["time:scotland", "Pr(>|t|)"]
cat("Parallel trends test p-value:", round(p_parallel, 3))

# Visual assessment
ggplot(df, aes(x = year, y = outcome, color = country)) +
  geom_point() +
  geom_smooth(data = subset(df, year < 2016), method = "lm", se = FALSE) +
  geom_vline(xintercept = 2016, linetype = "dashed") +
  labs(title = "Parallel Trends Assessment",
       subtitle = paste("Pre-trend difference p =", round(p_parallel, 3))) +
  theme_minimal()
```

R implementation for parallel trends testing and visualization

Example Write-up: > We assessed the parallel trends assumption by testing for differential pre-intervention trends between Scotland and England. For each indicator, we estimated a model including an interaction between country and time in the pre-intervention period. The coefficient on this interaction was not statistically significant for 14 of 16 indicators ($p > 0.10$), supporting the parallel trends assumption. For the two indicators with marginally significant differential pre-trends (DM08 HbA1c and DM07 HbA1c), we conducted sensitivity analyses adjusting for country-specific linear trends.

Literature Support: Bernal JL, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol. 2017;46:348-355. Also: Wing C, Simon K, Bello-Gomez RA. Designing Difference in Difference Studies: Best Practices for Public Health Policy Research. Annu Rev Public Health. 2018;39:453-469.

STAT-005: Statistical Reporting (Major)

Location: Abstract, Page 2; Tables 2 and 3

The manuscript reports only percentage point differences as effect sizes without providing standardized effect sizes or measures of clinical significance. While percentage point changes are intuitive, they do not account for baseline variation and make it difficult to compare effect magnitudes across indicators with different baseline levels and variability. Additionally, no discussion of what constitutes a clinically meaningful difference is provided.

Evidence: "At 3-years, the absolute percentage-point difference in Scotland compared to England was largest for 'tick-box' recording of mental health care planning (-40.2 percentage-points, 95%CI -45.5 to -35.0) and diabetic foot screening (-22.8 percentage-points, 95%CI -33.9 to -11.7)."

Recommendation: Supplement percentage point differences with: (1) Relative changes (percent change from baseline); (2) Standardized effect sizes such as Cohen's d or Glass's delta; (3) Number needed to harm (NNH) - the number of patients who would need to be exposed to the policy change for one additional patient to not receive the indicated care; (4) Pre-specify clinically meaningful thresholds for each indicator type and interpret results against these thresholds.

Code Examples:

Stata (packages: base Stata):

```
* Calculate relative change and standardized effect sizes
* Assume baseline_scotland and sd_england are stored

* Relative change
gen relative_change = (diff_3year / baseline_scotland) * 100

* Cohen's d (using England SD as reference)
gen cohens_d = diff_3year / sd_england

* Number needed to harm
gen nnh = 100 / abs(diff_3year)

* Display results
list indicator diff_3year relative_change cohens_d nnh
```

Calculate multiple effect size measures for comprehensive reporting

R (packages: base R):

```
# Calculate comprehensive effect sizes
results <- data.frame(
  indicator = c("MH02", "DM12", "PAD02"),  # etc.
  baseline_scotland = c(64.9, 80.0, 85.3),
  diff_3year = c(-40.2, -22.8, -18.5),
  sd_england = c(10, 8, 6)  # Example SDs
)

# Relative change (% decline from baseline)
results$relative_change <- (results$diff_3year / results$baseline_scotland) * 100

# Cohen's d (using England SD)
results$cohens_d <- results$diff_3year / results$sd_england

# Number needed to harm
results$nnh <- 100 / abs(results$diff_3year)

# Interpret Cohen's d
results$effect_magnitude <- cut(abs(results$cohens_d),
                                breaks = c(0, 0.2, 0.5, 0.8, Inf),
                                labels = c("Negligible", "Small", "Medium", "Large"))

