AI Manuscript Review

Manuscript

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

Editorial Decision: PENDING

EXECUTIVE SUMMARY

This large population-based cohort study provides valuable risk estimates for adverse events following common arthroscopic shoulder procedures. The study's strengths include its large sample size and a novel, clinically significant finding regarding postoperative pneumonia risk. However, the manuscript requires major revision before it can be accepted. The consensus among reviewers is that there are significant issues with the statistical methodology, most notably a lack of correction for multiple comparisons. Furthermore, the authors make causal inferences that are not supported by the observational study design. The presentation also needs significant improvement, as the key clinical messages are currently obscured by an overload of statistics and numerous numerical inconsistencies between the text and tables. The work is promising, but these fundamental issues must be addressed to ensure the validity and impact of the findings.

DECISION LETTER

Editorial Summary: [Manuscript Title]

Manuscript ID: bMpbXP1FsgiX13slUMVQ **Date:** 2024-06-18 **Editor:** AI Editor-in-Chief **Number of Reviews:** 7

Overview of Reviews

The manuscript has been evaluated by a comprehensive panel of seven reviewers, including domain experts, methodologists, and automated verification agents. There is a strong consensus that this large population-based study addresses an important and clinically relevant question. The use of a national dataset is a significant strength. However, there is also a clear consensus among the narrative reviewers (Skeptic, Pragmatic, Systematic) that the manuscript requires substantial revision before it can be considered for publication. While some reviewers recommended minor revisions or acceptance, the major issues raised by multiple independent reviewers regarding statistical methodology, causal inference, and clarity of presentation must be addressed.

Points of Consensus

Several critical issues were independently identified by multiple reviewers, highlighting them as the highest priority for revision.

1. **Unsupported Causal Inference** — Raised by Reviewers [Adversarial Skeptic, Systematic Reviewer, Statistical Methods Agent] - The manuscript makes causal attributions that are not supported by the retrospective cohort design. Specifically, the claim that the temporal decline in ASAD procedures is due to the "impact of...trials" and the direct comparison of adverse event rates with knee arthroscopy are inappropriate without a suitable study design (e.g., interrupted time-series) or formal statistical adjustment for confounding.

2. **Lack of Correction for Multiple Statistical Comparisons** — Raised by Reviewers [Systematic Reviewer, Statistical Methods Agent] - The study designates seven "primary outcomes" and conducts numerous statistical comparisons without adjusting for the inflated risk of Type I error (false positives). This is a significant methodological flaw that undermines the statistical validity of the findings.

3. **Insufficient Methodological Detail** — Raised by Reviewers [Domain Expert, Statistical Methods Agent] - Key methodological details are either missing from the main text or inadequately described. This includes the procedural classification hierarchy, the specifics of the statistical adjustment models, and the lack of reporting on model diagnostics (e.g., calibration, discrimination) or accounting for data clustering by hospital or patient.

4. **Poor Clarity and Data Presentation** — Raised by Reviewers [Pragmatic Reviewer, Results Accuracy Verification Agent, Scientific Technical Reviewer] - The abstract and results are overloaded with statistics, obscuring the key clinical messages. Furthermore, numerous numerical inconsistencies were found between the manuscript text and the tables, alongside inconsistent formatting of numbers and abbreviations, suggesting a need for careful proofreading.

Points of Divergence

The primary point of divergence was the final recommendation, which ranged from "Accept" to "Reject". The Statistical Methods and Results Accuracy Verification agents provided recommendations ("Accept" and "Reject", respectively) that were inconsistent with the severity of the issues they identified. The "Accept" recommendation from the Statistical Methods Agent was particularly incongruous with its identification of six major statistical flaws. In contrast, the narrative reviewers (Skeptic, Pragmatic, Systematic) were in agreement that the identified issues, while significant, were correctable and warranted a "Revise and Resubmit" decision.

- **Editor's assessment:** I am siding with the narrative reviewers. The issues raised are major and preclude acceptance in the current form. However, they are addressable through careful revision. The recommendation from the Statistical Methods Agent appears to be an error, and its detailed, high-quality critique will be treated as the primary contribution. The numerical errors found by the Verification Agent are serious but correctable, making rejection too severe.

Required Revisions

The following revisions are mandatory for the manuscript to be reconsidered for publication.

1. **Address Multiple Comparisons:** You must revise your statistical analysis plan. Either designate a single, pre-specified primary outcome and treat all others as secondary/exploratory, or apply a valid statistical correction for multiple comparisons (e.g., Bonferroni, Benjamini-Hochberg) to all primary outcomes and justify your choice. — Source: Reviewer(s) [Systematic Reviewer, Statistical Methods Agent]

2. **Remove Unsupported Causal Claims:** Rephrase the manuscript to remove all language that implies causality. The discussion of temporal trends for ASAD must be framed as an observation of an association, not an impact of trials. The comparison to knee arthroscopy must be presented cautiously as a hypothesis-generating comparison of unadjusted rates, not a formal assessment of differential risk. — Source: Reviewer(s) [Adversarial Skeptic, Systematic Reviewer, Statistical Methods Agent]

3. **Enhance Methodological Transparency:** The main Methods section must be expanded to include a clear and complete description of the procedural hierarchy, the full specification of the logistic regression models, and how you accounted for clustering of outcomes (e.g., bilateral procedures within patients, patients within hospitals). You must also report standard model diagnostics (e.g., C-statistic, calibration plots). — Source: Reviewer(s) [Domain Expert, Statistical Methods Agent]

4. **Improve Clarity and Readability:** The Abstract must be rewritten to emphasize the key clinical findings and their implications, rather than presenting a dense list of statistics. The most important takeaways for a general clinical audience should be immediately apparent. — Source: Reviewer(s) [Pragmatic Reviewer]

5. **Correct All Numerical Inconsistencies:** A thorough audit of the manuscript is required to ensure that all numbers, percentages, and confidence intervals reported in the text are identical to those in the tables and supplementary materials. All numerical formatting must be made consistent. — Source: Reviewer(s) [Results Accuracy Verification Agent, Scientific Technical Reviewer]

Recommended Revisions

These revisions are strongly encouraged to improve the manuscript's impact and robustness.

- **Define Terminology:** Provide clear, concise definitions for the different types of shoulder procedures in the main text to ensure the manuscript is accessible to a non-specialist audience. — Source: Reviewer [Pragmatic Reviewer]

- **Contextualize Novel Findings:** The discussion of the unexpected finding of postoperative pneumonia should be expanded to cautiously explore potential mechanisms, framing this clearly as a hypothesis for future investigation. — Source: Reviewer [Adversarial Skeptic, Pragmatic Reviewer]

Optional Suggestions

Authors may consider these suggestions to further strengthen the paper.

- **Clarify Acronyms:** For an international audience, provide a brief clarifier for "Hospital Episode Statistics (HES)". — Source: Reviewer [Pragmatic Reviewer]

- **Address Minor Writing Issues:** Please address the minor points of writing style, formatting, and abbreviation consistency raised by the Scientific Technical Reviewer.

Editor's Comments to Authors

The reviewers and I agree that your study addresses an important clinical question using a valuable, large-scale dataset. The work has clear potential to contribute significantly to the literature on surgical safety. However, the manuscript in its current form contains several major, interconnected issues spanning statistical methodology, interpretation, and presentation that must

be resolved.

The required revisions are substantial but achievable. Your primary focus should be on ensuring the statistical analysis is robust and transparent, particularly regarding the issue of multiple comparisons and the handling of clustered data. Equally important is reframing the manuscript's claims to align strictly with what your observational data can support.

We hope you will find the reviewers' detailed comments constructive. We believe that if you can fully address the required revisions, the resulting manuscript will be a strong candidate for publication. We look forward to receiving a revised version along with a point-by-point response to the issues raised.

Editorial Decision

Decision: Revise and Resubmit

Rationale: The manuscript presents a clinically important study using a powerful dataset, but it is undermined by major, correctable flaws. Key concerns shared across multiple reviewers include a failure to adjust for multiple statistical comparisons, causal claims that overstep the observational study design, a lack of essential methodological detail, and numerous numerical inconsistencies. A major revision is required to address these fundamental issues of statistical validity, interpretation, and clarity before the manuscript can be reconsidered for publication.

REQUIRED CHANGES

1. Address the lack of adjustment for multiple comparisons by either specifying a single primary outcome or applying a formal statistical correction.

2. Remove all language that implies causality from the observational data, particularly regarding temporal trends and comparisons to other surgical procedures.

3. Expand the Methods section in the main text to provide full transparency on the procedural classification hierarchy and the statistical models used, including how data clustering was addressed and the results of model diagnostics.

4. Rewrite the abstract to prioritize clear, clinical messages over a dense list of statistics to improve accessibility and impact.

5. Conduct a full review of the manuscript to find and correct all numerical inconsistencies between the text and tables.

SUGGESTED IMPROVEMENTS

1. Provide clear definitions of the different surgical procedures for a non-specialist audience.

2. Expand the discussion of the novel pneumonia finding, cautiously exploring potential mechanisms while framing it as a hypothesis.

3. Add a brief description of the Hospital Episode Statistics (HES) dataset for international readers.

Domain Expert

Model: gemini/gemini-2.5-pro

Journal Article Review: Manuscript Under Review

Summary Assessment

This is a well-motivated and important population-based cohort study that leverages a large national dataset to provide much-needed, precise risk estimates for common arthroscopic shoulder procedures. The study's key strengths are its massive sample size and the novel, clinically significant finding regarding postoperative pneumonia risk. However, the manuscript's conclusions are undermined by a lack of critical methodological detail in the main text regarding how procedures were classified and how statistical adjustments were performed.

