#Reviewer 1

1.Comments to the Author:

The Consent form and IRB are both in non-English language. Please translate them and submit them again

Author response and action taken:

We are truly grateful for this insightful feedback, and we sincerely appreciate your expertise in this field. I have translated the current non-English informed consent form into English with the IRB document and resubmitted it.

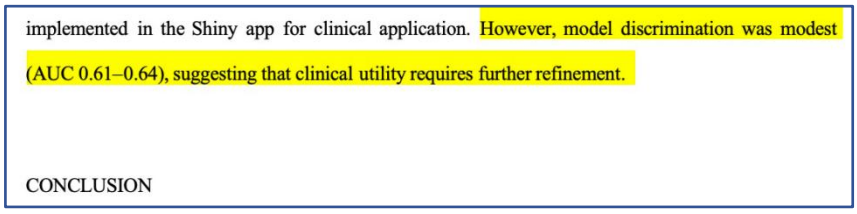
2.Comments to the Author:

XGBoost AUCs (0.640 training, 0.610 testing) are low for clinical prediction. This is mentioned in the conclusion but could be emphasized more openly in both Abstract and Discussion to manage the expectations of the readers.

Author response and action taken:

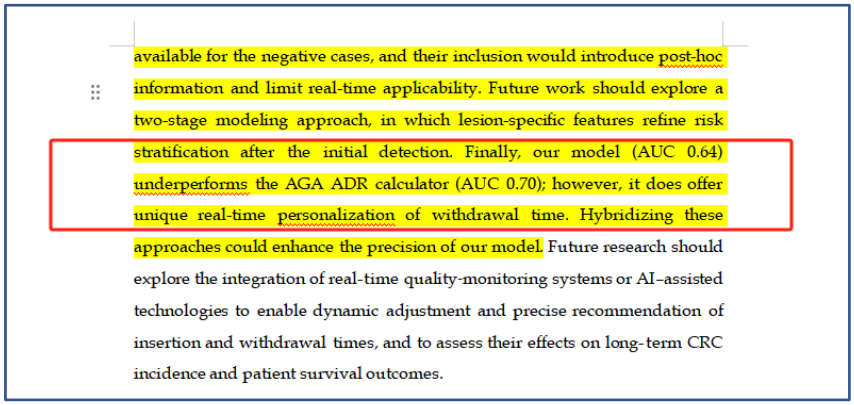
We highly appreciate this suggestion, and thank you once again for your in-depth academic insights. We've made it clearer in the Abstract and Discussion sections as required to have a low AUC for the XGBoost model (0.640 for the training set; 0.610 for the test set) to manage reader expectations.

[Abstract]:


implemented in the Shiny app for clinical application. However, model discrimination was modest (AUC 0.61–0.64), suggesting that clinical utility requires further refinement.

CONCLUSION

[Discussion]:


available for the negative cases, and their inclusion would introduce post-hoc information and limit real-time applicability. Future work should explore a two-stage modeling approach, in which lesion-specific features refine risk stratification after the initial detection. Finally, our model (AUC 0.64) underperforms the AGA ADR calculator (AUC 0.70); however, it does offer unique real-time personalization of withdrawal time. Hybridizing these approaches could enhance the precision of our model. Future research should explore the integration of real-time quality-monitoring systems or AI-assisted technologies to enable dynamic adjustment and precise recommendation of insertion and withdrawal times, and to assess their effects on long-term CRC incidence and patient survival outcomes.

3.Comments to the Author:

Consider making insertion time an explicit feature in the machine learning model if possible or clarify why stratification was preferred.

Author response and action taken:

This feedback is highly valuable, and we genuinely appreciate your thoughtful academic input and support. During model development, insertion time (IT) was first evaluated as a categorical variable for sample stratification—dividing IT into discrete bins —but this produced uneven strata with small group sizes and yielded no AUC improvement. Next, these same IT bins were one-hot encoded as explicit features, yet model discrimination remained unchanged and IT's feature importance stayed low. Finally, we treated IT as a continuous predictor, allowing the algorithm to learn potential nonlinear relationships; this approach delivered the best (albeit modest) AUC gain and avoided the instability observed with sparse categorical groups. Consequently, to maximize statistical power, stability, and clinical interpretability, we retained insertion time as a continuous feature in our final machine-learning models.

I've modified the expression in the summary section.

> AIM
>
> To identify independent predictors of ≥25% ADR, develop a risk model, and propose withdrawal durations based on different insertion times.

4.Comments to the Author:

More information about shinny app will be useful.

Author response and action taken:

Thanks for the reviewer's valuable comments on our work. This is a very insightful suggestion, and we truly appreciate your high level academic insights once again. We've added more information about the Shiny app's features, interface, and usage scenarios.