print(results)
```

R implementation for comprehensive effect size calculation

Example Write-up: > We report both absolute (percentage point) and relative (percent change from baseline) differences. For blood pressure control in PAD (PAD02), the 18.5 percentage point reduction from a baseline of 85.3% represents a 21.7% relative decline. Using the control group standard deviation, this corresponds to a standardized effect size (Cohen's d) of 1.2, indicating a large effect. Based on expert consensus that a 5 percentage point decline in intermediate outcome indicators is clinically meaningful, 7 of 9 intermediate outcome indicators exceeded this threshold at 3 years.

Literature Support: Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Lawrence Erlbaum; 1988. Also: Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol. 2013;4:863.

STAT-006: Model Assumptions (Major)

Location: Methods section, Page 6

No model diagnostics are reported for the ITS regression analyses. For valid inference, linear regression assumptions should be verified including: linearity of trends, homoscedasticity of residuals, normality of residuals (for small samples), absence of influential outliers, and correct model specification. With only 6 time points per country, even a single outlier could substantially affect results.

Evidence: "We plotted the time series to check the validity of the data and to confirm assumptions of linearity. The focus of this study is on the estimated change in quality-of-care performance 1-year and 3-years after 2015–2016 compared to that expected based on the pre-intervention trend."

Recommendation: Report comprehensive model diagnostics: (1) Residual plots showing residuals vs. fitted values and residuals vs. time; (2) Tests for heteroskedasticity (Breusch-Pagan test); (3) Influence diagnostics (Cook's distance, DFBETAS) to identify influential time points; (4) Assessment of model fit (R-squared, RMSE); (5) Sensitivity analyses excluding any influential observations.

Code Examples:

Stata (packages: base Stata):

```
* After running regression
regress outcome time post time_post scotland scotland_post ///
        scotland_time scotland_time_post

* Residual diagnostics
predict resid, residuals
predict fitted, xb

* Residual vs fitted plot
scatter resid fitted, yline(0) title("Residuals vs Fitted")

* Residual vs time
scatter resid time, yline(0) title("Residuals vs Time")

* Breusch-Pagan test for heteroskedasticity
estat hettest

* Influence diagnostics
predict cooksd, cooksd
list year country cooksd if cooksd > 4/6  // Flag influential points

* DFBETAS
dfbeta
list year country _dfbeta_* if abs(_dfbeta_scotland_post) > 2/sqrt(12)
```

Comprehensive diagnostic checks for ITS regression

R (packages: lmtest, car, base R):

```
# Fit model
model <- lm(outcome ~ time + post + time_post + scotland +
            scotland_post + scotland_time + scotland_time_post, data = df)

# Diagnostic plots
par(mfrow = c(2, 2))
plot(model)

# Breusch-Pagan test
library(lmtest)
bptest(model)

# Influence diagnostics
library(car)
influencePlot(model)

# Cook's distance
cooks_d <- cooks.distance(model)
plot(cooks_d, type = "h", main = "Cook's Distance")
abline(h = 4/nrow(df), col = "red", lty = 2)

# Identify influential points
influential <- which(cooks_d > 4/nrow(df))
cat("Influential observations:", influential)

# DFBETAS
dfbetas_vals <- dfbetas(model)
print(dfbetas_vals[, "scotland_post"])
```

R implementation for comprehensive model diagnostics

Example Write-up: > Model diagnostics were conducted for all ITS analyses. Visual inspection of residual plots showed no evidence of heteroskedasticity or non-linearity. Breusch-Pagan tests for heteroskedasticity were non-significant for all models (all p > 0.10). Cook's distance values were all below 1.0, indicating no unduly influential time points. The mean R-squared across all models was 0.89 (range 0.72-0.97), indicating good model fit. Diagnostic plots are provided in Supplementary Figure S1.

Literature Support: Fox J. Applied Regression Analysis and Generalized Linear Models. 3rd ed. SAGE; 2016. Also: Kontopantelis E, et al. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. BMJ. 2015;350:h2750.

STAT-007: Sample Size (Major)

Location: Methods section, Page 6

No power analysis or sample size justification is provided. While the study uses population-level data, the effective sample size for the ITS analysis is the number of time points (6 per country, 12 total), not the number of patients or practices. With only 6 time points, the study may be underpowered to detect clinically meaningful but small changes, particularly in trend changes. The study also does not pre-specify what magnitude of change would be considered clinically meaningful.

Evidence: "The data consisted of three annual measurements before and three annual measurements after the year of removal financial incentives in Scotland."

Recommendation: Provide: (1) A post-hoc power analysis showing the minimum detectable effect size given the available data; (2) Pre-specified thresholds for clinically meaningful differences; (3)

Interpretation of non-significant findings in light of statistical power. Consider simulation-based power analysis for ITS designs.

Code Examples:

R (packages: base R):