Major Concerns

1. **Insufficient Methodological Detail on Procedure Classification and Statistical Adjustment** — Methods Section:

- **Problem**: The manuscript's core findings depend on two critical methodological steps that are not adequately described in the main text, hindering the reader's ability to assess the validity of the results.
- **Procedural Hierarchy**: The paper states, "Procedures were grouped for analysis according to the main therapeutic procedure code recorded, based on a hierarchy described in the supplementary file." As many shoulder surgeries involve multiple concurrent procedures (e.g., rotator cuff repair with subacromial decompression), the logic of this hierarchy is fundamental to interpreting the procedure-specific risks. Relegating this core definitional logic to a supplement is insufficient; without it, the reader cannot determine if the analytic groups are clinically meaningful or artifactual.
- **Statistical Adjustment**: The description of the statistical analysis is incomplete, cutting off at "After adjustment for age, comorbid...". The novel comparison between shoulder and knee arthroscopy is a key finding, but its validity rests entirely on the appropriateness of the statistical adjustment. The manuscript must explicitly detail all covariates included in the models to allow for proper evaluation of potential confounding.
- **Recommendation**: - In the main Methods section, provide a concise summary of the procedural classification logic. For instance: "Procedures were classified based on a pre-specified hierarchy of clinical complexity, where rotator cuff repair was considered the primary procedure if coded concurrently with less complex procedures like subacromial decompression or ACJ excision." - Complete the description of the statistical models, explicitly listing all variables used for adjustment in the main text.

Strengths

- **Large Population-Based Cohort**: The use of the English national hospital dataset (HES) provides a massive sample size, allowing for robust and precise risk estimation, particularly for rare adverse events.
- **Clinically Relevant Research Question**: The study addresses a genuine evidence gap, providing data that is essential for patient counseling, the informed consent process, and shared decision-making.
- **Novel and Important Finding**: The identification of an elevated pneumonia risk following shoulder arthroscopy compared to knee arthroscopy is a novel and clinically important finding that warrants further investigation and could influence perioperative protocols.

Recommendation

Major Revision

Justification: The study is based on a powerful dataset and addresses a clinically important question with a novel finding. However, the manuscript in its current form lacks the essential methodological detail required for a reader to independently assess the validity of its central claims. The concerns regarding the procedural hierarchy and statistical adjustment are significant but addressable through textual revisions without requiring new analysis. Once these methodological descriptions are clarified in the main manuscript, the paper will represent a substantial contribution to the literature.

Adversarial Skeptic

Model: openai/gpt-5.2

Journal Article Review: Manuscript Under Review

Summary Assessment

The manuscript has a clear aim and appears to use an appropriate large-scale dataset for estimating *inpatient-coded* serious adverse events. However, the central interpretive claim about why procedure volumes changed over time overreaches what the described design can support. Several essential methodological details (e.g., procedure classification, risk modeling, and comparator validity) are referenced as important but are not available in the provided excerpt, limiting the ability to fully evaluate internal validity.

Major Concerns

1. **Attribution/causal overreach for temporal trends ("trials caused the decline")**

- **Location**: Abstract, Paragraph 3; "What this study adds," Paragraph 4 - **Quote**: "**The temporal drop in ASAD numbers observed is in keeping with the impact of conducting high-quality national trials on its effectiveness.**" - **Challenge**: The authors imply attribution from a temporal pattern ("impact of…trials") without presenting (in the provided Methods excerpt) a design that can separate trial impact from other contemporaneous drivers. *What if* the apparent decline reflects changes in commissioning/policy, reimbursement incentives, capacity constraints, guideline updates, media coverage, or—particularly concerning given the noted coding-update context—ascertainment/coding artifacts rather than "trial impact"? - **Why it matters**: This is not a cosmetic wording issue: attributing changes to "high-quality national trials" invites causal interpretation and downstream policy conclusions. If the decline is partly (or largely) explained by non-trial factors, the manuscript's main "impact" narrative is overstated and potentially misleading. - **Fix needed**: 1) **Reframe the claim** to a strictly descriptive, non-causal statement (e.g., "coincided with," "temporally aligned with") unless stronger identification is added. 2) If the authors wish to retain an attribution claim, add a **formal quasi-experimental analysis** (e.g., interrupted time series with *pre-specified* interruption points tied to trial publication/implementation; controls/negative controls such as other shoulder procedures unlikely to be affected by the trials; sensitivity analyses to alternative interruption dates; modeling of autocorrelation/seasonality). 3) Explicitly address the risk that **coding or recording changes** could produce spurious "drops," including sensitivity analyses around any coding-update boundary.

Minor Issues

Because only an excerpt and one verbatim manuscript quote were provided, I cannot responsibly enumerate the full set of minor, line-level issues referenced in the prompt (19 minor issues) without inventing text. If you share the remaining section-by-section reviews (or the full manuscript text), I can consolidate those minor items into this unified format.

Strengths

- Clear clinical/service-relevant question and an apparently suitable dataset for capturing *inpatient-coded* serious adverse events at scale.
- The manuscript (as excerpted) is readable and structured in a way that should support translation to practice, provided causal language is calibrated to the design.

Questions for Authors

1. What is the explicit identifying strategy for linking temporal changes in ASAD use to the conduct/publication/implementation of trials (as opposed to contemporaneous policy/guideline/capacity/coding changes)?
2. Were any coding updates, tariff changes, or commissioning guidance introduced during the study window that could alter recorded ASAD volumes or adverse event capture? How were these handled analytically?
3. Please specify (ideally in the main text, not only supplements) the procedure classification rules, risk-adjustment model specification, and how comparator groups (if any) were selected/validated.

Recommendation

Major Revision

Justification: The excerpted manuscript makes a key interpretive claim that reads as causal attribution from temporal association without (in the provided Methods) a design capable of supporting that attribution. This is correctable via (i) reframing to non-causal language and/or (ii) adding an appropriate interrupted time-series/control-series framework and sensitivity analyses, alongside clearer reporting of key methodological details.

Systematic Reviewer

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

Journal Article Review: Manuscript Under Review

Summary Assessment

This large retrospective cohort study examining adverse events following arthroscopic shoulder surgery in England addresses an important clinical question with a substantial dataset (288,250 procedures over 8 years). The study leverages Hospital Episode Statistics data effectively and has the potential to inform clinical practice.

However, significant methodological concerns regarding multiple hypothesis testing and statistical reporting require resolution before publication. The manuscript would benefit from clearer specification of the primary outcome, appropriate adjustment for multiple comparisons, and more detailed methodological transparency regarding temporal trend analysis and adjusted analyses.

Major Concerns

Multiple Comparisons Without Appropriate Statistical Correction

1. Problem with Primary Outcome Designation

- **Problem**: Seven outcomes are designated as "primary outcomes" without any mention of adjustment for multiple testing. This designates all seven as co-primary, which substantially inflates Type I error risk.
- With 7 comparisons at ± = 0 . 05, the probability of at least one false positive is approximately 30% (familywise error rate = 1 - (1-0.05)^7 "H 0.30). Standard practice dictates either a single pre-specified primary outcome or explicit correction for multiplicity when testing multiple endpoints.
- **Recommendation**: - Option 1 (Preferred): Designate ONE outcome as the primary outcome (e.g., the most clinically important or the one used for sample size calculation) and explicitly label the other six as secondary or exploratory outcomes. - Option 2: If all seven outcomes are truly co-primary, apply and report appropriate multiple comparison correction. Add to Methods: "To control the familywise error rate across seven primary outcomes, we applied [Bonferroni/Holm-Bonferroni/false discovery rate] correction. The adjusted significance threshold was ±= [0.007 for Bonferroni; report actual values for other methods]." Report both unadjusted and adjusted p-values in results tables. - Provide justification for the chosen approach in the statistical analysis section.
- **Severity**: MAJOR — This directly affects the validity of statistical inference and the interpretation of findings. Without correction, some "significant" findings may be false positives.

Note: The issue summary indicates 4 additional major issues that were not included in the provided section review text. These should be addressed by the authors upon receipt of the complete review.

Minor Issues

[The section review indicates 18 minor issues were identified, but details were not included in the provided text. These should be communicated to the authors in the complete review.]

Strengths

- **Robust sample size**: The study analyzes 288,250 procedures in 261,248 patients, providing substantial statistical power for detecting clinically meaningful differences in adverse event rates.
- **Comprehensive data source**: Use of Hospital Episode Statistics (HES) data over an 8-year period provides population-level coverage and minimizes selection bias inherent in single-center studies.
- **Clinically relevant question**: The study addresses an important gap in understanding adverse events following a common surgical procedure, with direct implications for patient counseling and clinical decision-making.
- **Multiple adverse outcomes assessed**: The comprehensive assessment of seven clinically important adverse events (mortality, PE, respiratory infection, MI, AKI, stroke, UTI) provides a thorough safety profile.

Questions for Authors

1. Was a single primary outcome pre-specified in a study protocol or analysis plan? If so, which outcome was designated as primary and what was the rationale?
2. Was a formal sample size calculation performed? If so, which outcome was used for the power calculation and what effect size was considered clinically meaningful?
3. Were any adjustments made for multiple testing in the analysis? If adjustments were made but not reported, please provide adjusted p-values.
4. How were temporal trends in adverse event rates over the 8-year study period assessed and controlled for in the analysis?
5. What variables were included in adjusted analyses? Please provide complete details of all covariates included in multivariable models.

Recommendation

Major Revision

Justification: The study leverages an excellent dataset to address an important clinical question and appears methodologically sound in most respects. However, the critical issue of multiple comparisons without adjustment must be resolved, as this directly affects the validity of statistical conclusions.

This is addressable through either reframing the hierarchy of outcomes (primary vs. secondary) or applying appropriate statistical corrections. The revision required is substantive but feasible, and the underlying study has merit.

Once the multiple testing issue and other methodological clarifications (noted in the issue summary but not detailed in the provided text) are addressed, this manuscript would make a valuable contribution to the surgical literature.

Pragmatic Reviewer

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

Journal Article Review: Manuscript Under Review

Summary Assessment

This population-based cohort study examining adverse events in 261,248 patients undergoing arthroscopic shoulder surgery addresses an important knowledge gap with robust methods and comprehensive data. The work has clear clinical value and the findings warrant publication. However, the presentation significantly undermines accessibility and impact. The abstract and results sections suffer from statistical overload that buries clinically meaningful messages, making it difficult for non-specialist readers (including referring clinicians) to extract actionable information. The most striking findings—particularly the 3-fold higher pneumonia rate compared to knee arthroscopy and the 1-in-26 reoperation rate—deserve more prominent explanation and mechanistic discussion.

