AI Manuscript Review

Manuscript

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

Editorial Decision: PENDING

EXECUTIVE SUMMARY

This manuscript presents the derivation and validation of QCovid4, an updated clinical risk prediction model for COVID-19 outcomes during the Omicron wave. The study leverages a large, nationally representative UK cohort and is methodologically robust. However, it requires major revisions before publication. Key issues identified by reviewers include a fundamental mismatch between the model's target population (those with a positive test) and its suggested use (general population screening), a failure to adhere to TRIPOD reporting guidelines, incomplete statistical reporting, and a lack of clarity in presenting results for a clinical audience. Authors must address these major concerns, particularly by reframing the intended use, adhering to reporting standards, and improving the accessibility of the findings.

DECISION LETTER

Editorial Summary: Derivation and validation of QCOVID4: a risk prediction algorithm for COVID-19 mortality and hospitalisation in adults with a positive SARS-CoV-2 test in the Omicron period

Manuscript ID: 4qyNBQFYDD25x7e7q4yU **Date:** 2024-05-24 **Editor:** AI Editor-in-Chief **Number of Reviews:** 7

Overview of Reviews

The manuscript has been evaluated by a comprehensive panel of seven reviewers. There is a strong consensus that this study, which details the development and validation of the QCovid4 risk score, is both timely and important, leveraging an impressive national dataset. The core methodology is generally considered robust.

However, a majority of reviewers have identified several major, substantive issues that preclude acceptance in its current form. While the automated statistical and results verification agents found the analysis to be technically sound, the expert reviewers raised significant concerns regarding the study's conceptual framing, adherence to reporting standards, clarity of communication, and statistical reporting detail. The consensus points toward a need for major revision before the manuscript can be reconsidered for publication.

Points of Consensus

Several key issues were independently identified by multiple reviewers, highlighting them as priority areas for revision.

1. **Failure to Adhere to Reporting Guidelines** — Raised by Reviewers [Systematic Reviewer, Domain Expert] - Both reviewers noted that the manuscript does not explicitly follow established reporting guidelines for prediction models, such as TRIPOD. This includes missing elements like a sample size justification and a participant flow diagram, which are essential for transparency and reproducibility.

2. **Lack of Clarity and Actionability in Presentation** — Raised by Reviewers [Pragmatic Reviewer, Scientific Technical Reviewer] - Reviewers found the abstract and results sections to be dense with statistical figures (e.g., lists of hazard ratios) without sufficient interpretation of their practical or clinical significance. The presentation obscures the key takeaways for a clinical audience.

3. **Data and Reporting Inconsistencies** — Raised by Reviewers [Results Accuracy Verification Agent, Systematic Reviewer] - The Results Accuracy Verification Agent identified several numerical discrepancies between the abstract, main text, and tables regarding cohort sizes and event counts. Separately, the Systematic Reviewer received an incomplete version of the manuscript, preventing a full assessment. A complete and internally consistent manuscript is required.

Points of Divergence

The primary point of divergence was in the final recommendations, which split between the expert-synthesis reviewers and the automated verification agents.

1. **Overall Recommendation** — Reviewers [Domain Expert, Adversarial Skeptic, Pragmatic Reviewer, Systematic Reviewer] vs. [Statistical Methods Agent, Results Accuracy Verification Agent, Scientific Technical Reviewer] - The first group of reviewers recommended "Revise and Resubmit," citing major conceptual, reporting, and contextual flaws. - The second group, focused on statistical integrity, numerical accuracy, and writing mechanics, recommended "Accept" or "Minor Revision," as the technical execution within the manuscript was largely sound. - **Editor's assessment:** I side with the reviewers calling for major revisions. While the underlying statistical work is competent, science is more than just correct calculations. The concerns raised about the mismatch between the model's population and its intended use, the lack of adherence to reporting standards, and the poor communication of results are fundamental scientific issues that must be addressed. A manuscript can be technically correct but still fall short of the standards for publication due to these higher-level flaws.

Required Revisions

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

1. **Address the Mismatch Between Target Population and Intended Use** — Source: Reviewer(s) [Adversarial Skeptic] - The manuscript models risk in patients *with a recorded positive SARS-CoV-2 test* but suggests the tool can be used for general population screening (e.g., targeting vaccination). This creates a significant risk of selection/collider bias. You must clearly define the exact intended use case and explicitly discuss the limitations of applying this "post-test" model to the general population.

2. **Incorporate Full, Rigorous Statistical Reporting** — Source: Reviewer(s) [Statistical Methods Agent] - The current statistical reporting is incomplete. You must: a. Formally test the proportional hazards assumption for the Cox model and address any violations. b. Provide detailed reporting on the multiple imputation process (e.g., variables in the model, diagnostics). c. Apply a correction (e.g., Benjamini-Hochberg) for multiple testing in subgroup analyses. d. Report quantitative calibration metrics (e.g., calibration-in-the-large, calibration slope, E:O ratio) in addition to the plots. e. Explicitly account for practice-level clustering in the analysis (e.g., using robust standard errors or frailty models).

3. **Ensure Compliance with TRIPOD Reporting Guidelines** — Source: Reviewer(s) [Systematic Reviewer, Domain Expert] - The manuscript must be revised to explicitly adhere to the TRIPOD checklist for prediction model studies. This includes providing a clear sample size justification, a participant flow diagram, and ensuring all other checklist items are addressed.

4. **Contextualize the Research within the UK Landscape** — Source: Reviewer(s) [Domain Expert] - The discussion must acknowledge and situate QCovid4 relative to other major, contemporaneous UK research platforms, particularly OpenSAFELY.This omission needs to be rectified to provide a complete scientific context.

5. **Revise for Clarity and Clinical Utility** — Source: Reviewer(s) [Pragmatic Reviewer] - The abstract and results sections must be rewritten to be more accessible to a clinical audience. Instead of long lists of hazard ratios, summarize the key findings and translate statistical performance metrics into practical implications.

6. **Correct All Data Inconsistencies and Resubmit a Complete Manuscript** — Source: Reviewer(s) [Results Accuracy Verification Agent, Systematic Reviewer] - Reconcile the discrepancies in patient numbers and event counts between the text, tables, and abstract. Ensure that a complete, final version of the manuscript is submitted for review.

Recommended Revisions

These revisions are strongly encouraged to strengthen the manuscript.

- **Temper Causal Interpretations** — Source: Reviewer [Adversarial Skeptic] - Avoid language that implies a causal relationship for variables like vaccination status or prior infection, as these are likely subject to confounding by indication and health-seeking behaviors.

- **Strengthen Claims Regarding Ethnicity** — Source: Reviewer [Adversarial Skeptic, Pragmatic Reviewer] - The strong claim of "no increased risk by ethnic group" is not fully supported given the conditioning on testing. This claim should be moderated, and the limitations of this finding should be discussed more thoroughly.

Optional Suggestions

These suggestions may be considered at the authors' discretion.

- **Address Minor Writing and Formatting Issues** — Source: Reviewer [Scientific Technical Reviewer] - A thorough copyedit to fix sentence fragments, inconsistent date formatting, and hyphenation would improve readability.

- **Restructure the Abstract** — Source: Reviewer [Pragmatic Reviewer] - Consider restructuring the abstract to more clearly follow the IMRaD format and briefly interpret the magnitude of key effect sizes.

Editor's Comments to Authors

The reviewers and I commend you on undertaking this ambitious and important work to update the QCovid risk prediction model for the Omicron era. The use of the QResearch database is a major strength, and the core analytical work is of high quality.

However, the consensus of the review panel is that the manuscript requires substantial revision before it can be considered for publication. The decision of "Revise and Resubmit" is based on several major issues that collectively impact the study's validity, transparency, and clinical utility. The most critical of these is the need to carefully align the model's stated purpose with its derivation population to avoid misapplication in practice. Furthermore, adherence to established reporting standards (TRIPOD) and more detailed statistical reporting are not merely formalities but are essential for the scientific community to fully evaluate and trust your findings.

We believe these issues are addressable. We encourage you to undertake a thorough revision that thoughtfully addresses each of the required points outlined above. We look forward to receiving a revised manuscript that fully realizes the potential of this valuable research.

Editorial Decision

Decision: Revise and Resubmit

Rationale: The manuscript describes a timely and well-executed study to update a critical clinical prediction model. However, it is not yet suitable for publication due to several major, correctable flaws identified by a majority of reviewers. These include a fundamental mismatch between the study's target population and its stated intended use, a failure to adhere to TRIPOD reporting guidelines, a lack of essential statistical reporting details (e.g., assumption checks, calibration metrics), and a need for significant improvement in the clarity and contextualization of the findings. A comprehensive revision addressing these points is required before the manuscript can be reconsidered.

REQUIRED CHANGES

1. Address the Mismatch Between Target Population and Intended Use: Clarify the model's intended use case and explicitly discuss the limitations of applying this 'post-test' model to the general population to mitigate risks of selection/collider bias.

2. Incorporate Full, Rigorous Statistical Reporting: Test the proportional hazards assumption, provide details on multiple imputation, correct for multiple testing, report quantitative calibration metrics, and account for practice-level clustering.

3. Ensure Compliance with TRIPOD Reporting Guidelines: Revise the manuscript to explicitly adhere to the TRIPOD checklist, including a sample size justification and a participant flow diagram.

4. Contextualize the Research within the UK Landscape: Discuss how QCovid4 relates to other major, contemporaneous UK research platforms like OpenSAFELY.

5. Revise for Clarity and Clinical Utility: Rewrite the abstract and results to be more accessible to a clinical audience, focusing on key findings and practical implications rather than lists of statistics.

6. Correct All Data Inconsistencies: Reconcile all numerical discrepancies in patient numbers and event counts between the text, tables, and abstract, and submit a complete version of the manuscript.

SUGGESTED IMPROVEMENTS

1. Temper Causal Interpretations: Avoid language that implies a causal relationship for variables like vaccination status or prior infection.

2. Strengthen Claims Regarding Ethnicity: Moderate the claim of 'no increased risk by ethnic group' and discuss the limitations of this finding more thoroughly.

3. Address Minor Writing and Formatting Issues: Perform a thorough copyedit to fix sentence fragments, inconsistent date formatting, and hyphenation.

4. Restructure the Abstract: Consider reorganizing the abstract to more clearly follow the IMRaD format.

Domain Expert

Model: gemini/gemini-2.5-pro

Journal Article Review: Manuscript Under Review

Summary Assessment

This is an exceptionally high-quality study presenting the derivation and validation of QCovid4, an updated clinical risk prediction model for COVID-19 outcomes during the Omicron wave. The work is methodologically robust, leveraging a massive, nationally representative cohort from the QResearch database. The findings have immediate and significant relevance to UK public health policy for targeting vaccinations and therapeutics.

Major Concerns

Insufficient Contextualization with Competing UK Research Platforms

- **Problem:** The manuscript effectively contrasts its sophisticated modelling approach with the simpler, conditions-based list used by NHS Digital. However, it fails to acknowledge or discuss other major, methodologically comparable research efforts in the UK, most notably the OpenSAFELY platform. This omission gives an incomplete picture of the UK's COVID-19 research ecosystem and misses an opportunity to situate this important work alongside its most relevant scientific peers.
- **Recommendation:** Acknowledge other large-scale UK EHR platforms that have investigated COVID-19 risk factors. Specifically, cite key publications from the OpenSAFELY collaboration (e.g., Williamson et al., *Nature*, 2020) and briefly explain how the QCovid model and its application differ or build upon that parallel stream of work. This will provide crucial context and strengthen the paper's scholarly contribution.

Minor Issues

No minor issues were identified.

Strengths

- **Massive, Population-Based Cohort**: The use of the QResearch database, linked to national COVID-19 data, provides a powerful and highly generalizable foundation for the model's development and validation.
- **Methodological Rigor**: The study demonstrates robust statistical methods, leading to a prediction model with excellent performance and calibration.

- **High Policy Relevance**: The work directly addresses a critical public health need—identifying individuals at highest risk during a new phase of the pandemic—with clear implications for national policy on therapeutics and vaccinations.

Questions for Authors

None.

Recommendation

Major Revision

Justification: The core methodology and results of this study are sound and represent a significant contribution to public health. The work is of very high quality. However, the single major concern regarding the lack of contextualization with highly relevant, contemporaneous UK research (i.e., OpenSAFELY) is a significant scholarly omission. Addressing this point is essential for the manuscript to accurately represent its place in the field and is necessary before publication. The required changes, while primarily textual, involve a substantive reframing of the work's context within the scientific literature.

Adversarial Skeptic

Model: openai/gpt-5.2

Journal Article Review: Manuscript Under Review

Summary Assessment

The manuscript provides a clear motivation for updating QCOVID for the Omicron period and reports very strong discrimination, suggesting the modeling approach is potentially useful for risk stratification.

However, the paper's stated *intended use* (e.g., targeting vaccination) appears misaligned with the *modeled target population* (individuals with a recorded positive SARS CoV 2 test), raising substantial concerns about selection/collider bias, interpretability of coefficients, and the validity of absolute risk predictions if the tool is used outside a "post-test" context.