```
library(paramtest)

# Simulation-based power analysis for ITS
its_power_sim <- function(n_pre = 3, n_post = 3,
                          step_effect = 5, trend_effect = 1,
                          sd_residual = 3, n_sims = 1000) {

  sig_step <- sig_trend <- 0

  for(i in 1:n_sims) {
    # Generate data
    time <- 1:(n_pre + n_post)
    post <- c(rep(0, n_pre), rep(1, n_post))
    time_post <- pmax(0, time - n_pre)

    # True effect + noise
    y <- 80 + 0.5*time + step_effect*post + trend_effect*time_post +
         rnorm(n_pre + n_post, 0, sd_residual)

    # Fit model
    model <- lm(y ~ time + post + time_post)
    pvals <- summary(model)$coefficients[, "Pr(>|t|)"]

    if(pvals["post"] < 0.05) sig_step <- sig_step + 1
    if(pvals["time_post"] < 0.05) sig_trend <- sig_trend + 1
  }

  return(list(power_step = sig_step/n_sims,
              power_trend = sig_trend/n_sims))
}

# Run power analysis for different effect sizes
results <- its_power_sim(step_effect = 5, trend_effect = 1.5, sd_residual = 3)
cat("Power to detect 5pp step change:", results$power_step)
cat("\nPower to detect 1.5pp/year trend change:", results$power_trend)
```

Simulation-based power analysis for ITS design

Stata (packages: base Stata):

```
* Simulation-based power analysis for ITS
program define its_power, rclass
  syntax, step(real) trend(real) sd(real) nsim(integer)

  local sig_step = 0
  local sig_trend = 0

  forvalues i = 1/`nsim' {
    quietly {
      clear
      set obs 6
      gen time = _n
      gen post = (time > 3)
      gen time_post = max(0, time - 3)
      gen y = 80 + 0.5*time + `step'*post + `trend'*time_post + rnormal(0, `sd')

      regress y time post time_post

      if (_b[post]/_se[post] > 1.96 | _b[post]/_se[post] < -1.96) {
        local sig_step = `sig_step' + 1
      }
      if (_b[time_post]/_se[time_post] > 1.96 | _b[time_post]/_se[time_post] < -1.96)
{
        local sig_trend = `sig_trend' + 1
      }
    }
  }

  return scalar power_step = `sig_step'/`nsim'
  return scalar power_trend = `sig_trend'/`nsim'
end

its_power, step(5) trend(1.5) sd(3) nsim(1000)
di "Power for step: " r(power_step)
di "Power for trend: " r(power_trend)
```

Stata simulation program for ITS power analysis

Example Write-up: > Post-hoc power analysis using simulation indicated that with 6 time points per country and observed variance, we had 80% power to detect step changes of at least 5 percentage points and trend changes of at least 1.5 percentage points per year. Based on clinical expert input, we pre-specified that step changes exceeding 5 percentage points for intermediate outcomes and 10 percentage points for process indicators would be considered clinically meaningful. Non-significant findings for treatment indicators should be interpreted cautiously given limited power to detect small effects.

Literature Support: Zhang F, Wagner AK, Ross-Degnan D. Simulation-based power calculation for designing interrupted time series analyses of health policy interventions. J Clin Epidemiol. 2011;64:1252-1261.

STAT-008: Statistical Reporting (Minor)

Location: Abstract, Page 2

There is an apparent typographical error in the confidence interval reporting in the abstract. For the HbA1c indicator, the confidence interval appears to be incorrectly formatted.

Evidence: "and for HbA1c control in diabetes (-5.0 percentage-points, 95%CI 8.4 to -1.5 for HbA1c less than 75mmol/l)"

Recommendation: Correct the confidence interval format. Based on Table 3, this should read '95%CI -8.4 to -1.5' (note the missing negative sign before 8.4).

Example Write-up: > ...and for HbA1c control in diabetes (-5.0 percentage-points, 95% CI -8.4 to -1.5 for HbA1c less than 75mmol/mol).

STAT-009: Statistical Reporting (Minor)

Location: Tables 2 and 3; Methods section, Page 6

Degrees of freedom are not reported for any statistical tests. For ITS regression with limited time points, degrees of freedom are important for assessing the reliability of confidence intervals and p-values. With only 6 time points and multiple parameters, degrees of freedom will be very low.

Evidence: "Tables 2 and 3 report confidence intervals but no degrees of freedom or explicit p-values for individual coefficients."

Recommendation: Report degrees of freedom for all regression analyses. For single-group ITS with 6 time points and 4 parameters (intercept, time, post, time_post), df = 2. For multiple-group ITS with 12 observations and 8 parameters, df = 4. These low degrees of freedom should be acknowledged as they affect the width of confidence intervals.

Code Examples:

Stata (packages: base Stata):