Major Concerns

1. **Abstract Statistical Overload Obscures Key Clinical Messages** — Abstract, Results paragraph: - Problem: The abstract presents eight different confidence intervals in rapid succession ("0.64% (0.54 to 0.76)...1.65% (1.48 to 1.83)...0.33% (0.31 to 0.35)...0.07%...3.82% (3.75 to 3.90)...2.73% (2.50 to 2.98)..."), making it impossible for readers to identify the headline findings. The most clinically significant message—that approximately 1 in 26 patients require reoperation within 12 months—is buried in this numerical barrage. A GP deciding whether to refer a patient for shoulder arthroscopy cannot extract usable information from this density of statistics. - Recommendation: Restructure to lead with headline findings in plain language, then provide supporting detail. Suggested revision: "Within 90 days of surgery, complications occurred in approximately 1 in 80 patients (1.23%), with pneumonia being the most common problem (1 in 300 patients). Surprisingly, pneumonia rates were three times higher than in knee arthroscopy patients. By 12 months, nearly 1 in 26 patients (3.82%) required reoperation, with rates varying by procedure type. The most common complications were... [then provide key confidence intervals for 2-3 most important findings only]." This approach makes the clinical significance immediately clear while preserving statistical rigor.

2. **Insufficient Contextualization of Procedure Types for Non-Specialist Readers** — Introduction and Methods: - Problem: The study compares adverse event rates across different shoulder procedures (glenohumeral stabilisation, frozen shoulder release, subacromial decompression, etc.), but assumes readers understand what these procedures involve, why they might have different risk profiles, and which patients typically undergo each procedure. Non-specialist readers in general medical journals cannot interpret why frozen shoulder release carries higher risk than stabilisation without basic context about procedural invasiveness, patient populations, or anatomical approaches. - Recommendation: Add 2-3 sentences to the Introduction contextualizing the main procedure types: "Shoulder arthroscopy encompasses procedures ranging from relatively simple diagnostic arthroscopy and subacromial decompression (removing bone spurs to relieve impingement) to more complex procedures such as frozen shoulder release (dividing scar tissue) and rotator cuff repair. These procedures differ substantially in duration, tissue trauma, and typical patient age and comorbidity profiles, potentially leading to different risk profiles." This allows readers to interpret the comparative findings meaningfully.

3. **Unexplained Mechanistic Finding: 3-Fold Higher Pneumonia Rates** — Results and Discussion: - Problem: The finding that pneumonia rates are three times higher following shoulder arthroscopy compared to knee arthroscopy is clinically striking and counterintuitive (neither procedure directly involves the chest). This demands mechanistic explanation, yet the manuscript appears to present this as a descriptive finding without exploring potential causes. Is this due to positioning (shoulder procedures typically performed in beach-chair position affecting respiratory mechanics)? Patient selection (older, more comorbid patients for shoulder vs. knee)? Post-operative immobilization? Without addressing this, readers cannot judge whether this represents a true procedural risk or confounding. - Recommendation: Add a dedicated paragraph to the Discussion exploring potential mechanisms: "The elevated pneumonia risk following shoulder arthroscopy warrants careful interpretation. Potential contributing factors include: (1) Patient characteristics—shoulder surgery patients tend to be older with more comorbidities [cite data from your Table 1]; (2) Positioning—the beach-chair or lateral decubitus position may compromise respiratory mechanics during surgery; (3) Post-operative factors—shoulder immobilization may limit deep breathing and coughing effectiveness. Our data suggest [X] is most likely because [evidence from subgroup analyses or comparisons]. Clinicians should consider pneumonia prevention strategies particularly in [high-risk subgroups]." This transforms a puzzling finding into actionable clinical insight.

4. **Reoperation Findings Lack Clinical Context** — Results presentation: - Problem: The 3.82% overall reoperation rate and procedure-specific variations are presented as raw percentages without clinical context. What constitutes "expected" vs. "concerning" reoperation rates for these procedures? Are these rates consistent with individual surgeon series, or do they reveal a quality gap? Without benchmarks, readers cannot judge whether 2.73% reoperation after stabilisation represents good outcomes or room for improvement. - Recommendation: Add contextualizing sentences to Results or Discussion: "The observed reoperation rate of 3.82% at 12 months compares to [X%] reported in registry data from [country] and [Y%] in meta-analyses of RCTs. The higher rates observed for [procedure type] align with surgical series reporting [Z%], suggesting this reflects known technical challenges rather than systematic quality issues. However, the secular trend showing [increasing/decreasing] reoperation rates warrants further investigation into whether this represents changing patient selection, surgical indications, or quality of care."

5. **"What This Study Adds" Box Understates Impact** — Context box: - Problem: The "What this study adds" section lists contributions as bland descriptive statements ("This study reports complication rates...") that fail to convey why this matters clinically. The transformative aspect—that this provides the first population-level benchmarks for informed consent conversations and quality monitoring—is not articulated. - Recommendation: Reframe to emphasize practical impact: - "Provides population-level complication benchmarks that clinicians can use during informed consent discussions (e.g., 'About 1 in 26 patients need another operation within a year')" - "Identifies pneumonia as an under-recognized risk following shoulder arthroscopy, occurring three times more frequently than after knee arthroscopy" - "Enables hospitals to benchmark their outcomes against national rates for quality improvement" - "Reveals substantial variation in reoperation rates by procedure type, informing surgical decision-making"

Minor Issues

- **Abstract, Methods**: "Population-based cohort study using Hospital Episode Statistics" assumes international readers know what HES is. Add brief clarifier: "...using Hospital Episode Statistics (the mandatory administrative database covering all NHS hospital care in England)."

- **Abstract, Results**: The phrase "ranging from 2.73% (2.50 to 2.98)..." ends mid-sentence without completing the range. Presumably this should read "ranging from 2.73%...to X% following [procedure]."

- **Introduction**: Consider adding one sentence on why now: "With shoulder arthroscopy rates increasing substantially over the past two decades, population-level safety data are increasingly important for informed decision-making."

- **Methods, Data Source**: Clarify what "finished consultant episodes" means for non-UK readers: "finished consultant episodes (complete hospital admissions under a single consultant's care)."

- **Methods, Study Period**: Justify the 2001-2017 timeframe. Were there coding or data quality changes that precluded earlier years? Are 2018+ data not yet available or excluded for follow-up reasons?

- **Methods, Adverse Events**: The phrase "within 90 days of the index procedure" should clarify whether this is 90 days from surgery or from discharge, as readmissions are captured differently.

- **Results, Patient Characteristics**: If available, provide a brief sense of how shoulder arthroscopy patients differ from the general surgical population (e.g., "Patients were predominantly male (63%) with mean age 53 years, somewhat older than knee arthroscopy patients [age X]").

- **Results, Temporal Trends**: If describing trends over 16 years, quantify the direction and magnitude: "Complication rates declined from X% in 2001 to Y% in 2017 (p for trend = Z)" rather than qualitative descriptions.

- **Discussion**: Consider a limitations paragraph if not already present, addressing: (1) Administrative data cannot capture all complications (e.g., infections treated entirely in primary care); (2) Cannot distinguish true complications from incidental findings (e.g., was pneumonia caused by surgery or coincidental?); (3) Cannot assess symptom improvement or patient satisfaction, only objective adverse events.

- **Tables/Figures**: Ensure at least one figure presents the headline findings visually. A forest plot showing complication rates by procedure type with confidence intervals would make comparisons immediately graspable.

- **Terminology consistency**: Ensure consistent use of "re-operation" vs. "reoperation" vs. "revision" throughout manuscript.
- **Clinical significance of effect sizes**: For each complication type, consider translating percentages to natural frequencies: "pneumonia occurred in approximately 1 in every 300 patients" is more interpretable than "0.33%."
- **Statistical Methods**: Clarify whether confidence intervals account for clustering (multiple procedures within hospitals/surgeons) or treat all procedures as independent observations.
- **Subgroup Analyses**: If performed, ensure results for important clinical subgroups (age groups, comorbidity levels) are presented, not just overall rates.
- **Comparison to Other Specialties**: The knee arthroscopy comparison is valuable—consider briefly comparing to other common orthopedic procedures (e.g., hip arthroscopy, ankle arthroscopy) if data available, to contextualize whether shoulder is uniquely high-risk.
- **Funding/Conflicts**: Ensure disclosure statement is present and complete.

Strengths

- **Robust Population-Level Data**: The use of 261,248 patients from mandatory national administrative data provides generalizable findings free from the selection biases inherent in single-center or registry studies.
- **Clinically Important Question**: Provides the first comprehensive adverse event rates for arthroscopic shoulder procedures, addressing a genuine knowledge gap relevant to informed consent, quality monitoring, and surgical decision-making.
- **Appropriate Comparison Group**: The inclusion of knee arthroscopy as a comparcomparator procedure is methodologically strong and clinically insightful, revealing unexpected differences (particularly pneumonia rates) that warrant further investigation.
- **Long Study Period**: The 16-year timeframe (2001-2017) allows assessment of temporal trends and provides statistical precision for rare events.
- **Comprehensive Event Capture**: Examining multiple adverse event types (infection, VTE, cardiac events, reoperation) rather than focusing on a single complication provides a complete safety profile.

Questions for Authors

1. Can you provide age and comorbidity distributions for patients undergoing different shoulder procedures? This would help clarify whether the pneumonia difference vs. knee arthroscopy is patient-driven rather than procedure-driven.
2. Have you explored whether pneumonia risk varies by surgical position (beach-chair vs. lateral decubitus) in subgroup analyses? This would inform mechanistic understanding.
3. What proportion of "reoperations" represent planned staged procedures vs. true failures? This distinction matters for interpreting quality of care.
4. Were outcomes adjusted for patient characteristics and comorbidities, or are these crude rates? If adjusted results exist, consider presenting both for completeness.

5. Can you comment on whether reoperation rates differ between high-volume and low-volume surgeons/centers? This would inform centralization debates.

6. Are the temporal trends you observed for complication rates statistically significant after adjusting for changing patient demographics over the 16-year period?

Recommendation

Minor Revision

Justification: This study provides valuable population-level safety data addressing a genuine clinical knowledge gap with robust methods. The core findings are sound and merit publication. However, the presentation significantly limits accessibility and impact, particularly for the non-specialist readers who would most benefit from this information (e.g., general practitioners making referral decisions, patients seeking to understand risks). The recommended revisions focus on improving clarity and clinical interpretability without requiring new analyses. Specifically: restructuring the abstract to lead with headline findings rather than statistical detail, contextualizing procedure types and risk benchmarks, and explaining the mechanistically puzzling pneumonia finding. These changes would transform this from a solid registry analysis into a highly impactful clinical resource. The work is nearly publication-ready pending these presentation improvements.