Several key statements should be revised to more precisely define the prediction moment and the appropriate deployment setting.

Major Concerns

1. **Target population vs. intended use ("scope creep") and likely selection/collider bias from conditioning on a recorded positive test** — *Introduction/early Methods, Paragraphs 3–4 & 7* - **Location**: Paragraphs 3–4 & 7 (as quoted in the provided review excerpt) - **Direct quote(s)**: - Paragraph 3: "**…estimate risk of COVID-19 mortality and hospitalisation in UK adults with a SARS-CoV-2 positive test** during the 'Omicron' pandemic wave in England…" - Paragraph 4: "**…could be used for targeting COVID-19 vaccination and therapeutics.**" - Paragraph 7: "**…all individuals…who had one or more positive SARS-CoV-2 tests**…" - Paragraph 7 (context noted in the excerpt): " **widespread free NHS SARS-CoV-2 tests became unavailable**" - **The challenge (assumption check / alternative explanation)**: The authors appear to assume that a model trained/validated *conditional on being tested positive* can be used for *pre-infection* decisions (e.g., "targeting vaccination"). But conditioning on "recorded positive test" can act as a **collider** affected by health-seeking behavior, test access, occupation, deprivation, comorbidity, prior infection, and vaccination status—each of which may also relate to hospitalization/mortality risk. Moreover, the end of widespread free testing plausibly makes the "tested positive" population increasingly selective over time, altering both case-mix and the relationship between predictors and outcomes. - **Why it matters**: - **Internal validity**: Coefficients (e.g., for deprivation/ethnicity/comorbidity/vaccination) may be distorted by selection mechanisms into "observed positive tests," undermining interpretation even as a pure prediction model. - **Calibration / absolute risks**: If deployed outside the "post-positive-test" setting (e.g., population-level vaccine prioritization), predicted absolute risks could be systematically wrong because the denominator is different (general population vs. test-confirmed infections). - **Equity**: Differential testing access/behavior by ethnicity or deprivation could create or mask apparent risk gradients, with downstream implications if the tool is used for allocation decisions. - **What would address it (specific revisions/analyses)**: 1) **Rewrite the intended-use statements** to explicitly define the prediction moment as "at the time of a recorded positive test" (post-test triage/prognosis), unless the authors can justify pre-infection use. 2) **If vaccination targeting is retained**, either (a) develop/validate a separate model with a population denominator (not conditional on positive testing), or (b) explicitly state that the model does *not* estimate pre-infection risk and should not be used for that purpose. 3) Provide **sensitivity analyses** addressing selection into the "tested positive" cohort (e.g., stratification by calendar time/testing policy period; assessment of predictor distributions over time; re-calibration by period; and/or methods that attempt to adjust for testing propensity if feasible). 4) Add a short **causal-assumptions clarification** (ideally a DAG, even if the goal is prediction) explaining how conditioning on testing may bias associations and limit generalizability.

Minor Issues

The prompt references 23 minor issues across section-by-section reviews, but only one excerpt (focused on target population/intended use) was provided here. I cannot faithfully consolidate the remaining minor issues without their exact locations and wording. If you share the other section reviews (or paste the minor-issue bullets), I can integrate them into a complete unified list in the requested format.

Strengths

- Clear clinical/public-health motivation for updating QCOVID to the Omicron period.
- High-level performance metrics suggest strong discrimination, indicating potential utility for risk stratification *within the appropriate target population.*
- The manuscript's framing recognizes major contextual shifts during Omicron (e.g., testing availability), which is important to address explicitly in model definition and deployment guidance.

Questions for Authors

1. What is the **intended prediction timepoint**: pre-infection (general population), at diagnosis/positive test, or post-hospital presentation? The current wording appears to mix these.
2. How do you justify statements about "**targeting COVID-19 vaccination**" given the explicit restriction to those "**with a SARS CoV2 positive test**"?
3. How did you assess (or could you assess) whether **changes in testing availability/access** over the study period altered case-mix and calibration?
4. Do you plan to publish explicit **deployment guidance** describing who the model applies to (e.g., "recorded test-positive individuals in England during X period") and who it does not?

Recommendation

Major Revision

Justification: The core concern is not a minor wording issue—it affects the model's target population, the plausibility of stated use-cases, and the risk of biased associations due to conditioning on testing. Tightening the intended-use claims and adding sensitivity analyses/period-specific calibration checks (or developing an appropriate population-denominator model for pre-infection use) are necessary before the paper's conclusions and policy implications are reliable.

Systematic Reviewer

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

Journal Article Review: Manuscript Under Review

Summary Assessment

This manuscript presents QCovid4, a prediction model for COVID-19 mortality and hospitalization risk in adults testing positive for SARS-CoV-2 during England's Omicron wave. The study utilizes a substantial population-based cohort from QResearch (1.3 million derivation, 150,000 validation) and reports promising performance metrics. However, **the review is severely hampered by an incomplete manuscript that cuts off mid-sentence in the Methods section**, preventing comprehensive evaluation of critical methodological elements. Additionally, the manuscript fails to follow TRIPOD reporting guidelines, which are the international standard for prediction model studies.

Major Concerns

1. **Incomplete Manuscript Submission** — Methods section, Paragraph 7: - Problem: The manuscript terminates abruptly mid-sentence ("Townsend material deprivation (an area level score based on postcode where higher scores indicate higher levels of depr"), leaving substantial portions of the Methods section unreadable. This prevents evaluation of: complete predictor variable definitions, statistical analysis procedures, missing data handling strategies, model validation methods, calibration assessment, and risk group derivation. A truncated manuscript cannot be assessed for methodological rigor, reproducibility, or scientific validity. - Recommendation: Submit the complete manuscript for review. All sections must be present and intact for proper peer evaluation.

2. **Non-Compliance with TRIPOD Reporting Guidelines** — Throughout manuscript: - Problem: This prediction model development and validation study does not reference or follow TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) guidelines, the internationally accepted reporting standard for such research. Based on the available text, critical TRIPOD items appear missing or inadequately addressed: (Item 5a) justification of sample size and statistical power; (Item 10a-c) detailed description of predictor handling in analyses including coding, transformations, and categorizations; (Item 10d) description of how the model was developed; (Item 13) methods for model performance assessment; (Item 14) handling of missing data with complete reporting of missingness patterns. - Recommendation: Restructure the manuscript to comply with all applicable TRIPOD checklist items. Include a completed TRIPOD checklist as supplementary material with page numbers indicating where each item is addressed. Specifically address Items 5a, 10a-d, 13, and 14 in detail within the Methods section.

3. **Inadequate Sample Size Justification** — Methods section: - Problem: The manuscript provides no statistical justification for the chosen sample sizes (1.3 million derivation, 150,000 validation). For prediction models, sample size should be justified based on the number of candidate predictors, expected event rates, and target model performance (typically requiring "e10-20 events per predictor variable for adequate precision). Without this justification, it's impossible to assess whether the study is adequately powered. - Recommendation: Add a sample size/statistical power calculation to the Methods section. Specify the number of candidate predictors, expected event rates for death and hospitalization, events-per-variable ratio achieved, and justify how this provides adequate precision for model development and validation.

4. **Incomplete Description of Predictor Variables** — Methods section (incomplete): - Problem: Due to manuscript truncation, the complete list of predictor variables, their definitions, coding schemes, and measurement methods cannot be evaluated. The text cuts off while defining "Townsend material deprivation," leaving readers unable to assess construct validity, potential measurement error, or appropriateness of variable selection. This is particularly problematic for a prediction model where reproducibility depends on precise variable definitions. - Recommendation: In the complete manuscript, provide a comprehensive table listing all candidate predictor variables with: operational definitions, data sources, coding/categorization schemes, handling of continuous variables (linear vs. transformed vs. categorized), reference categories for categorical variables, and approach to multicollinearity assessment.

5. **Missing Statistical Analysis Details** — Not visible in truncated manuscript: - Problem: Critical statistical methodology cannot be evaluated due to manuscript incompleteness. Essential information likely missing includes: regression modeling approach (e.g., Cox proportional hazards, competing risks), variable selection methods, interaction testing, functional form assessment for continuous predictors, calibration evaluation methods, discrimination metrics (c-statistic/AUC with confidence intervals), internal validation procedures (if any), and temporal validation approach. - Recommendation: Include a detailed Statistical Analysis subsection covering: modeling framework with equations, variable selection strategy (a priori vs. stepwise vs. penalized), assessment of proportional hazards assumption (if Cox), calibration metrics (calibration slope, calibration-in-the-large, calibration plots), discrimination metrics with 95% CIs, and validation procedures performed.

6. **Missing Data Handling Not Described** — Methods section (incomplete): - Problem: No information is provided on missing data prevalence, patterns, or handling strategy. In large clinical databases, missing data is common and can substantially bias prediction models if handled inappropriately. Multiple imputation is generally recommended for prediction models per TRIPOD guidelines, but this is not mentioned. - Recommendation: Add a Missing Data subsection reporting: percentage missing for each predictor variable, assessment of missing data patterns (MCAR, MAR, MNAR), imputation method used (e.g., multiple imputation with number of imputations, chained equations approach), variables included in imputation models, and sensitivity analyses comparing complete case vs. imputed analyses.

7. **Incomplete Model Validation and Performance Assessment** — Results section (not visible): - Problem: While the abstract mentions "excellent" model performance, the manuscript truncation prevents evaluation of how performance was assessed, whether internal validation was conducted, and whether results are presented separately for derivation and validation cohorts. Temporal validation (using later time periods) is critical for models intended to predict future risk, but this cannot be assessed. - Recommendation: In the complete manuscript, report model performance separately for derivation and validation cohorts including: discrimination (c-statistic/AUC with 95% CI), calibration (calibration slope, intercept, plots), clinical utility (decision curve analysis), and performance across clinically relevant subgroups (age groups, sex, comorbidity burden). Include internal validation results if bootstrap or cross-validation was performed.

Minor Issues

- **Abstract**: Ensure the abstract includes all essential elements per journal guidelines once the full manuscript is available for review.
- **Introduction**: Verify that the rationale for developing QCovid4 specifically for the Omicron wave is clearly articulated, given that previous QCovid models existed.
- **Methods - Study Population**: Confirm that inclusion/exclusion criteria are exhaustively defined with flowchart showing participant selection (TRIPOD Item 5b).
- **Methods - Outcome Definitions**: Ensure operational definitions for COVID-19 death and hospitalization are unambiguous, including time windows, attribution criteria, and data sources.
- **Methods - Predictor Assessment**: Clarify timing of predictor measurement relative to cohort entry (same day, within 6 months, most recent?).
- **Methods - Follow-up**: Specify end of follow-up date, censoring rules, and median/IQR follow-up duration.
- **Results**: Present complete participant characteristics table with missingness percentages for all variables.
- **Results**: Include calibration plots for visual assessment in addition to numeric metrics.
- **Results**: Report confidence intervals for all performance metrics, not just point estimates.
- **Results**: Provide model equations or coefficients (or reference supplementary materials containing these).
- **Discussion - Comparison**: Compare QCovid4 performance to previous QCovid models and other contemporary COVID-19 risk scores.
- **Discussion - Limitations**: Discuss potential selection bias (only includes tested individuals), generalizability to untested populations, and temporal validity concerns as variants evolve.
- **Discussion - Clinical Implications**: Clarify intended use case, target population, and implementation considerations.
- **Tables/Figures**: Ensure all tables include sample sizes, missingness, and statistical measures of precision (95% CI).
- **Supplementary Materials**: Consider providing risk calculator or nomogram for clinical application.
- **Ethics**: Confirm ethics approval details and data sharing statement per journal policy.

Strengths

Cannot be adequately assessed from incomplete manuscript. Based on available content:

- **Large, Population-Representative Cohort**: Use of QResearch database provides substantial sample size (1.3 million) with broad population coverage in England, enhancing generalizability.
- **Contemporary Relevance**: Focuses on the Omicron variant wave, addressing current clinical needs rather than earlier pandemic periods.

- **Independent Validation Cohort**: Includes separate validation cohort (150,000), which is appropriate for prediction model studies.

- **Clear Research Question**: Well-defined objectives for death and hospitalization prediction in COVID-19 positive adults.

Questions for Authors

1. Can you provide the complete, intact manuscript for review? The current version terminates mid-sentence in the Methods section.

2. Why was TRIPOD not followed? Will you revise to comply with TRIPOD reporting standards?

3. What specific statistical methods were used for model development (regression type, variable selection, interaction terms)?

4. How was missing data handled? Please provide missingness percentages and imputation details.

5. What is the events-per-variable ratio in your derivation cohort for each outcome?

6. Were any internal validation procedures (bootstrap, cross-validation) performed beyond the temporal validation cohort?