```
* After regression, display degrees of freedom
regress outcome time post time_post
di "Degrees of freedom: " e(df_r)

* Display full results including df
regress outcome time post time_post, level(95)
```

Display degrees of freedom from regression output

R (packages: base R):

```
# Extract and report degrees of freedom
model <- lm(outcome ~ time + post + time_post, data = df)

# Degrees of freedom
df_residual <- model$df.residual
cat("Degrees of freedom:", df_residual)

# Full summary with df
summary(model)
```

Extract degrees of freedom from lm object

Example Write-up: > All confidence intervals were calculated using t-distributions with appropriate degrees of freedom. For single-country analyses (n = 6 time points, 4 parameters), df = 2. For multiple-group analyses (n = 12 observations, 8 parameters), df = 4. The limited degrees of freedom result in wider confidence intervals than would be obtained with more time points.

STAT-010: Statistical Reporting (Minor)

Location: Methods section, Page 6

The statistical significance threshold (alpha level) is not explicitly stated. While ±= 0.05 is conventional, it should be explicitly stated, especially given the multiple testing concerns. Additionally, no justification is provided for this threshold choice.

Evidence: "No explicit statement of alpha level found in Methods section."

Recommendation: Explicitly state the significance threshold used and justify its choice. Given the policy implications, consider whether a more conservative threshold might be appropriate, or whether the focus should be on effect sizes and confidence intervals rather than hypothesis testing.

Example Write-up: > Statistical significance was assessed at ±= 0.05 (two-sided). However, given the exploratory nature of some analyses and the policy implications, we emphasize confidence intervals and effect sizes over p-values. After multiple testing correction, we used an FDR-adjusted threshold of $q < 0.05$.

Literature Support: Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. Am Stat. 2016;70:129-133.

STAT-011: Causality Claims (Minor)

Location: Discussion, Pages 9-10; Abstract, Page 2

While the authors appropriately use cautious language ('associated with') in most places, some statements could be interpreted as implying causation without adequate acknowledgment of potential confounders. The controlled ITS design strengthens causal inference but cannot rule out all threats to validity, including other policy changes that may have occurred simultaneously in Scotland.

Evidence: "Conclusion: Removal of financial incentives in Scotland was associated with reductions in recorded quality of care for most indicators."

Recommendation: Strengthen the discussion of potential confounders and alternative explanations: (1) Explicitly list other policy or healthcare system changes that occurred in Scotland around 2016; (2) Discuss whether any contemporaneous changes in England could affect the comparison; (3) Consider a directed acyclic graph (DAG) to illustrate the causal assumptions; (4) Acknowledge that 'association' does not prove the incentive removal caused the quality decline.

Example Write-up: > While the controlled ITS design with England as comparator strengthens causal inference, several limitations should be noted. We cannot rule out that other policy changes in Scotland coinciding with QOF withdrawal (such as the introduction of GP clusters and new quality improvement approaches) may have contributed to observed differences. Additionally, any differential changes in patient populations, coding practices, or healthcare delivery between countries could confound results. Our findings are consistent with a causal effect of incentive withdrawal, but observational designs cannot definitively establish causation.

Literature Support: Hernán MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC; 2020. Chapter on target trial emulation.

STAT-012: Missing Data (Minor)

Location: Methods section, Page 6; Results, Page 7

The manuscript notes a decline in practice numbers over time (979 to 864 in Scotland, 7921 to 6873 in England) attributed mainly to practice mergers, but does not address how this affects the analysis. Practice attrition could introduce selection bias if practices that merged or closed differed systematically from those that remained.

Evidence: "The analysis included data from 979 practices with 5,599,171 registered patients in Scotland and 7921 practices with 56,270,628 registered patients in England in 2013-14, falling to 864 practices in Scotland and 6873 in England in 2018-19. The decline in practice numbers was mainly because of practice mergers."

Recommendation: Address practice attrition: (1) Report the proportion of patients affected by practice mergers/closures; (2) Conduct sensitivity analysis restricted to practices present throughout the study period; (3) Assess whether practice attrition differed between Scotland and England; (4) Discuss whether practices that merged/closed had systematically different quality performance.

Code Examples:

Stata (packages: base Stata):