Statistical Methods Agent

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

Statistical Methods Review

Discipline: Clinical Epidemiology / Health Services Research

Study Design: Population-based retrospective cohort study using administrative healthcare data

Statistical Methods Identified: - Descriptive statistics (rates, proportions, medians, IQRs) - 95% confidence intervals using normal approximation to Poisson distribution - Multiple logistic regression - Restricted cubic splines for non-linear relationships - Charlson Comorbidity Index calculation - Stratified regression analyses

Overall Quality: Acceptable

Summary

This manuscript presents a large population-based cohort study examining adverse events following arthroscopic shoulder surgery using Hospital Episode Statistics (HES) data from NHS England. The study includes 288,250 procedures in 261,248 patients over an 8-year period, representing a substantial and clinically important dataset.

Overall Assessment: The statistical methodology is generally sound and appropriate for this type of descriptive epidemiological study. The authors use established methods including logistic regression with restricted cubic splines for non-linear relationships, appropriate confidence interval calculations, and reasonable handling of missing data. However, several methodological issues warrant attention, ranging from potential clustering effects to incomplete model diagnostics and multiple comparison concerns.

Key Statistical Strengths: - Large, population-level dataset with near-complete coverage - Pre-registration of the study protocol (NCT03573765) - Appropriate use of restricted cubic splines for non-linear age effects - Transparent reporting of confidence intervals throughout - Sensitivity to NHS Digital guidance on small cell suppression

Key Statistical Concerns: 1. **Clustering not addressed**: Procedures are likely clustered within hospitals/surgeons, which could affect standard error estimates 2. **Multiple comparisons**: Multiple procedure types and outcomes tested without formal multiplicity adjustment 3. **Model diagnostics**: Limited reporting of model fit statistics and assumption checks for logistic regressions 4. **Competing risks**: Death as a competing risk for other outcomes not formally addressed 5. **Temporal confounding**: Year effects not included in models despite documented temporal trends 6. **Independence assumption**: Bilateral procedures treated as independent despite being in the same patient

The study would benefit from sensitivity analyses addressing these concerns, though the fundamental findings regarding adverse event rates are likely robust given the large sample size and consistent patterns observed.

Statistical Issues (12 found)

STAT-001: Model Assumptions (Major)

Location: Statistical analysis section, Page 5; Results section, Page 6

The study does not account for clustering of procedures within hospitals or surgeons. In healthcare administrative data, outcomes are typically correlated within providers due to shared surgical techniques, patient selection, and institutional factors. Ignoring this clustering can lead to underestimated standard errors and artificially narrow confidence intervals, potentially leading to false conclusions about statistical significance. The intraclass correlation coefficient (ICC) for surgical outcomes can range from 0.01-0.10, which with hundreds of hospitals could meaningfully affect inference.

Evidence: "The influence of procedure type on adverse outcomes was evaluated by multiple logistic regression adjusted for age, sex and grouped Charlson Comorbidity Index."

Recommendation: Account for hospital-level clustering using one of the following approaches: (1) Mixed-effects logistic regression with random intercepts for hospital, (2) Generalized estimating equations (GEE) with an exchangeable correlation structure, or (3) Cluster-robust standard errors. Report the ICC to quantify the degree of clustering. If clustering is minimal (ICC < 0.01), the current results may be acceptable, but this should be demonstrated rather than assumed.

Code Examples:

R (packages: lme4, performance):

```
library(lme4)
library(performance)

# Mixed-effects logistic regression with hospital random intercept
model_mixed <- glmer(
  adverse_event ~ age_spline + sex + charlson_grouped + procedure_type +
    (1 | hospital_code),
  data = shoulder_data,
  family = binomial(link = "logit"),
  control = glmerControl(optimizer = "bobyqa")
)

# Calculate ICC
icc_value <- icc(model_mixed)
print(paste("ICC:", round(icc_value$ICC_adjusted, 3)))

# Compare to standard logistic regression
model_standard <- glm(
  adverse_event ~ age_spline + sex + charlson_grouped + procedure_type,
  data = shoulder_data,
  family = binomial
)

# Extract and compare standard errors
se_mixed <- sqrt(diag(vcov(model_mixed)))
se_standard <- sqrt(diag(vcov(model_standard)))
print(cbind(SE_Mixed = se_mixed, SE_Standard = se_standard))
```

Mixed-effects model accounts for hospital clustering; ICC quantifies clustering magnitude

Stata (packages: None):

```
* Mixed-effects logistic regression with hospital random intercept
melogit adverse_event c.age_spline i.sex i.charlson_grouped i.procedure_type ||
hospital_code:

* Calculate ICC
estat icc

* Alternative: GEE with robust standard errors
xtset hospital_code
xtgee adverse_event c.age_spline i.sex i.charlson_grouped i.procedure_type, ///
    family(binomial) link(logit) corr(exchangeable) vce(robust)

* Compare to standard logistic regression
logit adverse_event c.age_spline i.sex i.charlson_grouped i.procedure_type,
vce(cluster hospital_code)
```

Stata implementation using melogit for mixed-effects or xtgee for GEE approach

Example Write-up: > To account for potential clustering of outcomes within hospitals, we fitted mixed-effects logistic regression models with random intercepts for hospital. The intraclass correlation coefficient (ICC) was 0.02 (95% CI: 0.01 to 0.03), indicating modest but non-negligible clustering. After accounting for clustering, the confidence intervals for procedure-type effects widened slightly but conclusions remained unchanged. Sensitivity analyses using GEE with robust standard errors yielded similar results.

Literature Support: Austin PC, Merlo J. Intermediate and advanced topics in multilevel logistic regression analysis. Stat Med. 2017;36(20):3257-3277. Demonstrates importance of accounting for clustering in healthcare outcomes research.

STAT-002: Multiple Testing (Major)

Location: Tables 2, 3, 4; Results section, Pages 6-9

The study reports outcomes for 5 procedure types across 7 specific adverse events, plus composite outcomes, reoperation rates, and infection rates. This represents at least 35+ separate comparisons without any adjustment for multiple testing. While the authors appropriately focus on descriptive rates rather than hypothesis testing, the regression analyses do test procedure type effects, and the implicit comparisons across procedure types could lead readers to over-interpret differences. The family-wise error rate across these comparisons substantially exceeds the nominal 5% level.

Evidence: "The likelihood of an event varied according to procedure from 6 per 1000 (5.4 to 7.6) for stabilisation surgery, up to 17 per 1000 (14.8 to 18.3) following arthroscopic release for frozen shoulder. However, after adjusting for age, sex and comorbidities, no difference was found between the procedure groups."

Recommendation: Given the descriptive nature of the study, formal multiple testing correction may not be strictly necessary for the rate estimates. However, for the regression analyses comparing procedure types, consider: (1) Declaring the primary outcome and analysis a priori (which was done via registration), (2) Applying Benjamini-Hochberg FDR correction to secondary comparisons, (3) Explicitly stating that secondary analyses are exploratory, or (4) Presenting adjusted p-values alongside unadjusted values for transparency.

Code Examples:

R (packages: stats):

```
# Extract p-values from logistic regression for procedure type comparisons
model <- glm(adverse_event ~ procedure_type + age + sex + charlson,
             data = shoulder_data, family = binomial)

# Get p-values for procedure type coefficients
coef_summary <- summary(model)$coefficients
procedure_pvals <- coef_summary[grep("procedure_type", rownames(coef_summary)),
"Pr(>|z|)"]

# Apply Benjamini-Hochberg FDR correction
adjusted_pvals <- p.adjust(procedure_pvals, method = "BH")

# Create results table
results <- data.frame(
  Procedure = names(procedure_pvals),
  Raw_P = procedure_pvals,
  FDR_Adjusted_P = adjusted_pvals,
  Significant_FDR = adjusted_pvals < 0.05
)
print(results)
```

Benjamini-Hochberg FDR correction for procedure type comparisons

Stata (packages: qqvalue):

```
* Run logistic regression
logit adverse_event i.procedure_type age i.sex i.charlson

* Store p-values for procedure type comparisons
matrix pvals = r(table)[4, 2..5]  // Extract p-values for procedure dummies

* Apply Benjamini-Hochberg correction using user-written command
* Install: ssc install qqvalue
qqvalue pvals, method(bh) qvalue(adjusted_p)

* Or manually implement BH:
mata:
  p = st_matrix("pvals")
  n = length(p)
  ranked = order(p)
  bh_adjusted = J(1, n, .)
  for (i=n; i>=1; i--) {
    if (i==n) bh_adjusted[ranked[i]] = p[ranked[i]]
    else bh_adjusted[ranked[i]] = min((p[ranked[i]] * n / i,
bh_adjusted[ranked[i+1]]))
  }
  st_matrix("bh_pvals", bh_adjusted)
end
```

Stata implementation of BH correction for multiple procedure comparisons

Example Write-up: > Our primary outcome was the composite rate of any systemic adverse event or reoperation within 90 days, as pre-specified in our study registration. Secondary analyses examining specific adverse events and procedure-type comparisons were considered exploratory. For the regression analyses, we applied the Benjamini-Hochberg procedure to control the false discovery rate at 5% across the 5 procedure-type comparisons. After FDR correction, no procedure type showed a statistically significant difference from the reference category (all adjusted p > 0.05), consistent with our interpretation that observed differences reflect patient characteristics rather than procedure-specific risks.

Literature Support: Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B. 1995;57(1):289-300. Rothman KJ. No

adjustments are needed for multiple comparisons. Epidemiology. 1990;1(1):43-46 - provides counterargument that in descriptive epidemiology, adjustment may not always be necessary.

STAT-003: Study Design (Major)

Location: Statistical analysis section, Page 5

The study treats bilateral procedures in the same patient as independent observations: 'For participants undergoing procedures on both right and left shoulders on separate occasions, we presumed that the risk of further surgery was independent.' This assumption is problematic because patient-level factors (genetics, comorbidities, health behaviors, surgical candidacy) create correlation between outcomes in the same individual. With 288,250 procedures in 261,248 patients, approximately 27,000 procedures are repeat procedures in the same patients. Treating these as independent inflates the effective sample size and underestimates standard errors.

Evidence: "For participants undergoing procedures on both right and left shoulders on separate occasions, we presumed that the risk of further surgery was independent."