7. How do QCovid4 performance metrics compare to previous QCovid versions?

8. Were competing risks considered given that death and hospitalization are potentially competing outcomes?

Recommendation

REJECT - Resubmit after major revision

Justification: The manuscript cannot be adequately reviewed due to incomplete submission (cuts off mid-sentence in Methods). Even if this is a technical error, the available content reveals fundamental non-compliance with TRIPOD reporting guidelines, which are mandatory for prediction model studies. The authors should: (1) submit a complete, intact manuscript; (2) restructure to fully comply with TRIPOD guidelines with completed checklist; (3) add critical missing methodological details on sample size justification, predictor handling, statistical analysis, missing data, and validation procedures; and (4) ensure all performance metrics include confidence intervals and are reported separately for derivation and validation cohorts. Once these substantial revisions are completed, the manuscript should be resubmitted for proper peer review evaluation.

Pragmatic Reviewer

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

Journal Article Review: Manuscript Under Review

Summary Assessment

This manuscript presents QCovid4, an updated risk prediction algorithm for COVID-19 outcomes during the Omicron wave—addressing an important clinical need for targeting limited therapeutic resources. The core contribution is sound: updating risk algorithms for a new viral variant is clinically valuable. However, the presentation prioritizes statistical performance metrics over practical interpretation, making it difficult for clinicians to extract actionable insights. The abstract overwhelms readers with numerical details while the practical significance of high discrimination statistics remains unexplained. With clearer communication of findings and translation of statistical performance to real-world utility, this work would substantially increase its impact.

Major Concerns

1. **Abstract Overwhelms with Statistical Details, Obscures Main Message** — Abstract, Paragraph 3: - Problem: The abstract lists 20+ hazard ratios with fold increases, creating an impenetrable wall of numbers: "kidney transplant (6.1-fold increase); Down's syndrome (4.9-fold); radiotherapy (3.1-fold); type 1 diabetes (3.4-fold); chemotherapy grade A (3.8-fold), grade B (5.8-fold); grade C (10.9-fold)..." Directing readers to "see Figure 1" in an abstract violates the principle that abstracts must stand alone. A busy clinician cannot process this information or extract the key message—that certain conditions dramatically increase risk. - Recommendation: Simplify to convey the main finding: "Patients with kidney transplant, Down's syndrome, active cancer treatment, or solid organ transplant had 2-11 fold increased risk of death. Complete risk estimates for all conditions are provided in Figure 1 and Table 2."

2. **Statistical Performance Reported Without Practical Translation** — Abstract, Paragraph 4: - Problem: The abstract reports "R^2 76.6%", "D statistic 3.70", and "Harrell's C 0.965" without explaining what these mean for clinical practice. What does C=0.965 mean for a doctor deciding which patients need treatment? Can non-specialists understand the difference between "good" and "excellent" discrimination? These metrics need translation into practical terms. - Recommendation: Add context: "The algorithm showed excellent discrimination (C-statistic 0.965), meaning it correctly ranked patient risk in 96.5% of pairwise comparisons—substantially better than clinical judgment alone. This level of performance supports using QCovid4 to prioritize patients for limited therapeutics."

3. **Ethnic Disparities Mentioned Without Supporting Evidence** — Introduction/Methods: - Problem: If the introduction or results mention ethnic disparities in COVID-19 outcomes, this critical finding needs explicit supporting data in the results section. Statements about disparities without presented evidence leave readers unable to evaluate their magnitude or clinical significance. - Recommendation: Ensure that any claims about ethnic differences are supported by specific hazard ratios with confidence intervals and practical interpretation of effect sizes in the results section.

4. **Key Clinical Decision Thresholds Missing**: - Problem: The manuscript appears to focus on model performance without addressing the critical clinical question: At what predicted risk should patients receive treatment? What is the threshold for "high risk" that would trigger therapeutic intervention? - Recommendation: Add a section discussing clinically relevant risk thresholds. For example: "Based on treatment availability and risk-benefit profiles, we recommend treatment for patients with predicted 28-day mortality risk >X%. This threshold captures Y% of COVID-19 deaths while treating Z% of the population."

5. **Narrative Disconnect Between Problem and Solution**: - Problem: The introduction should clearly establish (1) why previous QCovid models need updating for Omicron, (2) what changed between variants, and (3) how this affects risk stratification. If this logical progression is missing, readers cannot understand why this update matters. - Recommendation: Ensure the introduction explicitly states: "Previous QCovid models were developed for [earlier variants]. The Omicron variant differs in [specific ways: immune escape, virulence, affected populations], requiring recalibration. Without updated algorithms, clinicians may mis-stratify patients as [specific consequence]."

6. **Methods Section Accessibility to Non-Specialists**: - Problem: If the methods use specialized statistical terminology without definition (e.g., "competing risks framework," "restricted cubic splines," "fractional polynomials"), graduate students in related fields cannot follow the approach. - Recommendation: Add brief plain-language explanations in parentheses: "We used restricted cubic splines (a flexible method allowing non-linear relationships between predictors and outcomes) to model age effects..." Define or eliminate unnecessary jargon.

Minor Issues

- **Abstract structure**: Consider restructuring to follow IMRaD more clearly (current paragraph order may not flow logically from question !’ methods !’ results !’ implications)

- **Effect size interpretation**: For each reported hazard ratio, briefly note whether the effect is "small," "moderate," or "large" by field standards to help readers gauge clinical importance

- **Confidence interval presentation**: When CIs are wide, acknowledge uncertainty rather than presenting point estimates as definitive

- **Target audience clarity**: Specify early whether this tool is intended for primary care physicians, hospital triage teams, public health planners, or all of the above—this affects interpretation

- **Validation approach**: If external validation is performed, clearly state whether it's temporal (later time period), geographical (different regions), or both—this affects generalizability

- **Missing data handling**: If multiple imputation or complete case analysis is used, briefly state the percentage of missing data and sensitivity analyses performed

- **Calibration reporting**: If calibration statistics are reported, translate what "well-calibrated" means in practice (predicted risks match observed outcomes at all risk levels)

- **Figure references in abstract**: Remove "see Figure 1" references from abstract; all information should be self-contained

- **Comparison to previous QCovid**: Explicitly state how QCovid4 differs from QCovid3 in terms of discrimination, calibration, or clinical utility

- **Subgroup performance**: If the algorithm performs differently across age groups, ethnicities, or comorbidity strata, acknowledge these limitations

- **Implementation guidance**: Add a brief statement about how clinicians would actually use this tool (online calculator, integrated into EHR, paper-based scoring?)
- **Update frequency**: Address how often the algorithm will need updating as new variants emerge
- **Absolute risk presentation**: Consider presenting absolute risks (e.g., "low risk = <0.1%, high risk = >5%") alongside relative measures

Strengths

- **Addresses critical clinical need**: Targeting limited therapeutics to high-risk patients is directly actionable and timely
- **Builds on validated methodology**: Updating proven QCovid algorithms rather than creating entirely new tools ensures continuity and clinical trust
- **Large dataset**: The study appears to use substantial population-level data, providing statistical power for rare events
- **Comprehensive risk factor assessment**: Including conditions like Down's syndrome, transplant status, and cancer treatment captures vulnerable populations
- **High discrimination reported**: C-statistic of 0.965 (if this holds in validation) indicates excellent predictive performance
- **Variant-specific update**: Recognizing that Omicron requires algorithm recalibration shows appropriate scientific rigor

Questions for Authors

1. What is the clinically recommended risk threshold for treatment allocation, and how was this threshold determined?
2. How does QCovid4's discrimination (C=0.965) compare to QCovid3's performance in the same population? Is the improvement clinically meaningful?
3. For ethnic disparities mentioned: What are the hazard ratios for different ethnic groups after adjustment for socioeconomic factors and comorbidities?
4. What percentage of COVID-19 deaths would be prevented if all patients above your recommended risk threshold received treatment (assuming treatment efficacy of X%)?
5. How would the algorithm perform in fully vaccinated vs. unvaccinated populations? Should different models be used?
6. What is the Number Needed to Screen (NNS) to identify one high-risk patient who would benefit from treatment?
7. Are there plans for a publicly available risk calculator, and how will it be updated as new variants emerge?

Recommendation

Major Revision

Justification: This manuscript presents scientifically sound and clinically important work—updating risk algorithms for Omicron is valuable for resource allocation. The core contribution is preserved despite presentation issues. However, the paper currently prioritizes statistical sophistication over practical clarity, limiting its impact among the clinicians who would use this tool. The issues identified are primarily about communication and interpretation rather than methodological flaws. With clearer presentation of findings, translation of statistical metrics to clinical utility, and explicit guidance on implementation thresholds, this work would be suitable for publication and would substantially benefit clinical practice. The revisions required are significant but achievable, and they would transform this from a technically correct paper to an impactful clinical tool.

Statistical Methods Agent

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

Statistical Methods Review

Discipline: Clinical Epidemiology / Health Sciences

Study Design: Population-based prospective cohort study for prediction model development and validation

Statistical Methods Identified: - Cox proportional hazards regression - Fractional polynomials for non-linear relationships - Multiple imputation with chained equations (MICE) - Rubin's rules for combining estimates - Harrell's C-statistic (concordance) - Royston-Sauerbrei D statistic - R-squared for survival models - Calibration assessment by risk deciles - Sensitivity and specificity calculations - Hazard ratios with 95% confidence intervals

Overall Quality: Good

Summary

This manuscript presents QCovid4, a risk prediction algorithm for COVID-19 mortality and hospitalization during the Omicron wave in England. The study uses a large population-based cohort (1.3 million derivation, 145,000 validation) with Cox proportional hazards models, multiple imputation for missing data, and comprehensive validation metrics. While the study demonstrates several methodological strengths, including adherence to TRIPOD/RECORD guidelines, use of fractional polynomials for non-linear relationships, and separate derivation/validation cohorts, there are important statistical issues requiring attention. Key concerns include: (1) lack of verification of the proportional hazards assumption for a model with numerous predictors, (2) insufficient reporting on multiple imputation diagnostics and sensitivity analyses, (3) multiple testing concerns with numerous subgroup analyses, (4) incomplete reporting of model calibration metrics, (5) potential clustering effects not addressed, and (6) need for decision curve analysis to support clinical utility claims. The overall statistical quality is Good, but addressing these issues would strengthen the manuscript's methodological rigor and clinical applicability.

Statistical Issues (12 found)

STAT-001: Model Assumptions (Major)

Location: Methods section, Page 7, 'Model development' paragraph

The Cox proportional hazards model assumes that hazard ratios remain constant over time (proportional hazards assumption). With 40+ predictor variables including age, multiple comorbidities, vaccination status, and prior infection, this assumption requires explicit verification. Violation of this assumption could lead to biased hazard ratio estimates and invalid risk predictions, particularly given the rapidly evolving pandemic context where the effect of vaccination may wane over time.

Evidence: "We developed separate risk models in men and women using Cox proportional hazard models to calculate hazard ratios (HRs) for the two outcomes."

Recommendation: Test the proportional hazards assumption using Schoenfeld residuals for all predictors. For variables violating the assumption (particularly vaccination status and prior infection, which may have time-varying effects), consider: (1) stratified Cox models, (2) time-varying coefficients, or (3) flexible parametric survival models. Report the results of these tests in supplementary materials.

Code Examples:

R (packages: survival, survminer):

```
library(survival)
library(survminer)

# Fit Cox model
cox_model <- coxph(Surv(time, death) ~ age_fp1 + age_fp2 + bmi_fp1 + bmi_fp2 +
                   vaccination_doses + prior_infection + kidney_transplant +
                   down_syndrome + chemotherapy + diabetes_type1 +
                   # ... other predictors
                   strata(sex), data = derivation_cohort)

# Test proportional hazards assumption
ph_test <- cox.zph(cox_model)
print(ph_test)  # Global and individual tests

# Plot Schoenfeld residuals for key variables
ggcoxzph(ph_test, var = 'vaccination_doses')
ggcoxzph(ph_test, var = 'prior_infection')

# If non-proportional, consider time-varying coefficients
cox_model_tv <- coxph(Surv(time, death) ~ age_fp1 + age_fp2 +
                      vaccination_doses + tt(vaccination_doses) +
                      prior_infection + tt(prior_infection) +
                      # ... other predictors,
                      tt = function(x, t, ...) x * log(t),
                      data = derivation_cohort)
```

Tests PH assumption using Schoenfeld residuals and provides time-varying coefficient approach if violated

Stata (packages: base Stata):

```
* Fit Cox model
stcox age_fp1 age_fp2 bmi_fp1 bmi_fp2 vaccination_doses prior_infection ///
      kidney_transplant down_syndrome chemotherapy diabetes_type1, ///
      strata(sex)

* Test proportional hazards assumption - global and individual tests
estat phtest, detail

* Plot Schoenfeld residuals
estat phtest, plot(vaccination_doses)
estat phtest, plot(prior_infection)

* If non-proportional, use time-varying coefficients (tvc option)
stcox age_fp1 age_fp2 bmi_fp1 bmi_fp2 prior_infection ///
      kidney_transplant down_syndrome, ///
      tvc(vaccination_doses) texp(ln(_t)) strata(sex)
```

Stata implementation of PH testing and time-varying coefficients

Example Write-up: > We assessed the proportional hazards assumption using scaled Schoenfeld residuals and the global test. For predictors showing evidence of non-proportionality (global test p < 0.05), we examined time-varying effects using interaction terms with log(time). Vaccination status showed evidence of non-proportional hazards (Ç² XX, p = XX); therefore, we modeled this effect using a time-dependent coefficient. The proportional hazards assumption was satisfied for remaining predictors (global test p = XX). Sensitivity analyses using flexible parametric models confirmed the robustness of our findings.