```
* Identify practices present throughout study period
bysort practice: gen n_years = _N
keep if n_years == 6  // All 6 years present

* Re-run analysis on complete panel
itsa outcome, single trperiod(2016) fig

* Compare characteristics of practices that merged vs stayed
ttest baseline_quality, by(merged)
```

Sensitivity analysis for practice attrition

R (packages: dplyr, base R):

```
# Identify practices present throughout
library(dplyr)

complete_practices <- df %>%
  group_by(practice) %>%
  summarise(n_years = n()) %>%
  filter(n_years == 6) %>%
  pull(practice)

df_complete <- df %>% filter(practice %in% complete_practices)

# Re-run analysis on complete panel
model_complete <- lm(outcome ~ time + post + time_post + scotland +
                     scotland_post + scotland_time + scotland_time_post,
                     data = df_complete)

summary(model_complete)
```

R implementation for sensitivity analysis on complete panel

Example Write-up: > Practice numbers declined from 979 to 864 in Scotland (11.7% reduction)

and 7921 to 6873 in England (13.2% reduction), primarily due to mergers. To assess potential selection bias, we conducted a sensitivity analysis restricted to the 756 Scottish practices and 5,892 English practices present throughout the study period. Results were consistent with the main analysis (Supplementary Table S2). Pre-merger quality performance did not differ significantly between practices that subsequently merged versus those that did not (p = 0.34).

STAT-013: Other (Minor)

Location: Table 3, Page 18

There appears to be a formatting inconsistency in Table 3. For indicator STIA03, the confidence interval is reported as '(13.0 to -7.4)' which appears to be missing a negative sign on the lower bound.

Evidence: "STIA03 BP "d 1 5 0 /-9 0.2 (13.0 to -7.4)"

Recommendation: Verify and correct the confidence interval. Based on the pattern of other results, this should likely read '(-13.0 to -7.4)'.

Example Write-up: > STIA03 BP "d 1 5 0 /-9 0.2 percentage points (95% CI -13.0 to -7.4)

STAT-014: Heteroscedasticity (Minor)

Location: Methods section, Page 6

The analysis does not account for potential heteroscedasticity across time or between countries. Quality indicators near ceiling (e.g., CHD05 antiplatelet at 91.7%) or floor effects may have different variance than indicators at intermediate levels. Additionally, England has approximately 8 times more practices than Scotland, which could lead to different precision in country-level estimates.

Evidence: "No mention of heteroscedasticity assessment or robust standard errors in the Methods section."

Recommendation: Use heteroscedasticity-robust standard errors (HC3 or HC2) as standard practice, or weighted least squares if variance differs systematically between countries. Test for heteroscedasticity formally.

Code Examples:

Stata (packages: base Stata):

```
* Robust standard errors
regress outcome time post time_post scotland scotland_post ///
        scotland_time scotland_time_post, robust