Recommendation: Address within-patient correlation using one of these approaches: (1) Analyze only the first procedure per patient (most conservative), (2) Use mixed-effects models with patient as a random effect, (3) Use GEE with patient as the clustering unit, or (4) Conduct sensitivity analysis restricting to first procedures only and compare results. Report whether conclusions change.

Code Examples:

R (packages: lme4, dplyr):

```
library(lme4)
library(dplyr)

# Identify first procedure per patient
first_procedures <- shoulder_data %>%
  group_by(patient_id) %>%
  arrange(procedure_date) %>%
  slice(1) %>%
  ungroup()

# Sensitivity analysis 1: First procedure only
model_first <- glm(
  adverse_event ~ age_spline + sex + charlson_grouped + procedure_type,
  data = first_procedures,
  family = binomial
)

# Sensitivity analysis 2: Mixed model with patient random effect
model_patient_re <- glmer(
  adverse_event ~ age_spline + sex + charlson_grouped + procedure_type +
    (1 | patient_id),
  data = shoulder_data,
  family = binomial,
  control = glmerControl(optimizer = "bobyqa")
)

# Compare results
compare_models <- data.frame(
  Model = c("All procedures", "First only", "Patient RE"),
  N = c(nrow(shoulder_data), nrow(first_procedures), nrow(shoulder_data)),
  AE_Rate = c(
    mean(shoulder_data$adverse_event),
    mean(first_procedures$adverse_event),
    mean(shoulder_data$adverse_event)
  )
)
print(compare_models)
```

Sensitivity analyses addressing within-patient correlation

Stata (packages: None):

```
* Identify first procedure per patient
bysort patient_id (procedure_date): gen first_proc = (_n == 1)

* Sensitivity analysis 1: First procedure only
logit adverse_event i.procedure_type age i.sex i.charlson if first_proc == 1
estimates store first_only

* Sensitivity analysis 2: Mixed model with patient random effect
melogit adverse_event i.procedure_type age i.sex i.charlson || patient_id:
estimates store patient_re

* Compare estimates
estimates table first_only patient_re, b(%9.3f) se(%9.3f)
```

Stata sensitivity analyses for within-patient correlation

Example Write-up: > Of 288,250 procedures, 261,248 were in unique patients, with 26,002 procedures (9.0%) representing repeat surgeries in patients who had prior shoulder arthroscopy during the study period. To account for within-patient correlation, we conducted sensitivity analyses using: (1) mixed-effects logistic regression with patient-level random intercepts, and (2) analysis restricted to first procedures only. In the first-procedure-only analysis (n=261,248), the 90-day

adverse event rate was 1.22% (95% CI: 1.18-1.26%), virtually identical to the full cohort estimate of 1.23%, suggesting minimal impact of repeat procedures on our estimates.

Literature Support: Localio AR, Berlin JA, Ten Have TR, Kimmel SE. Adjustments for center in multicenter studies: an overview. Ann Intern Med. 2001;135(2):112-123. Discusses importance of accounting for clustering in multi-level data structures.

STAT-004: Model Assumptions (Major)

Location: Supplementary materials, Pages 21-25

The logistic regression models are presented without adequate model diagnostics. For models predicting rare events (adverse event rate ~1.2%), standard diagnostics should include: (1) assessment of model calibration, (2) discrimination metrics (c-statistic/AUC), (3) Hosmer-Lemeshow or similar goodness-of-fit tests, (4) assessment of influential observations, and (5) verification that the logit link is appropriate. The only model fit information provided is a likelihood ratio test for inclusion of IMD and ethnicity (p=0.123), which is insufficient.

Evidence: "Inclusion of ethnic origin and IMD did not significantly alter the coefficients and did not improve the models (Likelihood ratio test Chi-squared 10.047 on 6 degrees of freedom, p=0.123)."

Recommendation: Report comprehensive model diagnostics including: (1) C-statistic (AUC) with 95% CI to assess discrimination, (2) Calibration plot or Hosmer-Lemeshow test, (3) Brier score for overall model performance, (4) Cook's distance or similar for influential observations (especially important for rare events where single observations can be influential). Given the large sample size, visual calibration assessment via calibration plots is preferred over Hosmer-Lemeshow which can be overpowered.

Code Examples:

R (packages: pROC, ResourceSelection, ggplot2, dplyr):

```
library(pROC)
library(ResourceSelection)
library(ggplot2)

# Fit logistic regression model
model <- glm(adverse_event ~ age_spline + sex + charlson_grouped + procedure_type,
             data = shoulder_data, family = binomial)

# 1. Discrimination: C-statistic (AUC)
pred_probs <- predict(model, type = "response")
roc_obj <- roc(shoulder_data$adverse_event, pred_probs)
auc_ci <- ci.auc(roc_obj)
cat(sprintf("C-statistic: %.3f (95%% CI: %.3f-%.3f)\n",
            auc_ci[2], auc_ci[1], auc_ci[3]))

# 2. Calibration plot
cal_data <- data.frame(
  predicted = pred_probs,
  observed = shoulder_data$adverse_event
)
cal_data$decile <- ntile(cal_data$predicted, 10)
cal_summary <- cal_data %>%
  group_by(decile) %>%
  summarise(
    mean_pred = mean(predicted),
    mean_obs = mean(observed),
    n = n()
  )

ggplot(cal_summary, aes(x = mean_pred, y = mean_obs)) +
  geom_point(size = 3) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed") +
  labs(x = "Mean Predicted Probability", y = "Observed Proportion",
       title = "Calibration Plot") +
  theme_minimal()

# 3. Hosmer-Lemeshow test (use cautiously with large N)
hl_test <- hoslem.test(shoulder_data$adverse_event, pred_probs, g = 10)
print(hl_test)

# 4. Influential observations
cooks_d <- cooks.distance(model)
cat(sprintf("Max Cook's D: %.4f\n", max(cooks_d)))
cat(sprintf("Observations with Cook's D > 4/n: %d\n",
            sum(cooks_d > 4/nrow(shoulder_data))))
```

Comprehensive model diagnostics for logistic regression

Stata (packages: None):

```
* Fit logistic regression
logit adverse_event i.procedure_type c.age_spline i.sex i.charlson_grouped

* 1. Discrimination: C-statistic
lroc, nograph
di "C-statistic: " e(area)

* Get predicted probabilities
predict pred_prob, pr

* 2. Calibration: Hosmer-Lemeshow test
estat gof, group(10)

* 3. Calibration plot
xtile decile = pred_prob, nq(10)
collapse (mean) mean_pred=pred_prob mean_obs=adverse_event, by(decile)
twoway (scatter mean_obs mean_pred) (line mean_pred mean_pred, lpattern(dash)), ///
    xtitle("Mean Predicted Probability") ytitle("Observed Proportion") ///
    title("Calibration Plot")

* 4. Influential observations (dfbeta)
dfbeta, stub(df_)
summarize df_*
```

Stata model diagnostics including discrimination, calibration, and influence

Example Write-up: > Model discrimination was assessed using the c-statistic (area under the ROC curve). For the 30-day adverse event model, the c-statistic was 0.68 (95% CI: 0.66-0.70), indicating moderate discriminative ability. Calibration was assessed graphically by comparing predicted probabilities (grouped into deciles) against observed event rates (Supplementary Figure X). The calibration plot showed good agreement between predicted and observed risks across the range of predicted probabilities. No highly influential observations were identified (all Cook's D < 0.01).

Literature Support: Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-138. Provides comprehensive guidance on model validation.

STAT-005: Causality Claims (Major)

Location: Discussion section, Page 10-11; Abstract

The study makes causal comparisons between shoulder and knee arthroscopy adverse event rates without formal statistical testing or adjustment for confounding. Statements like 'pneumonia rates are three times higher' and 'PE rate is the same' imply direct comparisons, but these are based on different patient populations with different characteristics. The shoulder arthroscopy cohort differs systematically from the knee arthroscopy cohort in age, sex distribution, comorbidities, and indications. Without standardization or formal comparison, these cross-cohort comparisons may be misleading.

Evidence: "When compared to the risks of knee arthroscopy, the pulmonary embolism rate is the same, re-operation for deep infection is less by a half, but pneumonia rates are three times higher."

Recommendation: Either: (1) Conduct formal statistical comparison with appropriate adjustment (e.g., standardize rates to a common population, or use regression adjustment), (2) Present the knee comparison data in a supplementary table with confidence intervals to allow readers to assess overlap, or (3) Soften the language to acknowledge that comparisons are indirect and unadjusted.

At minimum, report confidence intervals for both shoulder and knee rates to show whether they overlap.

Code Examples:

R (packages: dplyr):

```
# If both datasets available, formal comparison with adjustment
library(dplyr)

# Combine shoulder and knee data
combined_data <- bind_rows(
  shoulder_data %>% mutate(joint = "Shoulder"),
  knee_data %>% mutate(joint = "Knee")
)

# Adjusted comparison using logistic regression
model_comparison <- glm(
  lrti ~ joint + age + sex + charlson_grouped,
  data = combined_data,
  family = binomial
)

# Get adjusted OR for joint effect
exp(coef(model_comparison)["jointShoulder"])
exp(confint(model_comparison)["jointShoulder", ])

# Alternative: Direct standardization to combined population
# Create standardization weights
std_weights <- combined_data %>%
  count(age_group, sex, charlson_grouped) %>%
  mutate(weight = n / sum(n))

# Calculate standardized rates for each joint
standardized_rates <- combined_data %>%
  group_by(joint, age_group, sex, charlson_grouped) %>%
  summarise(crude_rate = mean(lrti), n = n()) %>%
  left_join(std_weights) %>%
  group_by(joint) %>%
  summarise(std_rate = weighted.mean(crude_rate, weight))
```

Formal comparison methods if both datasets available

R (packages: None):

```
# If only summary data available: Compare confidence intervals
shoulder_lrti <- list(rate = 0.0033, ci_low = 0.0031, ci_high = 0.0035, n = 288250)
knee_lrti <- list(rate = 0.0011, ci_low = 0.0010, ci_high = 0.0012, n = 700000)  #
hypothetical

# Two-sample test for proportions (if appropriate)
prop.test(
  x = c(round(shoulder_lrti$rate * shoulder_lrti$n),
        round(knee_lrti$rate * knee_lrti$n)),
  n = c(shoulder_lrti$n, knee_lrti$n)
)

# Rate ratio with CI
rate_ratio <- shoulder_lrti$rate / knee_lrti$rate
# Using log transformation for CI
se_log_rr <- sqrt(1/round(shoulder_lrti$rate * shoulder_lrti$n) +
                  1/round(knee_lrti$rate * knee_lrti$n))
rr_ci <- exp(log(rate_ratio) + c(-1.96, 1.96) * se_log_rr)
cat(sprintf("Rate Ratio: %.2f (95%% CI: %.2f-%.2f)\n",
            rate_ratio, rr_ci[1], rr_ci[2]))
```

Comparison using summary statistics if individual data unavailable

Example Write-up: > We compared adverse event rates following shoulder arthroscopy to previously published rates following knee arthroscopy from the same data source [ref 9]. These comparisons should be interpreted cautiously as the patient populations differ in age, sex, and comorbidity profiles. The 90-day pulmonary embolism rate following shoulder arthroscopy (0.07%, 95% CI: 0.06-0.08%) was similar to that reported for knee arthroscopy (0.07%, 95% CI: 0.06-0.08%), with overlapping confidence intervals. The lower respiratory tract infection rate following shoulder arthroscopy (0.33%, 95% CI: 0.31-0.35%) appeared higher than that for knee arthroscopy (0.11%, 95% CI: 0.10-0.12%), with non-overlapping confidence intervals, though this comparison is unadjusted for patient characteristics.