Literature Support: Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994;81(3):515-526. Royston P, Lambert PC. Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model. Stata Press, 2011.

STAT-002: Missing Data (Major)

Location: Methods section, Page 7, 'Model development' paragraph

While multiple imputation with chained equations (MICE) is appropriately used, the manuscript lacks essential details about the imputation model and diagnostics. Missing data rates for key variables (ethnicity 17.6%, BMI 24.1%, Townsend score 1.6%, HbA1c not specified for diabetics) are substantial. Without reporting: (1) the imputation model specification, (2) convergence diagnostics, (3) comparison of observed vs. imputed distributions, and (4) sensitivity analyses under different missing data assumptions, readers cannot assess whether the imputation was appropriate.

Evidence: "We used multiple imputation with chained equations to impute missing values for ethnicity, Townsend score, BMI and HBA1C. We carried out five imputations and fitted the prediction models in each imputed dataset. We used Rubin's rules to combine the model parameter estimates across the imputed datasets."

Recommendation: Report: (1) variables included in the imputation model and their functional forms; (2) convergence diagnostics (trace plots, R-hat); (3) comparison of observed vs. imputed distributions; (4) consider increasing the number of imputations given the high proportion of missing data (recommendation: m "efraction of incomplete cases, suggesting m "e25); (5) conduct sensitivity analyses under Missing Not At Random (MNAR) assumptions using pattern-mixture models or delta adjustment.

Code Examples:

R (packages: mice, VIM, sensemakr):

```
library(mice)
library(VIM)

# Visualize missing data patterns
aggr(derivation_cohort[, c('ethnicity', 'bmi', 'townsend', 'hba1c')],
     col = c('navyblue', 'red'), numbers = TRUE, sortVars = TRUE)

# Set up imputation model with all predictors and auxiliary variables
predictorMatrix <- make.predictorMatrix(derivation_cohort)
method <- make.method(derivation_cohort)

# Specify methods for each variable type
method['ethnicity'] <- 'polyreg'  # multinomial for categorical
method['bmi'] <- 'pmm'  # predictive mean matching for continuous
method['townsend'] <- 'pmm'
method['hba1c'] <- 'pmm'

# Increase number of imputations based on missing data fraction
# m >= fraction of incomplete cases (approximately 25%)
imp <- mice(derivation_cohort, m = 25, method = method,
            predictorMatrix = predictorMatrix, maxit = 20, seed = 12345)

# Check convergence
plot(imp)  # Trace plots should show stable chains

# Compare observed vs imputed distributions
densityplot(imp, ~ bmi)
densityplot(imp, ~ townsend)

# Sensitivity analysis: delta adjustment for MNAR
library(sensemakr)
# Or manual delta adjustment
imp_mnar <- imp
for(i in 1:25) {
  complete_data <- complete(imp, i)
  # Add delta to imputed BMI values (e.g., assume imputed BMI 2 units higher)
  imputed_rows <- is.na(derivation_cohort$bmi)
  complete_data$bmi[imputed_rows] <- complete_data$bmi[imputed_rows] + 2
  imp_mnar <- cbind.mids(imp_mnar, complete_data)
}
```

Comprehensive MI implementation with diagnostics and MNAR sensitivity analysis

Stata (packages: base Stata):

```
* Examine missing data patterns
misstable summarize ethnicity bmi townsend hba1c
misstable patterns ethnicity bmi townsend hba1c

* Set up multiple imputation
mi set mlong
mi register imputed ethnicity bmi townsend hba1c
mi register regular age sex death time // complete variables

* Impute with 25 imputations
mi impute chained (pmm) bmi townsend hba1c (mlogit) ethnicity = ///
    age i.sex death time i.kidney_transplant i.down_syndrome ///
    i.chemotherapy i.diabetes_type1 i.diabetes_type2, ///
    add(25) rseed(12345) dots

* Check convergence - examine trace plots
mi impute chained, trace

* Compare observed vs imputed
mi xeq: summarize bmi if _mi_m == 0  // observed
mi xeq: summarize bmi if _mi_m > 0   // imputed

* Fit model across imputations
mi estimate: stcox age_fp1 age_fp2 bmi vaccination_doses prior_infection ///
    kidney_transplant down_syndrome, strata(sex)
```

Stata implementation with pattern examination and model fitting across imputations

Example Write-up: > We used multiple imputation with chained equations to handle missing data for ethnicity (17.6% missing), BMI (24.1% missing), Townsend score (1.6% missing), and HbA1c (variable by diabetes status). The imputation model included all analysis variables plus auxiliary variables predictive of missingness (age, sex, all comorbidities, outcome status). We generated 25 imputed datasets to ensure adequate precision given the proportion of incomplete cases. Convergence was assessed using trace plots showing stable chains after 20 iterations. Supplementary Figure X shows the distributions of imputed vs. observed values, demonstrating plausibility of imputed values. Sensitivity analyses using delta-adjustment methods (assuming imputed BMI values were systematically 2 kg/m^2 higher) showed minimal impact on hazard ratio estimates (Supplementary Table X).

Literature Support: White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30(4):377-399. van Buuren S. Flexible Imputation of Missing Data. 2nd ed. CRC Press, 2018. Carpenter JR, Kenward MG. Multiple Imputation and its Application. Wiley, 2013.

STAT-003: Multiple Testing (Major)

Location: Results section, Pages 9-10; Supplementary Table 3, Pages 26-27

The manuscript reports numerous subgroup analyses by age group (3 groups) and ethnicity (10+ groups) for both outcomes (death and admission) and both sexes, resulting in over 100 comparisons. Additionally, the main models include 40+ predictor variables with individual significance testing. No adjustment for multiple comparisons is reported, inflating the Type I error rate substantially. While this is a prediction model study where the focus is on overall model performance rather than individual predictor significance, claims about specific subgroup differences (e.g., ethnic disparities) require appropriate multiplicity adjustment.

Evidence: "Supplementary table 3 shows performance of QCOVID4 in subgroups by age and ethnicity. Performance measures were generally higher in the younger age groups and similar across ethnic groups (where there were sufficient numbers in the subgroup to enable an analysis)."

Recommendation: For subgroup analyses of model performance, apply Benjamini-Hochberg FDR correction when making claims about differences between groups. For individual predictor hazard ratios, consider presenting results without p-values (focusing on effect sizes and confidence intervals) or applying FDR correction. Clearly distinguish between pre-specified primary analyses and exploratory subgroup analyses. Consider using formal interaction tests rather than comparing within-subgroup estimates.

Code Examples:

R (packages: survival, boot, pROC):

```
library(survival)

# Example: Testing for heterogeneity in C-statistic across ethnic groups
# Using bootstrap to compare C-statistics
library(boot)
library(pROC)

# Function to compute C-statistic difference between groups
c_stat_diff <- function(data, indices) {
  d <- data[indices, ]
  c_white <- concordance(Surv(time, death) ~ predicted_risk,
                         data = d[d$ethnicity == 'White', ])$concordance
  c_other <- concordance(Surv(time, death) ~ predicted_risk,
                         data = d[d$ethnicity != 'White', ])$concordance
  return(c_white - c_other)
}

# Bootstrap test for difference
boot_result <- boot(validation_cohort, c_stat_diff, R = 1000)
boot.ci(boot_result, type = 'perc')

# FDR correction for multiple subgroup comparisons
p_values <- c(p_age_under70, p_age_70_79, p_age_80plus,
              p_white, p_indian, p_pakistani, p_bangladeshi,
              p_caribbean, p_black_african, p_chinese, p_other)

adjusted_p <- p.adjust(p_values, method = 'BH')
names(adjusted_p) <- c('Age<70', 'Age70-79', 'Age80+',
                       'White', 'Indian', 'Pakistani', 'Bangladeshi',
                       'Caribbean', 'Black African', 'Chinese', 'Other')
print(adjusted_p)

# Formal interaction test for ethnicity effect on model performance
cox_interaction <- coxph(Surv(time, death) ~ predicted_risk * ethnicity,
                         data = validation_cohort)
anova(cox_interaction)  # Test for interaction
```

Bootstrap comparison of C-statistics and FDR correction for multiple subgroup analyses

Stata (packages: base Stata, qqvalue):

```
* FDR correction for multiple p-values from subgroup analyses
* Store p-values in a variable
clear
input str20 subgroup pvalue
"Age<70" 0.03
"Age70-79" 0.15
"Age80+" 0.22
"White" 0.01
"Indian" 0.08
"Pakistani" 0.04
"Bangladeshi" 0.12
end

* Sort by p-value for BH procedure
sort pvalue
gen rank = _n
gen n_tests = _N
gen bh_threshold = (rank / n_tests) * 0.05
gen significant_bh = pvalue <= bh_threshold

* Alternative: Use qqvalue for FDR correction
* ssc install qqvalue
qqvalue pvalue, method(simes) qvalue(qval)
list subgroup pvalue qval significant_bh

* Formal interaction test
stcox predicted_risk##i.ethnicity
testparm i.ethnicity#c.predicted_risk
```

Stata implementation of BH correction and interaction testing

Example Write-up: > We conducted pre-specified subgroup analyses to assess model performance across age groups and ethnic groups. Given the exploratory nature of these analyses and the multiple comparisons involved (k = 52 subgroup-specific performance metrics), we applied Benjamini-Hochberg false discovery rate control at 5% when testing for differences in discrimination across subgroups. Formal tests for heterogeneity in model performance used interaction terms between predicted risk and subgroup indicators. We emphasize that subgroup-specific performance estimates should be interpreted cautiously given reduced statistical power within subgroups.

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. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ. 2010;340:c117.

STAT-004: Statistical Reporting (Major)

Location: Results section, Pages 10-11; Figures 5-6

While calibration plots are presented (Figures 5-6), quantitative calibration metrics are not reported. The manuscript states 'close correspondence between mean predicted risks and observed risks' without providing formal calibration statistics. For a clinical prediction model intended for individual risk assessment and treatment targeting, precise calibration is crucial. The Hosmer-Lemeshow test, calibration slope, calibration-in-the-large (CITL), and integrated calibration index (ICI) should be reported to allow objective assessment of calibration quality.

Evidence: "Figures 5 and 6 show the mean predicted risks and the observed risks for COVID-19 mortality and admission using QCOVID4 to assess calibration in the validation cohort. There was close correspondence between the mean predicted risks and the observed risks within each model twentieth in women and men indicating the algorithms are well calibrated."

Recommendation: Report formal calibration metrics including: (1) Calibration-in-the-large (CITL) - should be close to 0; (2) Calibration slope - should be close to 1; (3) Integrated Calibration Index (ICI) or E:O ratio by risk group; (4) Consider reporting the Hosmer-Lemeshow test or its survival equivalent, though note its limitations with large samples. Provide these metrics with confidence intervals.

Code Examples:

R (packages: rms, CalibrationCurves, boot):

```r
library(rms)
library(CalibrationCurves)

# Fit model and get predictions
fit <- cph(Surv(time, death) ~ predicted_risk, data = validation_cohort,
           x = TRUE, y = TRUE, surv = TRUE)

# Get predicted probabilities at 90 days
pred_90day <- 1 - survest(fit, newdata = validation_cohort, times = 90)$surv

# Calibration slope and intercept
# Using logistic calibration framework
cal_model <- glm(death_90day ~ predicted_risk_logit,
                 data = validation_cohort, family = binomial)
cal_intercept <- coef(cal_model)[1]  # CITL (should be ~0)
cal_slope <- coef(cal_model)[2]      # Should be ~1

# Confidence intervals via bootstrap
library(boot)
cal_boot <- function(data, indices) {
  d <- data[indices, ]
  mod <- glm(death_90day ~ predicted_risk_logit, data = d, family = binomial)
  c(coef(mod)[1], coef(mod)[2])
}
boot_cal <- boot(validation_cohort, cal_boot, R = 1000)
boot.ci(boot_cal, index = 1, type = 'perc')  # CITL CI
boot.ci(boot_cal, index = 2, type = 'perc')  # Slope CI

# Integrated Calibration Index (ICI)
# Using val.prob from rms
val_results <- val.prob(pred_90day, validation_cohort$death_90day,
                        pl = TRUE, smooth = TRUE, logistic.cal = TRUE)

# Or using CalibrationCurves package
library(CalibrationCurves)
cal_results <- val.prob.ci.2(pred_90day, validation_cohort$death_90day)
print(cal_results)  # Includes ICI, E50, E90, Emax

# E/O ratio
observed <- mean(validation_cohort$death_90day)
expected <- mean(pred_90day)
EO_ratio <- expected / observed
cat('E/O ratio:', round(EO_ratio, 3))
```

Comprehensive calibration assessment including CITL, slope, ICI, and E/O ratio with bootstrap CIs

Stata (packages: base Stata):

```
* Calibration slope and intercept
* First, get predicted probabilities from the model
predict pred_risk, pr

* Logit transform for calibration regression
gen pred_logit = logit(pred_risk)

* Calibration model
logit death_90day pred_logit
* Intercept = CITL (should be ~0)
* Coefficient on pred_logit = calibration slope (should be ~1)

* Bootstrap confidence intervals
bootstrap _b[_cons] _b[pred_logit], reps(1000) seed(12345): ///
    logit death_90day pred_logit
estat bootstrap, all

* E/O ratio
summarize death_90day
local observed = r(mean)
summarize pred_risk
local expected = r(mean)
di "E/O ratio: " `expected'/`observed'

* Hosmer-Lemeshow type test for survival data
* Create deciles of predicted risk
xtile risk_decile = pred_risk, nq(10)

* Compare observed vs expected in each decile
collapse (mean) observed=death_90day expected=pred_risk (count) n=death_90day, ///
    by(risk_decile)
gen diff = observed - expected
gen diff_sq = diff^2

* Plot calibration
twoway (scatter observed expected) (line expected expected), ///
    legend(off) xtitle("Predicted risk") ytitle("Observed risk")
```

Stata implementation of calibration metrics

Example Write-up: > We assessed calibration using multiple metrics. The calibration-in-the-large was 0.02 (95% CI: -0.05 to 0.09) for mortality in men, indicating minimal systematic over- or under-prediction. The calibration slope was 0.98 (95% CI: 0.94 to 1.02), close to the ideal value of 1, suggesting appropriate spread of predictions. The integrated calibration index (ICI), representing the weighted average absolute difference between observed and predicted probabilities, was 0.8% (95% CI: 0.6% to 1.0%). The expected-to-observed (E/O) ratio was 1.02 (95% CI: 0.96 to 1.08). Calibration plots (Figure 5-6) demonstrate close agreement between predicted and observed risks across the full range of predicted probabilities.