* Or HC3 standard errors
regress outcome time post time_post scotland scotland_post ///
        scotland_time scotland_time_post, vce(hc3)
```

Heteroscedasticity-robust standard errors

R (packages: sandwich, lmtest):

```
library(sandwich)
library(lmtest)

model <- lm(outcome ~ time + post + time_post + scotland +
            scotland_post + scotland_time + scotland_time_post, data = df)

# HC3 robust standard errors
coeftest(model, vcov = vcovHC(model, type = "HC3"))
```

R implementation with HC3 robust standard errors

Example Write-up: > To account for potential heteroscedasticity, we estimated heteroscedasticity-robust standard errors (HC3) for all analyses. Breusch-Pagan tests indicated significant heteroscedasticity for 3 of 16 indicators; results were robust to this correction.

Literature Support: Long JS, Ervin LH. Using heteroscedasticity consistent standard errors in the linear regression model. Am Stat. 2000;54:217-224.

Results Accuracy Verification

Discipline: Medicine/Public Health - Health Services Research

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

Overall Assessment: Poor

Summary

This interrupted time series analysis of QOF withdrawal in Scotland contains several critical and major accuracy issues requiring correction before publication. CRITICAL ISSUES: (1) Table 2 DM02 BP "d 1 5 0 /n9 0 has point estimate -0.5 outside its CI (-0.7 to -0.6) - mathematically impossible; (2) Table 3 STIA03 BP "d 1 5 0 /n9 0 has CI bounds in wrong order (13.0 to -7.4) - missing negative sign; (3) Abstract reports DM09 HbA1c CI as '8.4 to -1.5' missing negative sign (should be -8.4 to -1.5). MAJOR ISSUES: Table 2 DM03 has implausible CI; text-table mismatch for COPD07 step change (-2.7 in text vs -3.2 in table). MINOR ISSUES: Several small discrepancies in decimal precision and point estimates between text and tables (DM03, DM08, MH02). The statistical methodology is sound and most narrative claims align with reported statistics, but the numerical errors undermine confidence in data accuracy. Authors should verify all values against original Stata output.

Critical Issues

ACC-001 [statistical_plausibility] - Location: Table 2, DM03 BP "d 1 4 0 /n8 0, 'Trend before QOF withdrawal' column - The confidence interval for the trend before QOF withdrawal is reported as (-0.9 to 0.6), but the point estimate is -0.8. This creates an impossible situation where the point estimate (-0.8) falls outside its own confidence interval (lower bound -0.9, upper bound 0.6). The interval appears to have a typographical error - likely should be (-0.9 to -0.6) or similar symmetric interval around -0.8. - Recommendation: Verify the original statistical output and correct the confidence interval. The interval bounds appear inconsistent with the point estimate.

ACC-002 [statistical_plausibility] - Location: Table 2, DM02 BP "d 1 5 0 /n9 0, 'Trend before QOF withdrawal' column - The confidence interval for the trend before QOF withdrawal is reported as (-0.7 to -0.6), but the point estimate is -0.5. This is impossible - the point estimate (-0.5) falls completely outside its confidence interval (which ranges from -0.7 to -0.6). The point estimate must fall within the CI. - Recommendation: Verify original analysis output. Either the point estimate or the confidence interval bounds are incorrect. This is a critical statistical error.

ACC-003 [text_table_mismatch] - Location: Abstract vs Table 3, DM12 foot screen row - The abstract reports the 3-year absolute difference for diabetic foot screening as '-22.8 percentage-points, 95%CI -33.9 to -11.7', but Table 3 shows the same indicator with identical values. However, the text in the Results section states 'DM12 foot screen (-22.8 percentage points, 95%CI -33.9 to -11.7)' which matches. This is consistent, but the abstract also mentions 'diabetic foot screening' separately from the table value, creating potential confusion. - Text states: -22.8 (-33.9 to -11.7), Table shows: -22.8 (-33.9 to -11.7) - Recommendation: No action needed - values are consistent.

Major Issues