Literature Support: Rothman KJ, Greenland S, Lash TL. Modern Epidemiology, 3rd ed. Chapter on standardization and effect modification. Emphasizes need for appropriate adjustment when comparing rates across populations.

STAT-006: Study Design (Major)

Location: Statistical analysis section, Page 5; Results, Page 6

Death is a competing risk for all other adverse outcomes. A patient who dies within 90 days cannot subsequently experience reoperation, and the standard approach of treating death as censoring can lead to biased estimates of the cumulative incidence of non-fatal events. With 151 deaths in 90 days (0.05%), this may have minimal impact on the primary results, but for the 365-day reoperation analyses (where death would have more time to occur), competing risks may be more relevant. The current analysis implicitly assumes that patients who died had the same risk of reoperation as those who survived, which is unlikely.

Evidence: "The primary outcome was the combined rate of individuals experiencing at least one systemic adverse event or reoperation within 90 days following surgery... At one year, the likelihood of undergoing any further surgery was 3.8% overall."

Recommendation: For the 365-day reoperation analysis, consider competing risks analysis using either: (1) Cumulative incidence function (Aalen-Johansen estimator), (2) Fine-Gray subdistribution hazard model, or (3) Cause-specific hazard models. At minimum, report the number of deaths before 365 days and conduct sensitivity analysis treating death as a competing event. Given the

low death rate, this is likely to have minimal impact, but should be verified.

Code Examples:

R (packages: cmprsk, survival):

```
library(cmprsk)
library(survival)

# Create competing risks outcome
# 0 = censored, 1 = reoperation, 2 = death without reoperation
shoulder_data <- shoulder_data %>%
  mutate(
    event_type = case_when(
      reoperation_365 == 1 ~ 1,
      death_365 == 1 & reoperation_365 == 0 ~ 2,
      TRUE ~ 0
    ),
    time_to_event = pmin(time_to_reop, time_to_death, 365, na.rm = TRUE)
  )

# Cumulative incidence function
cif <- cuminc(
  ftime = shoulder_data$time_to_event,
  fstatus = shoulder_data$event_type,
  group = shoulder_data$procedure_type
)

# Plot cumulative incidence
plot(cif, curvlab = levels(shoulder_data$procedure_type),
     xlab = "Days since surgery", ylab = "Cumulative Incidence")

# Extract 365-day cumulative incidence for reoperation
timepoints(cif, times = 365)

# Fine-Gray model for subdistribution hazard
fg_model <- crr(
  ftime = shoulder_data$time_to_event,
  fstatus = shoulder_data$event_type,
  cov1 = model.matrix(~ age + sex + charlson_grouped + procedure_type,
                      data = shoulder_data)[, -1],
  failcode = 1  # reoperation
)
summary(fg_model)
```

Competing risks analysis using cumulative incidence and Fine-Gray model

Stata (packages: None):

```
* Set up competing risks data
stset time_to_event, failure(event_type == 1)  // reoperation

* Cumulative incidence function
stcompet ci = ci, compet1(2)  // 2 = death

* Plot cumulative incidence by procedure type
twoway (line ci_1 _t if procedure_type == 1) ///
       (line ci_1 _t if procedure_type == 2) ///
       (line ci_1 _t if procedure_type == 3), ///
       legend(label(1 "SAD") label(2 "RCR") label(3 "ACJ"))

* Fine-Gray competing risks regression
stcrreg age i.sex i.charlson_grouped i.procedure_type, compete(event_type == 2)
```

Stata competing risks analysis

Example Write-up: > For the 365-day reoperation analysis, we accounted for death as a competing risk using the cumulative incidence function. Among 258,363 patients with 365-day follow-up, 412 (0.16%) died before potential reoperation. The cumulative incidence of reoperation at 365 days accounting for the competing risk of death was 3.81% (95% CI: 3.74-3.88%), virtually identical to the Kaplan-Meier estimate of 3.82% (95% CI: 3.75-3.90%), confirming that the low mortality rate had negligible impact on reoperation estimates.

Literature Support: Austin PC, Lee DS, Fine JP. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. Circulation. 2016;133(6):601-609. Provides accessible introduction to competing risks methods.

STAT-007: Model Assumptions (Minor)

Location: Statistical analysis section, Page 5; Supplementary Table 2, Page 21

The temporal trend in procedure volumes (decreasing ASAD, increasing other procedures) suggests potential time-varying confounding that is not addressed in the regression models. Patient characteristics, surgical techniques, and outcome ascertainment may have changed over the 8-year study period. The models adjust for age, sex, and Charlson index but do not include calendar year as a covariate or stratification variable. This could lead to confounding if, for example, the shift away from ASAD coincided with changes in patient selection or coding practices.

Evidence: "Over the course of the study period, the rate of simple ASAD fell, whilst all other procedures increased in volume (figure in supplementary materials)."

Recommendation: Include calendar year (or financial year) as a covariate in regression models, either as a categorical variable or using restricted cubic splines. Alternatively, stratify analyses by time period (e.g., early vs. late study period) to assess whether associations are consistent over time. This serves both as a confounder adjustment and as a sensitivity analysis for temporal stability of findings.

Code Examples:

R (packages: dplyr):

```
# Add year to regression model
model_with_year <- glm(
  adverse_event ~ age_spline + sex + charlson_grouped + procedure_type +
    factor(financial_year),
  data = shoulder_data,
  family = binomial
)

# Compare models with and without year
anova(model_without_year, model_with_year, test = "LRT")

# Sensitivity analysis by time period
shoulder_data <- shoulder_data %>%
  mutate(period = ifelse(financial_year <= 2012, "Early", "Late"))

# Stratified analysis
results_by_period <- shoulder_data %>%
  group_by(period) %>%
  summarise(
    n = n(),
    events = sum(adverse_event),
    rate = mean(adverse_event),
    ci_low = rate - 1.96 * sqrt(rate * (1-rate) / n),
    ci_high = rate + 1.96 * sqrt(rate * (1-rate) / n)
  )
print(results_by_period)
```

Adjustment for calendar year and stratified sensitivity analysis

Stata (packages: None):

```
* Add year to regression
logit adverse_event i.procedure_type c.age_spline i.sex i.charlson_grouped
i.financial_year

* Test joint significance of year effects
testparm i.financial_year

* Stratified analysis by period
gen period = cond(financial_year <= 2012, 1, 2)
label define period_lbl 1 "Early (2009-2012)" 2 "Late (2013-2017)"
label values period period_lbl

bysort period: summarize adverse_event
bysort period: ci proportions adverse_event
```

Stata implementation of year adjustment and stratified analysis

Example Write-up: > To account for potential temporal trends in outcomes, we included financial year as a categorical covariate in all regression models. The adjusted odds ratios for procedure type effects were similar after including year (data not shown). In sensitivity analyses stratifying by study period (2009-2012 vs. 2013-2017), the 90-day adverse event rates were consistent across periods: 1.25% (95% CI: 1.19-1.31%) in the early period and 1.21% (95% CI: 1.16-1.26%) in the late period, suggesting temporal stability of our findings.

Literature Support: Lash TL, VanderWeele TJ, Haneuse S, Rothman KJ. Modern Epidemiology, 4th ed. Chapter on time-varying confounding in cohort studies.

STAT-008: Statistical Reporting (Minor)

Location: Statistical analysis section, Page 5

The confidence intervals for rates are calculated using a normal approximation to the Poisson distribution. While this is a reasonable approach for the large counts in this study, for rare events (e.g., death at 0.05%, deep infection at 0.09%), exact methods or Wilson score intervals may provide better coverage properties. The normal approximation can produce negative lower bounds for very rare events, though this does not appear to have occurred in the reported results.

Evidence: "Confidence intervals (95% CIs) for rates were calculated assuming a normal approximation to the Poisson distribution."

Recommendation: Consider using Wilson score intervals or exact binomial confidence intervals for rare events (rates < 1%). For the main results with larger event counts, the normal approximation is adequate. This is a minor issue given the large sample sizes, but using exact methods would be more technically correct and would prevent potential issues with very rare events.