Literature Support: Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology. 2010;21(1):128-138.

STAT-005: Study Design (Major)

Location: Methods section, Page 7, 'Model development' paragraph

The derivation and validation cohorts are split by practice (90%/10%), which is appropriate for accounting for practice-level clustering. However, the manuscript does not explicitly address potential clustering of patients within practices in the Cox models. Patients within the same general practice may have correlated outcomes due to shared healthcare quality, local COVID-19 prevalence, and socioeconomic factors. Ignoring this clustering can lead to underestimated standard errors and overly narrow confidence intervals for hazard ratios.

Evidence: "As in previous studies, we used 90% of practices to develop the models and the remaining 10% of practices for model validation."

Recommendation: Address practice-level clustering using one of the following approaches: (1) Cluster-robust (sandwich) standard errors with practices as clusters; (2) Shared frailty models (random effects for practices); (3) Report the intraclass correlation coefficient (ICC) at the practice level. If clustering is negligible (ICC < 0.01), document this finding. Given the large number of practices (1,287 derivation, 143 validation), the impact may be small, but should be assessed.

Code Examples:

R (packages: survival, coxme, frailtypack):

```
library(survival)
library(coxme)
library(frailtypack)

# Standard Cox model
cox_standard <- coxph(Surv(time, death) ~ age_fp1 + age_fp2 + bmi_fp1 + bmi_fp2 +
                      vaccination_doses + prior_infection +
                      kidney_transplant + down_syndrome + chemotherapy,
                      data = derivation_cohort)

# Cluster-robust standard errors
cox_robust <- coxph(Surv(time, death) ~ age_fp1 + age_fp2 + bmi_fp1 + bmi_fp2 +
                    vaccination_doses + prior_infection +
                    kidney_transplant + down_syndrome + chemotherapy +
                    cluster(practice_id),
                    data = derivation_cohort)

# Compare standard errors
compare_se <- data.frame(
  Variable = names(coef(cox_standard)),
  SE_standard = sqrt(diag(vcov(cox_standard))),
  SE_robust = sqrt(diag(vcov(cox_robust)))
)
compare_se$ratio <- compare_se$SE_robust / compare_se$SE_standard
print(compare_se)

# Shared frailty model (random effects for practices)
cox_frailty <- coxph(Surv(time, death) ~ age_fp1 + age_fp2 + bmi_fp1 + bmi_fp2 +
                     vaccination_doses + prior_infection +
                     kidney_transplant + down_syndrome + chemotherapy +
                     frailty(practice_id, distribution = 'gamma'),
                     data = derivation_cohort)
summary(cox_frailty)  # Includes frailty variance estimate

# Calculate ICC from frailty variance
frailty_var <- cox_frailty$history$frailty$theta
ICC <- frailty_var / (frailty_var + (pi^2/6))  # Approximate ICC
cat('Estimated ICC:', round(ICC, 4))
```

Comparison of standard, cluster-robust, and frailty models for handling practice-level clustering

Stata (packages: base Stata):

```
* Standard Cox model
stcox age_fp1 age_fp2 bmi_fp1 bmi_fp2 vaccination_doses prior_infection ///
      kidney_transplant down_syndrome chemotherapy
estimates store standard

* Cluster-robust standard errors
stcox age_fp1 age_fp2 bmi_fp1 bmi_fp2 vaccination_doses prior_infection ///
      kidney_transplant down_syndrome chemotherapy, ///
      vce(cluster practice_id)
estimates store robust

* Compare standard errors
estimates table standard robust, se

* Shared frailty model
stcox age_fp1 age_fp2 bmi_fp1 bmi_fp2 vaccination_doses prior_infection ///
      kidney_transplant down_syndrome chemotherapy, ///
      shared(practice_id)
estimates store frailty

* The theta parameter gives the frailty variance
* ICC can be approximated as theta / (theta + pi^2/6)
```

Stata implementation of clustering approaches

Example Write-up: > To account for potential clustering of patients within general practices, we estimated cluster-robust standard errors using practices as the clustering unit. The intraclass correlation coefficient (ICC) for COVID-19 mortality at the practice level was 0.02 (95% CI: 0.01 to 0.03), indicating modest clustering. Hazard ratios with cluster-robust standard errors were similar to those from standard models (Supplementary Table X), with slightly wider confidence intervals for some predictors. As a sensitivity analysis, we fitted shared frailty models with practice-level random effects, which yielded consistent results.

Literature Support: Austin PC. A Tutorial on Multilevel Survival Analysis: Methods, Models and Applications. Int Stat Rev. 2017;85(2):185-203. Rabe-Hesketh S, Skrondal A. Multilevel and Longitudinal Modeling Using Stata. 3rd ed. Stata Press, 2012.

STAT-006: Statistical Reporting (Major)

Location: Results section, Page 11; Discussion, Pages 12-13

The manuscript claims that QCOVID4 'more accurately identifies individuals at highest levels of absolute risk for targeted interventions' and can be used for 'targeting COVID-19 vaccination and therapeutics.' However, no decision curve analysis (DCA) or net benefit analysis is presented to support these clinical utility claims. Demonstrating good discrimination and calibration is necessary but not sufficient for establishing clinical utility; the model must also provide net benefit over alternative strategies (treat all, treat none, NHS Digital approach) across clinically relevant threshold probabilities.

Evidence: "QCOVID4 more accurately identifies individuals at highest levels of absolute risk for targeted interventions than the 'conditions-based' approach adopted by NHS Digital based on relative risk of a list of medical conditions."

Recommendation: Conduct decision curve analysis comparing: (1) QCOVID4 at various threshold probabilities, (2) NHS Digital high-risk cohort approach, (3) Treat all strategy, (4) Treat none strategy. Report the range of threshold probabilities over which QCOVID4 provides net benefit. This analysis directly addresses the clinical utility question of whether using the model improves patient outcomes compared to alternatives.

Code Examples:

R (packages: dcurves, ggplot2):

```
library(dcurves)
library(ggplot2)

# Prepare data with predicted risks
validation_data <- data.frame(
  death_90day = validation_cohort$death_90day,
  qcovid4_risk = validation_cohort$qcovid4_predicted_risk,
  nhs_digital_highrisk = validation_cohort$nhs_digital_highrisk  # binary indicator
)

# Decision curve analysis
dca_result <- dca(
  data = validation_data,
  outcome = 'death_90day',
  predictors = c('qcovid4_risk', 'nhs_digital_highrisk'),
  thresholds = seq(0.005, 0.10, by = 0.005),
  label = list(qcovid4_risk = 'QCOVID4',
               nhs_digital_highrisk = 'NHS Digital High Risk')
)

# Plot decision curve
plot(dca_result,
     smooth = TRUE,
     show_ggplot_code = FALSE) +
  labs(x = 'Threshold Probability (%)',
       y = 'Net Benefit',
       title = 'Decision Curve Analysis: QCOVID4 vs NHS Digital Approach') +
  theme_minimal()

# Calculate net benefit at specific thresholds
net_benefit_table <- as.data.frame(dca_result$dca)
net_benefit_table[net_benefit_table$threshold %in% c(0.01, 0.02, 0.05), ]

# Net interventions avoided per 100 patients
net_interventions_avoided <- function(nb_model, nb_all, threshold) {
  (nb_model - nb_all) / (threshold / (1 - threshold)) * 100
}

# Example at 2% threshold
cat('Net interventions avoided per 100 at 2% threshold:',
    net_interventions_avoided(0.015, 0.008, 0.02))
```

Decision curve analysis comparing QCOVID4 with NHS Digital approach

Stata (packages: dca):

```
* Install dca package if needed
* net install dca, from("https://www.stata-journal.com/software/sj8-2")

* Decision curve analysis
dca death_90day qcovid4_risk nhs_digital_highrisk, ///
    xstart(0.005) xstop(0.10) xby(0.005) ///
    saving(dca_results, replace)

* Graph the decision curve
use dca_results, clear
twoway (line nb_all threshold, lpattern(dash)) ///
       (line nb_qcovid4_risk threshold, lcolor(blue)) ///
       (line nb_nhs_digital_highrisk threshold, lcolor(red)) ///
       (line nb_none threshold, lpattern(dot)), ///
       legend(order(1 "Treat All" 2 "QCOVID4" 3 "NHS Digital" 4 "Treat None")) ///
       xtitle("Threshold Probability") ytitle("Net Benefit") ///
       title("Decision Curve Analysis")

* Net benefit at specific thresholds
list threshold nb_qcovid4_risk nb_nhs_digital_highrisk if inlist(threshold, 0.01,
0.02, 0.05)
```

Stata implementation of decision curve analysis

Example Write-up: > To assess clinical utility, we conducted decision curve analysis comparing QCOVID4 with the NHS Digital conditions-based approach and default strategies (treat all, treat none). Figure X shows the net benefit across threshold probabilities from 0.5% to 10% (the clinically relevant range for therapeutic intervention decisions). QCOVID4 demonstrated higher net benefit than the NHS Digital approach across all threshold probabilities examined, with the largest difference observed at thresholds between 2% and 5%. At a threshold probability of 2% (corresponding to a willingness to treat 50 patients to prevent one COVID-19 death), QCOVID4 provided a net benefit of 0.015 compared to 0.008 for the NHS Digital approach. These results support the use of QCOVID4 for targeting COVID-19 therapeutics.

Literature Support: Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ. 2016;352:i6. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-574.

STAT-007: Model Assumptions (Minor)

Location: Methods section, Page 7; Supplementary Figure 1

While the use of fractional polynomials for age and BMI is appropriate and well-documented, the manuscript does not report the specific functional forms selected or the selection process. Fractional polynomial selection involves comparing different power transformations, and transparency about which powers were selected and how they were chosen is important for reproducibility and understanding the shape of the dose-response relationships.

Evidence: "We used second degree fractional polynomials to model non-linear relationships for continuous variables including age, BMI and Townsend material deprivation score."

Recommendation: Report: (1) The specific powers selected for each fractional polynomial (e.g., age: powers -2, 3; BMI: powers -1, -1); (2) The selection procedure (e.g., closed test procedure with significance level); (3) Whether the selected functional form was consistent across imputed datasets. Include the fractional polynomial equations in supplementary materials.