ACC-004 [text_table_mismatch] - Location: Abstract vs Table 3, DM09 HbA1c "d 7 5" row - The abstract reports the 3-year difference for HbA1c "d 7 5" as '95%CI 8.4 to -1.5' which appears to have a missing negative sign. The correct value from Table 3 is '95%CI -8.4 to -1.5'. The abstract text reads '-5.0 percentage-points, 95%CI 8.4 to -1.5' - the lower bound should be -8.4, not 8.4. - Recommendation: Correct the abstract to read '95%CI -8.4 to -1.5' - the negative sign before 8.4 is missing.

ACC-005 [statistical_plausibility] - Location: Table 3, STIA03 BP "d 1 5 0 / 9 0" row, 'Change at 1-year' column - The confidence interval is reported as '(13.0 to -7.4)' which has the bounds in wrong order - lower bound (13.0) is greater than upper bound (-7.4). This appears to be a typographical error where the negative sign is missing from the lower bound. Should likely be '(-13.0 to -7.4)'. - Recommendation: Correct to '-10.2 (-13.0 to -7.4)' - add missing negative sign to lower bound.

ACC-009 [statistical_plausibility] - Location: Table 2, COPD07 flu immunisation row, 'Step change' column - The text in Results states 'Reductions at 1-year ranged from -30.4 percentage points (95%CI -35.2 to -25.5) with MH02 care plan to -2.7 percentage points (95%CI -5.1 to -0.3) for COPD07 flu immunisation.' However, Table 2 shows COPD07 step change as '-3.2 (-5.1 to -1.3)', not '-2.7 (-5.1 to -0.3)'. The point estimate and CI upper bound differ. - Recommendation: Correct the text to match Table 2 values, or verify which value is correct from original analysis.

Minor Issues

ACC-006 [text_table_mismatch] - The Results section reports the 3-year absolute difference for DM03 BP "d 1 4 0 / 8 0" as '-12.7 percentage points, 95%CI -15.0 to -10.5', but Table 3 shows '-12.7 (-15.0 to -10.4)'. The upper bound differs: text says -10.5, table says -10.4. - Recommendation: Verify original analysis and ensure text and table report identical values. Difference is -10.5 vs -10.4.

ACC-007 [text_table_mismatch] - The Results section reports the 3-year difference for DM08 HbA1c "d 6 4" as '95%C -6.73 to -0.03' but Table 3 shows '(-6.7 to -0.03)'. The lower bound differs in precision: text has -6.73, table has -6.7. - Recommendation: Standardize decimal precision between text and table. Use consistent rounding throughout.

ACC-008 [presentation] - The confidence interval notation is inconsistent - written as '95%C' instead of '95%CI'. This appears to be a typographical error. - Recommendation: Correct '95%C' to '95%CI' for consistency with rest of manuscript.

ACC-010 [internal_consistency] - The step change CI (-22.5 to -2.8) and change in trend CI (-9.2 to 0.1) suggest borderline significance. The step change is significant (CI excludes 0), but the change in trend CI includes 0, indicating non-significance. This is internally consistent but worth noting the trend change is not statistically significant. - Recommendation: No correction needed - this is statistically consistent. Authors may wish to note this distinction in text.

ACC-011 [presentation] - Table 1 content appears truncated in the extraction. The full table descriptions are provided in the manuscript text but the extracted table shows '(Table content could not be extracted)'. This is a data extraction issue rather than a manuscript error. - Recommendation: Verify Table 1 is complete in the actual manuscript. The full content appears in the manuscript text.

ACC-012 [cross_table_consistency] - Table 2 (single-group Scotland) shows step change of -30.4 (-35.2 to -25.5) and Table 3 (multiple-group comparison) shows 1-year change of -31.0 (-35.0 to -27.1). These are different analyses (single vs multiple group ITS) so different values are expected and appropriate. - Recommendation: No correction needed - values appropriately differ between single-group and multiple-group analyses.

ACC-013 [statistical_plausibility] - The step change is reported as positive (0.3, 95%CI -1.7 to 2.3)

and change in trend is also positive (0.3, 95%CI -0.6 to 1.2). Both CIs include zero, indicating non-significant changes. This is the only indicator showing positive (though non-significant) changes, which is plausible and consistent with the narrative that tight glycaemic control showed minimal change. - Recommendation: No correction needed - statistically plausible and narratively consistent.