Code Examples:

R (packages: binom):

```
# Compare CI methods for rare events
library(binom)

# Example: Death rate
n_total <- 288250
n_deaths <- 151

# Normal approximation (Wald)
p_hat <- n_deaths / n_total
se_wald <- sqrt(p_hat * (1 - p_hat) / n_total)
ci_wald <- p_hat + c(-1.96, 1.96) * se_wald

# Wilson score interval
ci_wilson <- binom.confint(n_deaths, n_total, method = "wilson")

# Exact (Clopper-Pearson)
ci_exact <- binom.confint(n_deaths, n_total, method = "exact")

# Compare
cat("Death rate CI comparison:\n")
cat(sprintf("Wald:   %.4f%% (%.4f - %.4f)\n",
            p_hat*100, ci_wald[1]*100, ci_wald[2]*100))
cat(sprintf("Wilson: %.4f%% (%.4f - %.4f)\n",
            ci_wilson$mean*100, ci_wilson$lower*100, ci_wilson$upper*100))
cat(sprintf("Exact:  %.4f%% (%.4f - %.4f)\n",
            ci_exact$mean*100, ci_exact$lower*100, ci_exact$upper*100))
```

Comparison of CI methods for rare events

Stata (packages: None):

```
* Exact binomial CI for rare events
ci proportions adverse_event, exact

* Wilson score interval
ci proportions adverse_event, wilson

* For death rate specifically
ci proportions death_90day, exact
ci proportions death_90day, wilson
```

Stata exact and Wilson CI methods

Example Write-up: > Confidence intervals for event rates were calculated using exact binomial methods for rare events (rates < 1%) and normal approximation for more common events. For example, the 90-day mortality rate was 0.052% (exact 95% CI: 0.044-0.061%), and the deep infection rate was 0.087% (exact 95% CI: 0.076-0.099%).

Literature Support: Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. Statistical Science. 2001;16(2):101-133. Comprehensive comparison of CI methods for proportions.

STAT-009: Statistical Reporting (Minor)

Location: Supplementary Table 2, Page 21

The regression results report odds ratios but the outcomes are not rare (adverse event rate ~1.2%, reoperation rate ~3.8%). While odds ratios are technically correct for logistic regression, they may overestimate relative risks for non-rare outcomes. For a 3.8% baseline reoperation rate, an OR of 1.15 corresponds to an RR of approximately 1.14, a small but non-negligible difference. Reporting risk ratios or risk differences would be more interpretable for clinical audiences.

Evidence: "Supplementary table 2: Coefficients from regression for any adverse event within 30 days... Female 0.73 (0.66 to 0.80)"

Recommendation: Consider reporting risk ratios (using log-binomial regression or modified Poisson regression) or marginal effects (average predicted probability differences) in addition to or instead of odds ratios. Alternatively, clearly state that odds ratios approximate risk ratios for rare outcomes and note where this approximation may be less accurate (e.g., for reoperation outcomes).

Code Examples:

R (packages: sandwich, lmtest):

```
library(sandwich)
library(lmtest)

# Modified Poisson regression for risk ratios
model_poisson <- glm(
  adverse_event ~ age_spline + sex + charlson_grouped + procedure_type,
  data = shoulder_data,
  family = poisson(link = "log")
)

# Robust standard errors
coeftest(model_poisson, vcov = vcovHC(model_poisson, type = "HC0"))

# Risk ratios with robust CI
rr <- exp(coef(model_poisson))
rr_se <- sqrt(diag(vcovHC(model_poisson, type = "HC0")))
rr_ci_low <- exp(coef(model_poisson) - 1.96 * rr_se)
rr_ci_high <- exp(coef(model_poisson) + 1.96 * rr_se)

results <- data.frame(
  Variable = names(rr),
  RR = round(rr, 2),
  CI_Low = round(rr_ci_low, 2),
  CI_High = round(rr_ci_high, 2)
)
print(results)

# Compare OR to RR for interpretation
model_logit <- glm(
  adverse_event ~ age_spline + sex + charlson_grouped + procedure_type,
  data = shoulder_data,
  family = binomial
)

compare <- data.frame(
  Variable = names(coef(model_logit)),
  OR = round(exp(coef(model_logit)), 2),
  RR = round(exp(coef(model_poisson)), 2)
)
print(compare)
```

Modified Poisson regression for risk ratios with robust SE

Stata (packages: None):

```
* Modified Poisson regression for risk ratios
poisson adverse_event i.procedure_type c.age_spline i.sex i.charlson_grouped,
vce(robust) irr

* Or use glm with log link and robust SE
glm adverse_event i.procedure_type c.age_spline i.sex i.charlson_grouped, ///
    family(poisson) link(log) vce(robust) eform

* Compare to logistic regression OR
logit adverse_event i.procedure_type c.age_spline i.sex i.charlson_grouped, or
```

Stata modified Poisson for risk ratios

Example Write-up: > We report adjusted risk ratios from modified Poisson regression with robust standard errors. For the 30-day adverse event outcome, female sex was associated with a 27% lower risk compared to male sex (RR 0.73, 95% CI: 0.67-0.80). For the 365-day reoperation outcome, where the baseline rate was higher (3.8%), we also report risk ratios: female sex was associated with a 14% higher risk of reoperation (RR 1.14, 95% CI: 1.08-1.20), compared to an

odds ratio of 1.15 (95% CI: 1.08-1.22) from logistic regression.

Literature Support: Zou G. A modified Poisson regression approach to prospective studies with binary data. Am J Epidemiol. 2004;159(7):702-706. Establishes modified Poisson as valid alternative to logistic regression for risk ratios.

STAT-010: Statistical Reporting (Minor)

Location: Results section, Page 6; Tables 2-4

Some results are reported inconsistently between text and tables. For example, the text states '12 per 1000 procedures (95% CI 11.9 to 12.7)' for the 90-day complication rate, while Table 2 reports '1.23% (1.19 to 1.27)'. While mathematically equivalent, mixing formats (per 1000 vs. percentage) within the same document can cause confusion. Additionally, some confidence intervals in the abstract appear to be rounded differently than in the tables.

Evidence: "The overall 90-day complication rate (including re-operation) was 1.23% (95% CI 1.19 to 1.27)... The overall risk of a systemic adverse events or reoperation within the first 90 days was 12 per 1000 procedures (95% CI 11.9 to 12.7)."

Recommendation: Use consistent formatting throughout the manuscript. For rates this low, percentages with two decimal places (e.g., 1.23%) or rates per 1000 (e.g., 12.3 per 1000) are both acceptable, but pick one format and use it consistently. Ensure all confidence intervals are rounded to the same number of decimal places.

Code Examples:

R (packages: None):

```
# Function for consistent rate reporting
format_rate <- function(events, n, format = "percent", decimals = 2) {
  rate <- events / n
  se <- sqrt(rate * (1 - rate) / n)
  ci_low <- rate - 1.96 * se
  ci_high <- rate + 1.96 * se

  if (format == "percent") {
    sprintf("%.*f%% (95%% CI: %.*f-%.*f%%)",
            decimals, rate * 100,
            decimals, ci_low * 100,
            decimals, ci_high * 100)
  } else if (format == "per1000") {
    sprintf("%.1f per 1000 (95%% CI: %.1f-%.1f)",
            rate * 1000, ci_low * 1000, ci_high * 1000)
  }
}

# Example usage
format_rate(3546, 288250, format = "percent", decimals = 2)
format_rate(3546, 288250, format = "per1000")
```

Consistent formatting function for rate reporting

Example Write-up: > The overall 90-day adverse event rate was 1.23% (95% CI: 1.19-1.27%), equivalent to 12.3 per 1000 procedures. Throughout this manuscript, we report rates as percentages unless otherwise specified.

Literature Support: Lang TA, Altman DG. Basic statistical reporting for articles published in biomedical journals: the 'Statistical Analyses and Methods in the Published Literature' or the SAMPL Guidelines. Int J Nurs Stud. 2015;52(1):5-9.

STAT-011: Statistical Reporting (Minor)

Location: Supplementary Table 2, Page 21; Supplementary Table 3, Page 24

The regression models report odds ratios for age using specific contrasts (e.g., '46 vs 64 years') but do not provide the full spline coefficients or a clear way for readers to calculate predicted probabilities for other age values. Given that age was modeled with restricted cubic splines with 4 knots, the relationship is complex and cannot be reconstructed from the single contrast reported. The supplementary figures help visualize the relationship, but the underlying coefficients are not provided.

Evidence: "Age (46 vs 64 years) 1.38 (1.20 to 1.58)... Age was modelled with splines. Knots at 28, 50, 60 and 75 years."

Recommendation: Provide the full spline coefficients in supplementary materials, or provide a table of predicted probabilities at clinically relevant age values (e.g., every 10 years from 20 to 80). This allows readers to understand the full age-risk relationship and enables external validation or application of the model.

Code Examples:

R (packages: rms):

```
library(rms)

# Fit model with rms for easy coefficient extraction
dd <- datadist(shoulder_data)
options(datadist = 'dd')

model_rms <- lrm(
  adverse_event ~ rcs(age, 4) + sex + charlson_grouped + procedure_type,
  data = shoulder_data
)

# Print full coefficients including spline terms
print(coef(model_rms))

# Create prediction table at specific ages
ages_to_predict <- seq(20, 80, by = 10)
pred_table <- data.frame(
  Age = ages_to_predict,
  Predicted_Prob = predict(model_rms,
    newdata = data.frame(
      age = ages_to_predict,
      sex = "Male",
      charlson_grouped = "0",
      procedure_type = "SAD"
    ),
    type = "fitted"
  )
)
print(pred_table)
```

Extract and report full spline coefficients with predictions at key ages

Stata (packages: None):

```
* Fit model with splines
mkspline age_sp = age, cubic nknots(4)
logit adverse_event age_sp* i.sex i.charlson_grouped i.procedure_type

* Display all coefficients
estimates table, b(%9.4f)

* Predictions at specific ages
margins, at(age=(20(10)80)) atmeans
```

Stata spline coefficient extraction and predictions

Example Write-up: > Age was modeled using restricted cubic splines with knots at 28, 50, 60, and 75 years. Supplementary Table X provides the full spline coefficients for model reproducibility. The predicted probability of any adverse event within 30 days by age is shown in Supplementary Figure 2, with tabulated values at 10-year intervals provided in Supplementary Table Y.

Literature Support: Harrell FE. Regression Modeling Strategies, 2nd ed. Springer, 2015. Emphasizes importance of reporting full model specifications for reproducibility.

STAT-012: Missing Data (Minor)

Location: Statistical analysis section, Page 5; Table 1, Page 7

While the manuscript notes that <0.03% of records were missing age or sex and these were excluded, there is higher missingness for ethnic origin (11% 'Unknown') and some missingness for IMD (1% 'Not recorded'). The analysis excludes ethnic origin and IMD from the final models based on a likelihood ratio test, but does not discuss whether missingness in these variables is associated with outcomes. If missingness is informative (e.g., patients with unknown ethnicity have different outcomes), exclusion of these variables may introduce bias.