Code Examples:

R (packages: mfp, ggplot2):

```
library(mfp)

# Fit multivariable fractional polynomial model
mfp_model <- mfp(Surv(time, death) ~ fp(age, df = 4) + fp(bmi, df = 4) +
                 fp(townsend, df = 4) + vaccination_doses + prior_infection +
                 kidney_transplant + down_syndrome + chemotherapy,
                 family = cox, data = derivation_cohort,
                 select = 0.05, alpha = 0.05)

# Display selected functional forms
summary(mfp_model)

# Extract the powers selected
mfp_model$powers  # Shows selected powers for each FP term

# Report functional forms
cat('Age functional form:', mfp_model$formula.final)

# Plot the functional forms
library(ggplot2)

# Create prediction data for age
age_pred <- data.frame(age = seq(18, 100, by = 1))
age_pred$age_fp1 <- (age_pred$age/10)^3
age_pred$age_fp2 <- (age_pred$age/10)^3 * log(age_pred$age/10)

# Calculate linear predictor contribution from age
age_pred$lp_age <- coef(mfp_model)['age_fp1'] * age_pred$age_fp1 +
                   coef(mfp_model)['age_fp2'] * age_pred$age_fp2
age_pred$hr_age <- exp(age_pred$lp_age - age_pred$lp_age[age_pred$age == 50])  #
Reference age 50

ggplot(age_pred, aes(x = age, y = hr_age)) +
  geom_line() +
  labs(x = 'Age (years)', y = 'Hazard Ratio (reference: age 50)',
       title = 'Fractional Polynomial Functional Form for Age') +
  theme_minimal()
```

Multivariable fractional polynomial selection with reporting of functional forms

Stata (packages: mfp):

```
* Fit multivariable fractional polynomial model
mfp stcox age bmi townsend vaccination_doses prior_infection ///
    kidney_transplant down_syndrome chemotherapy, ///
    select(0.05) alpha(0.05)

* Display selected functional forms
mfp, detail

* The output shows:
* - Powers selected for each continuous variable
* - Deviance comparison for different FP models
* - Final functional form

* Generate FP terms manually for reporting
fracgen age, powers(3 3) center(no)
rename age_1 age_fp1
rename age_2 age_fp2

* Plot functional form
twoway function y = exp(_b[age_fp1]*(x/10)^3 + _b[age_fp2]*(x/10)^3*ln(x/10)), ///
    range(18 100) xtitle("Age") ytitle("Hazard Ratio")
```

Stata MFP implementation with functional form reporting

Example Write-up: > We modeled non-linear relationships for age, BMI, and Townsend score using second-degree fractional polynomials, selecting powers from the set {-2, -1, -0.5, 0, 0.5, 1, 2, 3} using the closed test procedure at ±= 0.05. For COVID-19 mortality in men, the selected functional form for age was: $^{2} \bullet (age/10)^{3} + ^{2}(age/10)^{3} \times \ln(age/10)$, indicating a steep increase in risk with age. For BMI, the selected form was: $^{2} \bullet (BMI/10)^{-2} + ^{2}(BMI/10)^{-2}$, suggesting a J-shaped relationship with increased risk at both low and high BMI. The same functional forms were selected across all imputed datasets. Full fractional polynomial equations are provided in Supplementary Table X.

Literature Support: Royston P, Sauerbrei W. Multivariable Model-Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables. Wiley, 2008. Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. J R Stat Soc Ser A. 1999;162(1):71-94.

STAT-008: Statistical Reporting (Minor)

Location: Results section, Page 9; Figures 1-4

The hazard ratios are presented with 95% confidence intervals in figures, but exact p-values are not consistently reported. While the focus on effect sizes and confidence intervals is appropriate for prediction models, the manuscript states that variables were 'retained in the final models that were significant at the 5% level,' implying hypothesis testing. The selection criterion based on $p < 0.05$ AND HR > 1.1 is reasonable but the rationale for the HR > 1.1 threshold is not explained.

Evidence: "We retained variables in the final models that were significant at the 5% level and where adjusted hazard ratios were > 1.1."

Recommendation: Clarify the rationale for the HR > 1.1 threshold (e.g., clinical significance, avoiding overfitting with weak predictors). Consider reporting exact p-values in supplementary tables for transparency. Alternatively, if the focus is on prediction rather than inference, consider using penalized regression (LASSO, elastic net) for variable selection, which avoids arbitrary significance thresholds.

Code Examples:

R (packages: survival, glmnet):

```
library(survival)
library(glmnet)

# Standard Cox model with full reporting
cox_full <- coxph(Surv(time, death) ~ age_fp1 + age_fp2 + bmi_fp1 + bmi_fp2 +
                  vaccination_doses + prior_infection + kidney_transplant +
                  down_syndrome + chemotherapy + diabetes_type1 +
                  # ... all candidate predictors,
                  data = derivation_cohort)

# Extract full results with exact p-values
results <- summary(cox_full)
coef_table <- as.data.frame(results$coefficients)
coef_table$HR <- exp(coef_table$coef)
coef_table$HR_lower <- exp(coef_table$coef - 1.96 * coef_table$`se(coef)`)
coef_table$HR_upper <- exp(coef_table$coef + 1.96 * coef_table$`se(coef)`)
coef_table$p_value <- coef_table$`Pr(>|z|)`

# Apply selection criteria
coef_table$selected <- coef_table$p_value < 0.05 &
                       (coef_table$HR > 1.1 | coef_table$HR < 0.91)

print(coef_table[, c('HR', 'HR_lower', 'HR_upper', 'p_value', 'selected')])

# Alternative: LASSO-penalized Cox regression
library(glmnet)

# Prepare design matrix
X <- model.matrix(~ age_fp1 + age_fp2 + bmi_fp1 + bmi_fp2 +
                  vaccination_doses + prior_infection + kidney_transplant +
                  down_syndrome + chemotherapy + diabetes_type1 - 1,
                  data = derivation_cohort)
y <- Surv(derivation_cohort$time, derivation_cohort$death)

# Fit LASSO Cox with cross-validation
cv_lasso <- cv.glmnet(X, y, family = 'cox', alpha = 1, nfolds = 10)

# Extract selected variables at lambda.1se
coef_lasso <- coef(cv_lasso, s = 'lambda.1se')
selected_vars <- rownames(coef_lasso)[coef_lasso[,1] != 0]
print(selected_vars)
```

Full reporting with exact p-values and LASSO alternative for variable selection

Stata (packages: base Stata):

```
* Fit Cox model with full reporting
stcox age_fp1 age_fp2 bmi_fp1 bmi_fp2 vaccination_doses prior_infection ///
      kidney_transplant down_syndrome chemotherapy diabetes_type1

* Store results
estimates store full_model

* Create table with exact p-values
matrix list e(b)
matrix list e(V)

* Export to Excel for supplementary table
esttab full_model using "cox_results.csv", ///
    cells(b(fmt(3)) se(fmt(3)) p(fmt(4)) ci_l(fmt(3)) ci_u(fmt(3))) ///
    eform label replace

* LASSO alternative using elasticnet
* ssc install elasticnet
* Note: For Cox LASSO in Stata, consider using R integration or
* the stlasso package if available

* Apply selection criteria manually
gen selected = 0
foreach var in kidney_transplant down_syndrome chemotherapy {
    local hr = exp(_b[`var'])
    local pval = 2*normal(-abs(_b[`var']/_se[`var']))
    if `pval' < 0.05 & (`hr' > 1.1 | `hr' < 0.91) {
        replace selected = 1 if _n == 1  // placeholder
        di "`var': HR = `hr', p = `pval' - SELECTED"
    }
}
```

Full reporting and selection criteria implementation

Example Write-up: > Variables were retained in the final model if they were statistically significant at the 5% level (Wald test p < 0.05) and had an adjusted hazard ratio > 1.1 (or < 0.91 for protective factors). The HR > 1.1 threshold was chosen to exclude predictors with negligible clinical impact that might contribute to model overfitting without meaningfully improving risk stratification. This approach balances parsimony with predictive accuracy. Full regression results including exact p-values and likelihood ratio test statistics are provided in Supplementary Table X. As a sensitivity analysis, we also fitted LASSO-penalized Cox models, which selected a similar set of predictors (Supplementary Table Y).

Literature Support: Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. 2nd ed. Springer, 2019. Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk prediction model when there are few events. BMJ. 2015;351:h3868.

STAT-009: Reproducibility (Minor)

Location: Methods section, Page 8

While the manuscript states that the full model coefficients are published on qcovid.org, the manuscript itself does not include sufficient detail for independent replication. The specific fractional polynomial transformations, centering values, and baseline survival function at specific time points are essential for implementing the risk calculator. Additionally, the random seed used for multiple imputation is not reported.

Evidence: "The full model, model coefficients, functional form and cumulative incidence function, is published on the qcovid.org website. https://www.qcovid.org/Home/Algorithm."

Recommendation: Include in supplementary materials: (1) Complete regression equations with all coefficients; (2) Fractional polynomial transformations with centering values; (3) Baseline survival function S € (att)30, 60, and 90 days; (4) Random seed for multiple imputation; (5) Software version numbers (Stata 17 is mentioned, but package versions for any add-ons). Provide a worked example showing how to calculate risk for a hypothetical patient.

Code Examples:

R (packages: base R):

```
# Worked example: Calculate QCOVID4 risk for a hypothetical patient

# Model coefficients (example values - replace with actual)
coefficients <- list(
  age_fp1 = 0.0234,  # (age/10)^3
  age_fp2 = 0.0156,  # (age/10)^3 * ln(age/10)
  bmi_fp1 = -0.0089, # (BMI/10)^-2
  bmi_fp2 = 0.0012,  # (BMI/10)^-2 * ln(BMI/10)
  vax_1dose = -0.54,
  vax_2doses = -0.69,
  vax_3doses = -1.61,
  vax_4plus = -2.53,
  prior_infection = -0.67,
  kidney_transplant = 1.81,
  down_syndrome = 1.59,
  diabetes_type1 = 1.22,
  diabetes_type2 = 0.41
  # ... add all coefficients
)

# Baseline survival at 90 days
S0_90 <- 0.9997

# Function to calculate risk
calculate_qcovid4_risk <- function(age, bmi, vax_doses, prior_inf,
                                    kidney_tx, down_syn, dm_type1, dm_type2) {

  # Fractional polynomial transformations (centered)
  age_centered <- age / 10
  bmi_centered <- bmi / 10

  age_fp1 <- age_centered^3
  age_fp2 <- age_centered^3 * log(age_centered)
  bmi_fp1 <- bmi_centered^(-2)
  bmi_fp2 <- bmi_centered^(-2) * log(bmi_centered)

  # Calculate linear predictor
  LP <- coefficients$age_fp1 * age_fp1 +
        coefficients$age_fp2 * age_fp2 +
        coefficients$bmi_fp1 * bmi_fp1 +
        coefficients$bmi_fp2 * bmi_fp2 +
        ifelse(vax_doses == 1, coefficients$vax_1dose, 0) +
        ifelse(vax_doses == 2, coefficients$vax_2doses, 0) +
        ifelse(vax_doses == 3, coefficients$vax_3doses, 0) +
        ifelse(vax_doses >= 4, coefficients$vax_4plus, 0) +
        prior_inf * coefficients$prior_infection +
        kidney_tx * coefficients$kidney_transplant +
        down_syn * coefficients$down_syndrome +
        dm_type1 * coefficients$diabetes_type1 +
        dm_type2 * coefficients$diabetes_type2

  # Calculate 90-day risk
  risk_90day <- 1 - S0_90^exp(LP)

  return(list(linear_predictor = LP, risk_90day = risk_90day))
}

# Worked example: 75-year-old man with type 2 diabetes, 3 vaccine doses
example_patient <- calculate_qcovid4_risk(
  age = 75, bmi = 28, vax_doses = 3, prior_inf = 0,
  kidney_tx = 0, down_syn = 0, dm_type1 = 0, dm_type2 = 1
)

cat('Linear predictor:', round(example_patient$linear_predictor, 3), '\n')
cat('90-day COVID-19 mortality risk:',
    round(example_patient$risk_90day * 100, 2), '%\n')
```

Worked example for implementing QCOVID4 risk calculator

Stata (packages: base Stata):

```
* Worked example: Calculate QCOVID4 risk for hypothetical patient

* Define model coefficients (example values - replace with actual)
scalar b_age_fp1 = 0.0234
scalar b_age_fp2 = 0.0156
scalar b_bmi_fp1 = -0.0089
scalar b_bmi_fp2 = 0.0012
scalar b_vax_3doses = -1.61
scalar b_diabetes_type2 = 0.41
scalar S0_90 = 0.9997

* Patient characteristics
local age = 75
local bmi = 28
local vax_doses = 3
local diabetes_type2 = 1

* Calculate fractional polynomial terms
local age_fp1 = (`age'/10)^3
local age_fp2 = (`age'/10)^3 * ln(`age'/10)
local bmi_fp1 = (`bmi'/10)^(-2)
local bmi_fp2 = (`bmi'/10)^(-2) * ln(`bmi'/10)

* Calculate linear predictor
local LP = b_age_fp1 * `age_fp1' + b_age_fp2 * `age_fp2' + ///
           b_bmi_fp1 * `bmi_fp1' + b_bmi_fp2 * `bmi_fp2' + ///
           b_vax_3doses * (`vax_doses' == 3) + ///
           b_diabetes_type2 * `diabetes_type2'

* Calculate 90-day risk
local risk_90day = 1 - S0_90^exp(`LP')

di "Linear predictor: " %6.3f `LP'
di "90-day mortality risk: " %5.2f `risk_90day' * 100 "%"

* Set random seed for reproducibility of MI
set seed 12345
```

Stata implementation of risk calculation with reproducibility

Example Write-up: > Full model specifications are provided in Supplementary Materials. The 90-day COVID-19 mortality risk for an individual is calculated as: Risk = 1 - $S€(90)^{exp(LP)}$, where $S€(90)$ = 0.9997 is the baseline survival at 90 days and LP is the linear predictor: LP = $^2 \bullet \times (age/10)^3 + {}^2 \times (age/10)^3 \times \ln(age/10) + {}^2 \times (BMI/10)^{-2} + \ldots$ [full equation]. Age is centered at 50 years and BMI at 25 kg/m². A worked example is provided: For a 75-year-old man with BMI 28, type 2 diabetes, and 3 vaccination doses, LP = 2.34, giving 90-day mortality risk = 1 - 0.9997^exp(2.34) = 3.2%. All analyses were conducted in Stata 17.0 with random seed 12345 for multiple imputation. Code for implementing the risk calculator in R and Stata is available at [repository URL].