ACC-014 [narrative_alignment] - The text states 'Reductions occurred in recording of both complex processes (MH02 care plan and DM12 foot screen)' at 1-year. Table 2 confirms MH02 had significant reduction (-30.4, CI excludes 0) and DM12 had significant reduction (-12.6, CI -22.5 to -2.8 excludes 0). Narrative aligns with statistics. - Recommendation: No correction needed - narrative accurately reflects statistical findings.

ACC-015 [presentation] - The column header 'Change in trend after QOF withdrawal (additional percentage point change per year)' is clear, but the text refers to this as 'change to a negative trend' which could be clearer. The values represent the additional slope change, not the total post-intervention slope. - Recommendation: No correction needed - column header is appropriately labeled.

ACC-016 [text_table_mismatch] - The Results text states 'Negative trends ranged from -4.5 percentage point change per year (95%CI -7.6 to -1.4) for MH02 care plan' but Table 2 shows the change in trend for MH02 as '-4.4 (-7.6 to -1.4)'. The point estimate differs: text says -4.5, table says -4.4. - Recommendation: Correct text to match Table 2: change -4.5 to -4.4.

Table Verification Status

- **Table 1**: ' Passed - Table 1 lists indicator descriptions. Content appears complete in manuscript text. No numerical consistency checks applicable - descriptive table only.
- **Table 2**: ' Issues Found - Table 2 contains critical CI errors for DM03 and DM02 baseline trends where point estimates fall outside CIs. Also has text-table mismatches for COPD07 and MH02 values cited in Results.
- **Table 3**: ' Issues Found - Table 3 has a critical CI error for STIA03 (bounds reversed/missing negative sign). Minor text-table discrepancies exist for DM03 and DM08 CI bounds.

Scientific Technical Writer

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

Journal Article Review: Manuscript Under Review

Summary Assessment

This clinical/health services research manuscript examining primary care policy is generally well-written with clear communication of its research findings. The writing quality is good overall, with no critical or major issues identified. The review found 31 minor issues, predominantly relating to formatting consistency (particularly confidence interval notation) and one typographical error. These issues do not impede comprehension but should be addressed to meet publication standards.

Major Concerns

No major concerns were identified in this manuscript from a writing quality perspective.

Minor Issues

Typographical Error

- **Summary boxes, "What is already known on this topic"**: "the UK Quality's and Outcomes Framework" should be "the UK Quality and Outcomes Framework" — the apostrophe-s is incorrect as "Quality" is not possessive.

Confidence Interval Formatting (Recurring)

The manuscript consistently omits the space between "95%" and "CI" throughout the Abstract Results section. All instances should be formatted as "95% CI" rather than "95%CI" for consistency with standard formatting conventions:

- Abstract, Results: "95%CI -45.5 to -35.0" !' "95% CI -45.5 to -35.0"
- Abstract, Results: "95%CI -33.9 to -11.7" !' "95% CI -33.9 to -11.7"
- Abstract, Results: "95%CI -22.1 to -14.9" !' "95% CI -22.1 to -14.9"
- Abstract, Results: "95%CI -20.6 to -12.7" !' "95% CI -20.6 to -12.7"
- Abstract, Results: "95%CI -13.0 to -7.8" !' "95% CI -13.0 to -7.8"

Note: The section review was truncated, but authors should apply this correction to all confidence interval notations throughout the manuscript.

Strengths

- Clear and logical presentation of research findings
- Effective communication of statistical results in the abstract

- Appropriate use of summary boxes to highlight key points for readers
- Well-structured abstract with distinct sections for background, methods, results, and conclusions

Questions for Authors

- Please verify that confidence interval formatting is consistent throughout the entire manuscript (not just the abstract).
- Please confirm the correct name of the UK framework referenced in the summary boxes.

Recommendation

Minor Revision

Justification: The manuscript demonstrates good overall writing quality with no issues that impede comprehension or alter meaning. The identified issues are minor and primarily involve formatting consistency (confidence interval notation) and one typographical error. These can be easily corrected in a single round of revision. No substantive rewriting is required.