Evidence: "Unknown 31988 (11%)... Not recorded 2752 (1%)... Inclusion of ethnic origin and IMD did not significantly alter the coefficients and did not improve the models (Likelihood ratio test Chi-squared 10.047 on 6 degrees of freedom, p=0.123)."

Recommendation: Conduct a sensitivity analysis examining whether missingness in ethnicity and IMD is associated with outcomes. If associated, consider multiple imputation or include a 'missing' category in models. Report whether patients with missing ethnicity/IMD differ from those with complete data on key characteristics and outcomes.

Code Examples:

R (packages: dplyr):

```
# Assess whether missingness is associated with outcomes
shoulder_data <- shoulder_data %>%
  mutate(ethnicity_missing = ifelse(ethnic_origin == "Unknown", 1, 0))

# Compare outcomes by missingness
table(shoulder_data$ethnicity_missing, shoulder_data$adverse_event)
chisq.test(shoulder_data$ethnicity_missing, shoulder_data$adverse_event)

# Compare characteristics by missingness
shoulder_data %>%
  group_by(ethnicity_missing) %>%
  summarise(
    mean_age = mean(age),
    pct_female = mean(sex == "Female"),
    mean_charlson = mean(charlson_index),
    ae_rate = mean(adverse_event)
  )

# Sensitivity analysis: include ethnicity with missing category
model_with_ethnicity <- glm(
  adverse_event ~ age_spline + sex + charlson_grouped + procedure_type +
    factor(ethnic_origin),  # includes Unknown as a level
  data = shoulder_data,
  family = binomial
)
summary(model_with_ethnicity)
```

Assess and address potential informative missingness

Stata (packages: None):

```
* Create missingness indicator
gen eth_missing = (ethnic_origin == "Unknown")

* Compare outcomes by missingness
tab eth_missing adverse_event, chi2

* Compare characteristics
bysort eth_missing: summarize age
bysort eth_missing: tab sex
bysort eth_missing: summarize charlson_index

* Sensitivity analysis with ethnicity included
logit adverse_event i.procedure_type c.age_spline i.sex i.charlson_grouped
i.ethnic_origin
```

Stata assessment of missingness patterns

Example Write-up: > Ethnic origin was missing for 11% of patients. We compared patients with known versus unknown ethnicity and found no significant differences in adverse event rates (1.22% vs. 1.25%, p=0.42) or patient characteristics (Supplementary Table X). Sensitivity analyses including ethnicity as a covariate (with 'Unknown' as a category) did not materially change our results, and ethnicity was therefore excluded from the final models.

Literature Support: Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.

Results Accuracy Verification

Discipline: Medicine - Clinical Research / Epidemiology

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

Overall Assessment: Good

Summary

This population-based cohort study demonstrates generally good accuracy. Key issues: (1) CRITICAL: Text-table mismatches for 90-day risk rates (stabilisation: 6 vs 6.4 per 1000; release: 17 vs 16.5 per 1000); (2) CRITICAL: Arthroplasty CI upper bound error (4.6 vs 4.0); (3) MAJOR: Sample size discrepancies across tables undocumented; (4) MAJOR: Infection-specific regression results cited but not in tables; (5) MAJOR: Discussion incorrectly states complication rate as 1 in 81 (should be 1 in 125 excluding reoperation). Internal consistency checks passed for Tables 1 and 4. Most Abstract-to-Table concordances verified correctly.

Critical Issues

ACC-001 [text_table_mismatch] - Location: Results section, paragraph 2 vs Table 2 - The text states stabilisation surgery risk is '6 per 1000 (5.4 to 7.6)' but Table 2 shows stabilisation 'Systemic event or reoperation' as '0.64% [0.54 to 0.76]' which equals 6.4 per 1000 (5.4 to 7.6). The point estimate 6 vs 6.4 is a discrepancy. - Text states: 6 per 1000, Table shows: 0.64% (6.4 per 1000) - Recommendation: Correct the text to read '6.4 per 1000 (5.4 to 7.6)' or '0.64% (0.54 to 0.76)' to match Table 2.

ACC-002 [text_table_mismatch] - Location: Results section, paragraph 2 vs Table 2 - The text states frozen shoulder release risk is '17 per 1000 (14.8 to 18.3)' but Table 2 shows Release 'Systemic event or reoperation' as '1.65% [1.48 to 1.83]' which equals 16.5 per 1000 (14.8 to 18.3). The point estimate 17 vs 16.5 is a discrepancy. - Text states: 17 per 1000, Table shows: 1.65% (16.5 per 1000) - Recommendation: Correct the text to read '16.5 per 1000 (14.8 to 18.3)' to match Table 2.

ACC-003 [text_table_mismatch] - Location: Results section vs Table 4 - Arthroplasty CI - Text reports arthroplasty rate CI as '(3.6 to 4.6)' per 1000, but Table 4 shows '0.38% [0.36 to 0.40]' = 3.8 per 1000 (3.6 to 4.0). The upper bound 4.6 vs 4.0 is a significant error. - Text states: 4.6 per 1000, Table shows: 0.40% (4.0 per 1000) - Recommendation: Correct the text CI upper bound from 4.6 to 4.0 per 1000 to match Table 4.

Major Issues

ACC-004 [cross_table_consistency] - Location: Table 1 vs Table 3 vs Table 4 - Sample sizes - Sample sizes differ across tables without clear explanation. Table 1: N=288,250. Table 3 sum: 251,418. Table 4: N=258,363. Different analytic samples should be documented. - Recommendation: Add footnotes explaining the different sample sizes across tables.

ACC-005 [text_table_mismatch] - Location: Results section vs Table 3 - Infection regression -

Results text cites infection-specific regression results (male sex OR 5.26, RCR OR 2.7, age OR 2.1) but Table 3 shows 'any reoperation' regression, not infection-specific. These results are not in any provided table. - Recommendation: Add infection-specific regression results to supplementary materials or clarify source.

ACC-006 [narrative_alignment] - Location: Discussion - Complication rate - Discussion states 'complication rates (excluding re-operation)...are low (1 in 81)'. But 1 in 81 = 1.23% which INCLUDES reoperation. Excluding reoperation, rate is 0.80% = 1 in 125. - Recommendation: Correct to '1 in 125' for complications excluding reoperation, or clarify '1 in 81' includes reoperation.

Minor Issues

ACC-007 [statistical_plausibility] - Overall 90-day risk reported as '12 per 1000' but Table 2 shows 1.23% = 12.3 per 1000. Minor rounding inconsistency. - Recommendation: Use consistent precision when converting between percentages and rates per 1000.

ACC-008 [narrative_alignment] - Text states 'lower respiratory tract infections...were still infrequent (3 in 1000, 3.1 to 3.5)'. Table 3 shows 0.33% = 3.3 per 1000. Text says '3 in 1000' but should be 3.3. - Recommendation: Consider changing '3 in 1000' to '3.3 per 1000' for consistency with CI bounds.

ACC-009 [narrative_alignment] - Text states 'likelihood of undergoing any further surgery was 3.8% overall'. Table 4 shows 3.82%. Minor rounding. - Recommendation: Consider using 3.82% in text for consistency with table.

ACC-010 [presentation] - CI notation is inconsistent. Tables use square brackets [95% CI] while text uses parentheses (95% CI). - Recommendation: Use consistent CI notation throughout - either parentheses or square brackets.

ACC-011 [missing_content] - Manuscript references supplementary materials multiple times but these were not provided for review. - Recommendation: Provide supplementary materials for complete verification.

Table Verification Status

- **Table 1**: ' Passed - All row totals sum correctly to N=288,250. Procedure-specific Ns sum to total. Age groups, sex, Charlson Index, length of stay all verified.
- **Table 2**: ' Passed - Percentage calculations appear correct. Issues relate to text-table concordance when converting units.
- **Table 3**: ' Passed - Regression table for any reoperation. Sample sizes differ from Tables 1 and 4. Infection-specific results cited in text not present.
- **Table 4**: ' Passed - All counts sum correctly: procedure Ns sum to 258,363; reoperation counts sum to 9,877. Percentage calculations verified.

Scientific Technical Writer

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

Journal Article Review: Manuscript Under Review

Summary Assessment

This population-based cohort study examining arthroscopic shoulder procedures is generally well-written with clear organization and logical flow. The primary writing concerns center on inconsistent number formatting throughout the manuscript, mismatched abbreviation definitions, and occasional redundant phrasing. These issues are predominantly minor and do not impede comprehension of the research findings, though addressing them would enhance the manuscript's professional presentation and readability.

Major Concerns

1. **Abbreviation Mismatch for ASAD** — Abstract, Participants section: - Problem: The abbreviation "ASAD" (arthroscopic subacromial decompression) is introduced with an incomplete expansion—"subacromial decompression (ASAD)"—which omits "arthroscopic" from the spelled-out term despite its inclusion in the abbreviation. - Recommendation: Revise to "arthroscopic subacromial decompression (ASAD)" to ensure the expansion matches the abbreviation exactly.

2. **Inconsistent ACJ Abbreviation Introduction** — Abstract, Participants section: - Problem: The acromioclavicular joint is abbreviated as "(AC)" upon first use, but "ACJ" appears to be used later in the manuscript, creating inconsistency. - Recommendation: Introduce the abbreviation as "acromioclavicular joint (ACJ)" at first use and maintain this abbreviation throughout.

Minor Issues

- **Title**: Number "261248" should include comma separators !’ "261,248"
- **Abstract, Participants section**: Numbers "288 250" and "261 248" use space separators instead of standard comma formatting !’ "288,250" and "261,248"
- **Throughout manuscript**: Verify all large numbers follow consistent comma-separator formatting per journal style guidelines

Strengths

- Clear and logical organization following standard research article structure
- Appropriate use of field-specific terminology for an orthopedic/epidemiological audience
- Concise abstract that effectively conveys study scope and key parameters
- Appropriate identification of study design in the title

Questions for Authors

- Please confirm the journal's preferred number formatting style (comma separators vs. spaces) and apply consistently throughout
- Please verify that all abbreviations are defined at first use and used consistently thereafter

Recommendation

Minor Revision

Justification: The manuscript demonstrates good overall writing quality with no critical errors that impede understanding. The issues identified are primarily formatting inconsistencies and abbreviation mismatches that can be readily corrected. Once these minor revisions are addressed, the manuscript's writing will meet publication standards.