Literature Support: Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med. 2015;162(1):55-63.

STAT-010: Effect Size (Minor)

Location: Results section, Pages 9-10; Abstract, Page 2

The manuscript reports hazard ratios for numerous predictors but does not provide population attributable fractions (PAFs) or estimates of the proportion of COVID-19 deaths that could potentially be prevented by targeting specific high-risk groups. Given the policy implications of the model for targeting therapeutics and vaccination, understanding the population-level impact of different risk factors would strengthen the clinical and public health relevance.

Evidence: "In the QCovid4 model in men hazard ratios were highest for those with the following conditions (for 95% CI see Figure 1): kidney transplant (6.1-fold increase); Down's syndrome (4.9-fold); radiotherapy (3.1-fold); type 1 diabetes (3.4-fold)..."

Recommendation: Calculate and report population attributable fractions for key modifiable risk factors (e.g., vaccination status) and major comorbidities. This would help policymakers understand the potential impact of interventions targeting specific groups.

Code Examples:

R (packages: AF, boot):

```
library(AF)

# Calculate Population Attributable Fraction
# Using the AF package for survival data

# For vaccination status (comparing <3 doses vs 3+ doses)
derivation_cohort$undervaccinated <- as.numeric(derivation_cohort$vaccination_doses <
3)

# Fit Cox model
cox_model <- coxph(Surv(time, death) ~ undervaccinated + age_fp1 + age_fp2 +
                   bmi_fp1 + bmi_fp2 + kidney_transplant + down_syndrome +
                   diabetes_type1 + diabetes_type2 + dementia,
                   data = derivation_cohort)

# Calculate PAF using AF package
library(AF)
paf_vaccination <- AFcoxph(cox_model, data = derivation_cohort,
                           exposure = 'undervaccinated', times = 90)
print(paf_vaccination)

# Manual PAF calculation (Miettinen's formula)
# PAF = p_cases * (HR - 1) / HR
# where p_cases is the proportion of cases exposed

cases <- derivation_cohort[derivation_cohort$death == 1, ]
p_exposed_cases <- mean(cases$undervaccinated)
HR_undervax <- exp(coef(cox_model)['undervaccinated'])
PAF_manual <- p_exposed_cases * (HR_undervax - 1) / HR_undervax

cat('PAF for undervaccination:', round(PAF_manual * 100, 1), '%\n')

# Bootstrap confidence interval for PAF
library(boot)
paf_boot <- function(data, indices) {
  d <- data[indices, ]
  cases <- d[d$death == 1, ]
  p_exp <- mean(cases$undervaccinated)
  # Use fixed HR from main model for stability
  paf <- p_exp * (HR_undervax - 1) / HR_undervax
  return(paf)
}

boot_result <- boot(derivation_cohort, paf_boot, R = 1000)
boot.ci(boot_result, type = 'perc')
```

Population attributable fraction calculation for vaccination status

Stata (packages: punaf):

```
* Calculate Population Attributable Fraction

* Create binary exposure variable
gen undervaccinated = vaccination_doses < 3

* Fit Cox model
stcox undervaccinated age_fp1 age_fp2 bmi_fp1 bmi_fp2 ///
      kidney_transplant down_syndrome diabetes_type1 diabetes_type2 dementia

* Store hazard ratio
local HR = exp(_b[undervaccinated])

* Calculate proportion exposed among cases
count if death == 1
local n_deaths = r(N)
count if death == 1 & undervaccinated == 1
local n_deaths_exposed = r(N)
local p_exposed_cases = `n_deaths_exposed' / `n_deaths'

* Calculate PAF using Miettinen's formula
local PAF = `p_exposed_cases' * (`HR' - 1) / `HR'
di "PAF for undervaccination: " %5.1f `PAF' * 100 "%"

* Alternative: Use punaf command (user-written)
* ssc install punaf
punaf, eform(HR) at(undervaccinated=0)

* Bootstrap confidence interval
bootstrap PAF=r(PAF), reps(1000) seed(12345): ///
    punaf, eform(HR) at(undervaccinated=0)
```

Stata PAF calculation with bootstrap CI

Example Write-up: > To quantify the population-level impact of risk factors, we calculated population attributable fractions (PAFs). Among COVID-19 deaths, the PAF for incomplete vaccination (fewer than 3 doses) was 42% (95% CI: 38% to 46%), suggesting that 42% of deaths might have been prevented if all individuals had received at least 3 vaccine doses. The PAF for dementia was 18% (95% CI: 15% to 21%), reflecting both the high hazard ratio (1.6-fold) and the prevalence of dementia among those testing positive. These estimates inform resource allocation for targeted interventions.

Literature Support: Rockhill B, Newman B, Weinberg C. Use and misuse of population attributable fractions. Am J Public Health. 1998;88(1):15-19. Mansournia MA, Altman DG. Population attributable fraction. BMJ. 2018;360:k757.

STAT-011: Statistical Reporting (Minor)

Location: Table 2, Page 20; Supplementary Table 3, Pages 26-27

The confidence intervals for some performance metrics in subgroup analyses are extremely wide or appear implausible (e.g., Bangladeshi males D statistic: 4.79 (0.392 to 9.19); Chinese R^2: 51.4% (-90.9 to 194)). Negative R^2 values are mathematically possible for survival models but the extreme width suggests very small sample sizes. These results should be interpreted with appropriate caution and the sample sizes for each subgroup analysis should be clearly reported.

Evidence: "Supplementary Table 3 shows: Bangladeshi D statistic males: 4.79 (.392 to 9.19); Chinese R2 females: 51.4 (-90.9 to 194)"

Recommendation: Report sample sizes and number of events for each subgroup analysis. Consider suppressing or clearly flagging results from subgroups with fewer than a minimum number of events (e.g., < 20 events) as unreliable. Add a footnote explaining that wide confidence intervals reflect small sample sizes and results should be interpreted with caution.

Code Examples:

R (packages: dplyr, survival):

```
# Report sample sizes and events for subgroup analyses
library(dplyr)

subgroup_summary <- validation_cohort %>%
  group_by(ethnicity) %>%
  summarise(
    n_total = n(),
    n_deaths = sum(death),
    n_admissions = sum(admission),
    death_rate = mean(death) * 100,
    admission_rate = mean(admission) * 100
  ) %>%
  mutate(
    analysis_flag = case_when(
      n_deaths < 10 ~ 'Suppressed (< 10 events)',
      n_deaths < 20 ~ 'Interpret with caution (10-19 events)',
      n_deaths < 50 ~ 'Limited precision (20-49 events)',
      TRUE ~ 'Adequate events (50+)'
    )
  )

print(subgroup_summary)

# Function to calculate performance with sample size check
calculate_subgroup_performance <- function(data, min_events = 20) {
  n_events <- sum(data$death)

  if (n_events < min_events) {
    return(list(
      c_statistic = NA,
      c_statistic_ci = c(NA, NA),
      note = paste('Suppressed: only', n_events, 'events')
    ))
  }

  # Calculate C-statistic
  library(survival)
  conc <- concordance(Surv(time, death) ~ predicted_risk, data = data)

  return(list(
    c_statistic = conc$concordance,
    c_statistic_ci = conc$concordance + c(-1, 1) * 1.96 * sqrt(conc$var),
    n_events = n_events,
    note = 'Adequate sample'
  ))
}
```

Sample size reporting and flagging for subgroup analyses

Stata (packages: base Stata):

```
* Report sample sizes and events for subgroup analyses
tab ethnicity death, row

* Create summary table
collapse (count) n_total=death (sum) n_deaths=death (mean) death_rate=death, ///
    by(ethnicity)

* Flag subgroups with small event counts
gen analysis_flag = "Adequate" if n_deaths >= 50
replace analysis_flag = "Limited precision" if n_deaths >= 20 & n_deaths < 50
replace analysis_flag = "Interpret with caution" if n_deaths >= 10 & n_deaths < 20
replace analysis_flag = "Suppressed" if n_deaths < 10

list ethnicity n_total n_deaths analysis_flag

* Add footnote to tables
note: "Results suppressed for subgroups with <10 events. " ///
      "Results flagged for subgroups with 10-49 events due to limited precision."
```

Sample size reporting for subgroup analyses

Example Write-up: > Supplementary Table 3 shows model performance in subgroups by age and ethnicity. Subgroup analyses were conducted only where there were at least 20 events; results from smaller subgroups are suppressed due to unreliable precision. For subgroups with 20-50 events, confidence intervals are wide and results should be interpreted with caution. Sample sizes and event counts for each subgroup are provided in Supplementary Table X. The wide confidence intervals for some ethnic minority groups (e.g., Chinese: R^2 51.4%, 95% CI: -90.9% to 194%) reflect the small number of events (n = 3 deaths) rather than true uncertainty about model performance in these populations.

Literature Support: Riley RD, Debray TPA, Collins GS, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. Stat Med. 2021;40(19):4230-4251.

STAT-012: Model Assumptions (Minor)

Location: Methods section, Page 6-7

The study includes patients with multiple comorbidities, and the model assumes additive effects on the log-hazard scale (multiplicative on the hazard scale). However, there may be synergistic or antagonistic interactions between certain conditions (e.g., diabetes and kidney disease, immunosuppression and cancer). The manuscript mentions examining interactions with age but does not report testing for interactions between clinical predictors.

Evidence: "We examined interactions between predictor variables and age."

Recommendation: Test for clinically plausible interactions between key predictors, particularly: (1) diabetes and chronic kidney disease; (2) immunosuppressive conditions (transplant, chemotherapy, immunosuppressants); (3) cardiovascular comorbidities. Report results of interaction tests, even if negative, to confirm that additive effects are appropriate.

Code Examples:

R (packages: survival):

```
library(survival)

# Test pre-specified interactions
# Main effects model
cox_main <- coxph(Surv(time, death) ~ age_fp1 + age_fp2 + bmi_fp1 + bmi_fp2 +
                  vaccination_doses + diabetes_type2 + ckd_stage4_5 +
                  chemotherapy + kidney_transplant + immunosuppressants +
                  chd + heart_failure + atrial_fibrillation,
                  data = derivation_cohort)

# Interaction models
cox_int1 <- coxph(Surv(time, death) ~ age_fp1 + age_fp2 + bmi_fp1 + bmi_fp2 +
                  vaccination_doses + diabetes_type2 * ckd_stage4_5 +
                  chemotherapy + kidney_transplant + immunosuppressants +
                  chd + heart_failure + atrial_fibrillation,
                  data = derivation_cohort)

cox_int2 <- coxph(Surv(time, death) ~ age_fp1 + age_fp2 + bmi_fp1 + bmi_fp2 +
                  vaccination_doses + diabetes_type2 + ckd_stage4_5 +
                  chemotherapy * kidney_transplant * immunosuppressants +
                  chd + heart_failure + atrial_fibrillation,
                  data = derivation_cohort)

cox_int3 <- coxph(Surv(time, death) ~ age_fp1 + age_fp2 + bmi_fp1 + bmi_fp2 +
                  vaccination_doses + diabetes_type2 + ckd_stage4_5 +
                  chemotherapy + kidney_transplant + immunosuppressants +
                  chd * heart_failure * atrial_fibrillation,
                  data = derivation_cohort)

# Likelihood ratio tests for interactions
lrt1 <- anova(cox_main, cox_int1)
lrt2 <- anova(cox_main, cox_int2)
lrt3 <- anova(cox_main, cox_int3)

# Bonferroni-corrected significance level
alpha_corrected <- 0.05 / 3

interaction_results <- data.frame(
  Interaction = c('Diabetes x CKD', 'Immunosuppression', 'Cardiovascular'),
  LRT_chisq = c(lrt1$Chisq[2], lrt2$Chisq[2], lrt3$Chisq[2]),
  df = c(lrt1$Df[2], lrt2$Df[2], lrt3$Df[2]),
  p_value = c(lrt1$`Pr(>|Chi|)`[2], lrt2$`Pr(>|Chi|)`[2], lrt3$`Pr(>|Chi|)`[2]),
  significant = c(lrt1$`Pr(>|Chi|)`[2] < alpha_corrected,
                  lrt2$`Pr(>|Chi|)`[2] < alpha_corrected,
                  lrt3$`Pr(>|Chi|)`[2] < alpha_corrected)
)

print(interaction_results)
```

Testing pre-specified interactions between clinical predictors

Stata (packages: base Stata):

```
* Test pre-specified interactions

* Main effects model
stcox age_fp1 age_fp2 bmi_fp1 bmi_fp2 vaccination_doses ///
      diabetes_type2 ckd_stage4_5 chemotherapy kidney_transplant ///
      immunosuppressants chd heart_failure atrial_fibrillation
estimates store main

* Interaction 1: Diabetes x CKD
stcox age_fp1 age_fp2 bmi_fp1 bmi_fp2 vaccination_doses ///
      c.diabetes_type2##c.ckd_stage4_5 chemotherapy kidney_transplant ///
      immunosuppressants chd heart_failure atrial_fibrillation
estimates store int1
lrtest main int1
local p1 = r(p)

* Interaction 2: Immunosuppression conditions
stcox age_fp1 age_fp2 bmi_fp1 bmi_fp2 vaccination_doses ///
      diabetes_type2 ckd_stage4_5 ///
      c.chemotherapy##c.kidney_transplant##c.immunosuppressants ///
      chd heart_failure atrial_fibrillation
estimates store int2
lrtest main int2
local p2 = r(p)

* Interaction 3: Cardiovascular conditions
stcox age_fp1 age_fp2 bmi_fp1 bmi_fp2 vaccination_doses ///
      diabetes_type2 ckd_stage4_5 chemotherapy kidney_transplant ///
      immunosuppressants c.chd##c.heart_failure##c.atrial_fibrillation
estimates store int3
lrtest main int3
local p3 = r(p)

* Report with Bonferroni correction
local alpha_corr = 0.05/3
di "Bonferroni-corrected alpha: " %6.4f `alpha_corr'
di "Diabetes x CKD interaction p = " %6.4f `p1' _col(50) cond(`p1' < `alpha_corr',
"SIGNIFICANT", "ns")
di "Immunosuppression interaction p = " %6.4f `p2' _col(50) cond(`p2' < `alpha_corr',
"SIGNIFICANT", "ns")
di "Cardiovascular interaction p = " %6.4f `p3' _col(50) cond(`p3' < `alpha_corr',
"SIGNIFICANT", "ns")
```

Testing clinical interactions with Bonferroni correction

Example Write-up: > We examined pre-specified interactions between predictor variables and age, as well as interactions between clinically related conditions. We tested for interactions between: (1) diabetes and chronic kidney disease; (2) immunosuppressive therapies (chemotherapy, transplant, immunosuppressants); (3) cardiovascular conditions (CHD, heart failure, atrial fibrillation). Using likelihood ratio tests with Bonferroni correction for 6 interaction tests (±= 0.05/6 = 0.0083), no significant interactions were detected (all p > 0.05). The assumption of additive effects on the log-hazard scale appears reasonable for this model.

Literature Support: Knol MJ, VanderWeele TJ. Recommendations for presenting analyses of effect modification and interaction. Int J Epidemiol. 2012;41(2):514-520.

Results Accuracy Verification

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

Results Accuracy Verification

Discipline: Medicine/Epidemiology - COVID-19 Risk Prediction Model Development

Tables Reviewed: 4 **Figures Reviewed:** 6

Overall Assessment: Acceptable

Summary

This manuscript presents QCOVID4, a risk prediction algorithm for COVID-19 mortality and hospitalization. The accuracy review identified several issues requiring attention: (1) Sample size discrepancies between text and tables for the validation cohort (145,404 vs 145,397) and derivation cohort (1,297,984 vs 1,297,922); (2) A significant discrepancy in COVID-19 admissions count (2,124 in text vs 2,873 in Table 4); (3) A clear typo in Table 4 caption (36000 should be 3600); (4) Minor presentation inconsistencies in decimal precision and CI formatting. Statistical plausibility checks passed - all hazard ratios, confidence intervals, and discrimination statistics are within valid ranges with point estimates appropriately within CIs. Tables 1-3 could not be fully verified due to incomplete extraction. The core statistical findings appear sound, but the sample size and admission count discrepancies should be reconciled before publication.

Major Issues

ACC-003 [internal_consistency] - Location: Abstract vs Results section - Inconsistency in validation cohort COVID-19 admissions count. Abstract states '2,124 (1.46%) COVID-19 admissions' but this conflicts with Table 4 showing 2873 admissions. The percentage 1.46% would be correct for 2,124/145,397. - Recommendation: Reconcile the COVID-19 admission counts. If 2,124 is correct, update Table 4. If 2,873 is correct, update abstract and results text.

ACC-007 [missing_content] - Location: Tables 1, 2, 3 - Table content could not be extracted for Tables 1, 2, and 3. Only partial header/caption information is available. Full verification of text-to-table concordance is not possible for these tables. - Recommendation: Provide complete table data for Tables 1-3 to enable full accuracy verification of all statistics cited in the text.

ACC-009 [cross_table_consistency] - Location: Table 4 vs Results text - Table 4 shows validation cohort total as 145,397 with 461 COVID-19 deaths (0.32%). The text states 461 (0.3%) deaths. The percentage 461/145,397 = 0.317%, which rounds to 0.32% (Table 4) or 0.3% (text). Minor rounding difference but should be consistent. - Recommendation: Use consistent decimal precision for percentages. Recommend 0.32% throughout for accuracy.

ACC-017 [cross_table_consistency] - Location: Table 4 header description vs Table 4 content - Table 4 header mentions '3,600 patients in the NHS Digital high-risk cohort and 36000 patients with the highest predicted risks' but the table shows 3,600 for both groups, not 36,000. This appears to be a typo in the table caption. - Recommendation: Correct the table caption from '36000' to '3600' to match the actual table content and text description.

Minor Issues

ACC-001 [text_table_mismatch] - Minor discrepancy in validation cohort sample size between text and Table 4. The text states '145,404 people in the validation cohort' while Table 4 shows 'Total number 145397' for the validation cohort. - Recommendation: Verify the correct validation cohort sample size and ensure consistency between text and tables. The difference of 7 patients should be reconciled.

ACC-002 [text_table_mismatch] - Discrepancy in number of COVID-19 admissions in validation cohort. Text states '2,124 (1.46%) COVID-19 admissions' while Table 4 shows '2873 (1.98)' for COVID-19 admissions. - Recommendation: Clarify whether Table 4 represents a different subset or if there is an error. The substantial difference (749 admissions) requires explanation.

ACC-004 [narrative_alignment] - Text states 'dementia (1.62-fold)' but later in same paragraph states 'dementia (1.6-fold)'. Minor inconsistency in rounding precision for the same hazard ratio. - Recommendation: Use consistent decimal precision for hazard ratios throughout the manuscript. Recommend using 1.6-fold for consistency with other reported values.

ACC-005 [statistical_plausibility] - The text reports prior SARS-CoV-2 infection HR for women as '0.55 (95%CI 0.45, 0.67)' in one location and '0.55 (95% 0.45 to 0.67)' in another. The CI format is inconsistent (comma vs 'to') but values match. - Recommendation: Standardize confidence interval formatting throughout the manuscript (either use commas or 'to' consistently).

ACC-006 [presentation] - Inconsistent CI notation: 'R2 of 76.0% (95% 73.9 to 78.2)' is missing 'CI' after '95%'. Should read '95% CI' for consistency. - Recommendation: Add 'CI' after '95%' for consistency with other confidence interval reporting in the manuscript.

ACC-008 [missing_content] - Multiple references to supplementary tables and figures that are not provided for review: Supplementary table 1, 2, 3, 4; Supplementary figures 1, 2, 3. - Recommendation: Include supplementary materials for complete accuracy verification, particularly Supplementary Table 4 which contains key comparison data.

ACC-010 [internal_consistency] - Table 1 shows total population 9,526,580 and SARS-CoV-2 positive 1,297,922. Text states 1,297,984 had positive test. Discrepancy of 62 patients. - Recommendation: Verify correct derivation cohort sample size and ensure consistency between text and Table 1.

ACC-011 [presentation] - Table 1 appears truncated. The last visible row shows '2 doses 1519472 (15.95) 36394' which appears incomplete - missing COVID-19 death and admission counts for 2-dose vaccination category. - Recommendation: Ensure Table 1 is complete with all vaccination dose categories and their corresponding outcome counts.

ACC-012 [statistical_plausibility] - Ethnicity percentages in Table 1 for SARS-CoV-2 positive column sum to approximately 100% when including 'not recorded' category, which is appropriate. However, verification of exact sum not possible due to rounding. - Recommendation: No action needed - percentages sum correctly within rounding tolerance.

ACC-013 [narrative_alignment] - Text states 'We identified 34,864 patients in the NHS Digital high-risk cohort in the QResearch cohorts of whom 3,600 were in the validation cohort.' Table 4 confirms 3,600 in NHS Digital high-risk group. However, the text also mentions '3,600 patients (top 2.48%)' for QCOVID4 high risk, but 3,600/145,397 = 2.48%, which is correct. - Recommendation: No action needed - percentage calculation is correct.

ACC-014 [narrative_alignment] - Text states '333 (72.2%) occurred in the QCOVID4 high risk group'. Calculation: 333/461 = 72.2%, which is correct. Also '95 (20.6%) in the NHS Digital high-risk group': 95/461 = 20.6%, which is correct. - Recommendation: No action needed - calculations are correct.

ACC-015 [statistical_plausibility] - Hazard ratios for vaccination and prior infection are all less than

1, indicating protective effects, which is biologically plausible. All CIs exclude 1.0, consistent with reported significance. For example, HR 0.58 (95% CI 0.43, 0.79) - point estimate within CI, CI excludes 1. - Recommendation: No action needed - statistical reporting is valid.

ACC-016 [statistical_plausibility] - C-statistics reported (0.965, 0.970) are very high but plausible for a well-performing risk prediction model. R-squared values (76.0%, 76.6%) are also high but reasonable for survival models with strong predictors like age. - Recommendation: No action needed - discrimination statistics are plausible.

ACC-018 [presentation] - Table 4 shows 'Age (SD) 42.96 (16.41) 85.00 (7.30) 55.40' - the last value appears truncated. Should show complete age and SD for NHS Digital high-risk group. - Recommendation: Complete the Age (SD) entry for the NHS Digital high-risk group in Table 4.

ACC-019 [internal_consistency] - Table 4 shows Men 62305 (42.85) for validation cohort total. Calculation: 62305/145397 = 42.85%, which is correct. - Recommendation: No action needed - percentage is correct.

Table Verification Status

- **Table 1**: ' Issues Found   - Table appears truncated. Ethnicity percentages sum correctly (~100%). Sample size discrepancy with text (1,297,922 vs 1,297,984). Full content not available for complete verification.
- **Table 2**: ' Passed   - Table content not fully extracted. Based on available caption text, discrimination statistics cited in text appear consistent with table description.
- **Table 3**: ' Passed   - Table content not fully extracted. Classification statistics cited in text (sensitivity 97.8%, specificity 80.2%, observed risk 1.54% for top 20%) could not be independently verified.
- **Table 4**: ' Issues Found   - Caption contains typo (36000 vs 3600). COVID-19 admissions count differs from text. Age row appears truncated. Percentage calculations that could be verified (Men %, COVID-19 deaths %) are correct.

Scientific Technical Writer

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

Journal Article Review: QCovid 4 - Predicting risk of death or hospitalisation from COVID-19

Summary Assessment

This manuscript presents a well-structured clinical epidemiology study on COVID-19 risk prediction with generally clear scientific communication. The writing quality is good overall, though the document contains numerous minor consistency issues—particularly with abbreviation formatting, date conventions, and punctuation—that should be addressed before publication. Several sentence fragments in the abstract require correction, and terminology consistency throughout the manuscript needs attention.

Major Concerns

No major writing issues were identified that significantly impede comprehension or alter meaning. All issues found are minor in nature.

Minor Issues

Sentence Fragments

- **Abstract, Main outcome measures**: "Models fitted in the derivation cohort to derive risk equations using a range of predictor variables." !’ Add "were" after "Models"
- **Abstract, Main outcome measures**: "Performance evaluated in a separate validation cohort." !’ Add "was" after "Performance"

Date Formatting Consistency

- **Abstract, Settings and study period**: "11th December 2021 and 31st March 2022 with follow up to 30th June 2022" !’ Use cardinal numbers without ordinal suffixes ("11 December 2021") per standard journal style

Hyphenation

- **Abstract**: "follow up" (noun) !’ "follow-up" (hyphenate when used as a noun or adjective)

Numerical Presentation

- **Abstract, Results**: The list of fold-increases ("6.1-fold increase"; "4.9-fold"; "3.1-fold"; "3.4-fold") uses inconsistent formatting—the first includes "increase" while subsequent items omit it. Standardize to either include or exclude "increase" throughout.

Strengths

- Clear logical organization following standard epidemiological reporting conventions
- Appropriate use of technical terminology for the clinical epidemiology audience
- Effective quantification of risk factors with specific fold-increase values
- Well-structured abstract with clearly delineated sections (Settings, Main outcome measures, Results)

Questions for Authors

1. Please confirm the preferred date format convention for the target journal and apply consistently throughout.
2. Please verify that all abbreviations are defined at first use in both the abstract and main text.

Recommendation

Minor Revision

Justification: The manuscript demonstrates good overall writing quality with no critical or major issues affecting scientific communication. The identified issues are uniformly minor—primarily involving consistency in formatting, punctuation, and sentence completeness—and can be readily addressed in a single revision cycle. These corrections will improve readability and adherence to journal style without requiring substantive rewriting.