> ***Shiny App Implementation and Usage***
>
> The Shiny application (https://xbxmodel.shinyapps.io/xbx_prediction/) offers a streamlined, single-page interface for real-time, personalized prediction of colonoscopy withdrawal time: users input cecal insertion time (seconds), patient age, sex (dropdown), and endoscopist experience level ("Low" or "High") via numeric fields and selectors, then click "Predict Withdrawal Time" to display the model-derived minimum WT required to achieve ≥25% ADR. In pre-procedure planning, anticipated insertion times can be used to establish WT targets; during live procedures, the app's countdown timer (not shown) ensures adherence to the calculated WT; and in post-procedure audit and training, exported predicted versus actual WT records enable quality-assurance teams to assess compliance, identify training needs, and refine endoscopy protocols.

5.Comments to the Author:

The role of insertion time is conceptually central to the study but remains underutilized in the final model. Integrating this variable directly into the machine learning pipeline could further refine predictive accuracy. AUCs of 0.640 (training) and 0.610 (testing) suggest limited discriminative power for XGBoost. While acceptable, this should be acknowledged or contextualized.

Author response and action taken:

Your suggestion is very insightful, and we greatly value your academic perspective and constructive feedback. In our research, we ultimately adopted a combination of SVM and XGBoost algorithms to establish the model. Throughout the two steps, we incorporated IT as a continuous variable into the model. Specifically, the first step involved predicting_WT based on IT, endoscopist experience level, patient age, and sex, then synthesizing these variables to forecast corresponding probabilities.

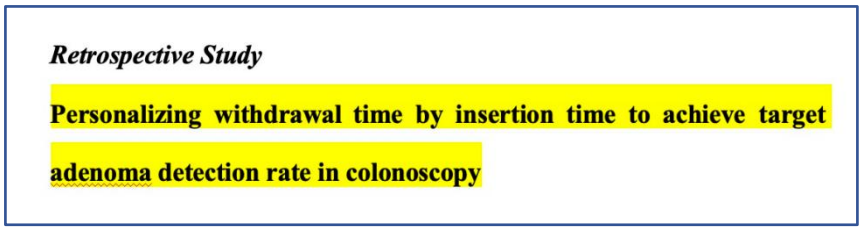
#Reviewer 2

1.Comments to the Author:

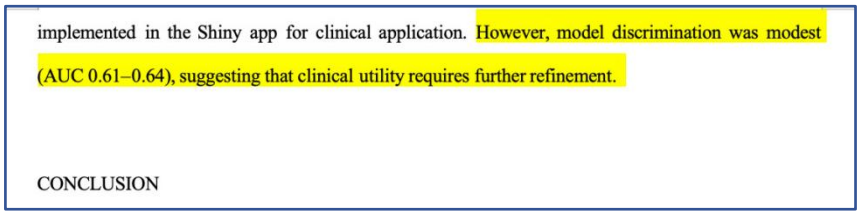
Title & Abstract Issue: "Optimal withdrawal time" is overstated; the model predicts minimum WT for ≥25% ADR, not true optimization. Correction: Title: Change to "Personalizing withdrawal time by insertion time to achieve target adenoma detection rate in colonoscopy." Abstract: Replace "optimal" with "target-achieving" and add: "Model discrimination was modest (AUC 0.61–0.64), suggesting clinical utility requires further refinement."

Author response and action taken:

We are truly grateful for this insightful feedback, and we sincerely appreciate your expertise in this field. Title has been changed to: “Personalizing withdrawal time by insertion time to achieve target adenoma detection rate in colonoscopy.” In abstract, we also replaced “Optimal” with “target-achieving” and added the sentence:“Model discrimination was modest (AUC 0.61–0.64), suggesting that clinical utility requires further refinement.” And I changed all the optimals in the article to target-achieving.


Retrospective Study

Personalizing withdrawal time by insertion time to achieve target adenoma detection rate in colonoscopy


implemented in the Shiny app for clinical application. However, model discrimination was modest (AUC 0.61–0.64), suggesting that clinical utility requires further refinement.

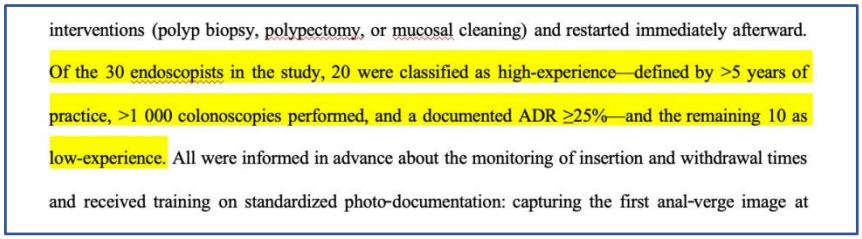
CONCLUSION

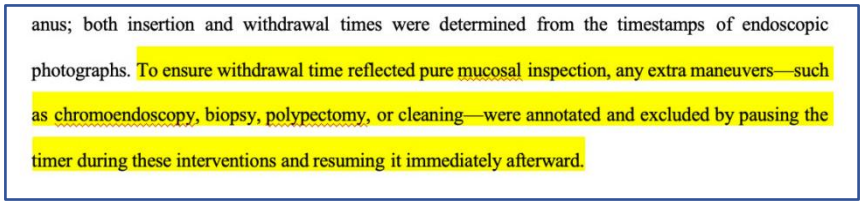
2.Comments to the Author:

Methods Endoscopist Stratification: Issue: Criteria for "high/low experience" undefined. Correction: Add: *"High-experience endoscopists were defined as those with >5 years of practice, >1,000 colonoscopies performed, and documented ADR ≥25%."* Withdrawal Time Measurement: Issue: Ambiguity in "negative WT" calculation. Correction: Specify: "The timer was paused during interventions (biopsy, polypectomy, cleaning) and resumed immediately after, ensuring WT reflected pure inspection time." Prospective Timeline: Issue: Prospective phase dates (Sept 2024–Apr 2025) imply future data. Correction: Revise dates to past tense (e.g., Sept 2023–Apr 2024) or clarify: "Prospective data collection was completed at submission, with analysis ongoing."

Author response and action taken:

We highly appreciate this suggestion, and thank you once again for your in-depth academic insights. We have added the following definition under Endoscopist Stratification: Of the 30 endoscopists in the study, 20 were classified as high-experience—defined by >5 years of practice, >1 000 colonoscopies performed, and a documented ADR ≥25%—and the remaining 10 as low-experience. Under Withdrawal Time Measurement, we clarified: To ensure withdrawal time reflected pure mucosal inspection, any extra maneuvers—such as chromoendoscopy, biopsy, polypectomy, or cleaning—were annotated and excluded by pausing the timer during these interventions and resuming it immediately afterward. This sentence has been added to the Prospective phase dates: Prospective data collection was completed at submission, with analysis ongoing.


interventions (polyp biopsy, polypectomy, or mucosal cleaning) and restarted immediately afterward. Of the 30 endoscopists in the study, 20 were classified as high-experience—defined by >5 years of practice, >1 000 colonoscopies performed, and a documented ADR ≥25%—and the remaining 10 as low-experience. All were informed in advance about the monitoring of insertion and withdrawal times and received training on standardized photo-documentation: capturing the first anal-verge image at


anus; both insertion and withdrawal times were determined from the timestamps of endoscopic photographs. To ensure withdrawal time reflected pure mucosal inspection, any extra maneuvers—such as chromoendoscopy, biopsy, polypectomy, or cleaning—were annotated and excluded by pausing the timer during these interventions and resuming it immediately afterward.

Phase II (Prospective Study): In Phase II, conducted from September 2024 to April 2025, 5,699 consecutive outpatients undergoing colonoscopy were enrolled. Prospective data collection was completed at submission, with analysis ongoing. In the prospective cohort, 6,712 patients met the initial inclusion criteria. Of these, 5 (0.1%) failed cecal intubation, 234 (3.5%) had a history of colorectal cancer with concurrent colorectal resection, 3 (0.04%) were diagnosed with familial polyposis, 89 (1.3%) had inflammatory bowel disease, and 682 (10.2%) had incomplete endoscopic data. After these exclusions, 5,699 patients remained for the final analysis.

3.Comments to the Author:

Results Table 3 (Multivariate Analysis): Issue: Implausible OR for WT ("1.00 [1.00–1.00]"). Correction: Correct to reported univariate OR (e.g., "1.02 [1.01–1.05]") and ensure consistent decimal precision. Figure 2 (Prospective Validation): Issue: Unconventional 7-subset validation lacks justification. Correction: Add: "We split the prospective cohort into 7 subsets (n≈800 each) to mirror the retrospective sample size, ensuring stability testing across heterogeneous groups." Model Performance: Issue: Modest AUC (0.61–0.64) framed as "stable." Correction: State: "The model showed consistent but modest discrimination (AUC 0.61–0.64), indicating potential for decision support pending optimization."

Author response and action taken:

This feedback is highly valuable, and we genuinely appreciate your thoughtful academic input and support. We corrected the OR of WT to "1.02 [1.01–1.05]" in the multivariate analysis. The rationale for the 7-group split validation was supplemented: We split the prospective cohort into 7 subsets (n≈800 each) to mirror the retrospective sample size, ensuring stability testing across heterogeneous groups. We added that the model showed consistent but modest discrimination (AUC 0.61 – 0.64), indicating potential for decision support pending optimization.

Intubation time	1.01(1.00–1.03)	0.621	-	
Withdrawal time	1.02 (1.01–1.05)	**<0.001**	1.02 (1.01–1.05)	**<0.001**

OR, Odds ratio; CI, Confidence interval; BMI, Body mass index; COPD, Chronic obstructive pulmonary disease; BBPS, Boston Bowel Preparation Scale

good model fit (Figure 2D). The decision-curve analysis demonstrated a positive net benefit across threshold probabilities of 10–50%, suggesting clinical utility (Figure 2E). We split the prospective cohort into 7 subsets (n≈800 each) to mirror the retrospective sample size, ensuring stability testing across heterogeneous groups. The overall AUC was 0.700, and the subset AUCs ranged from 0.690 to 0.730 (0.690, 0.730, 0.720, 0.710, 0.690, 0.700, and 0.710 for groups 1–7, respectively; Figure 2F),

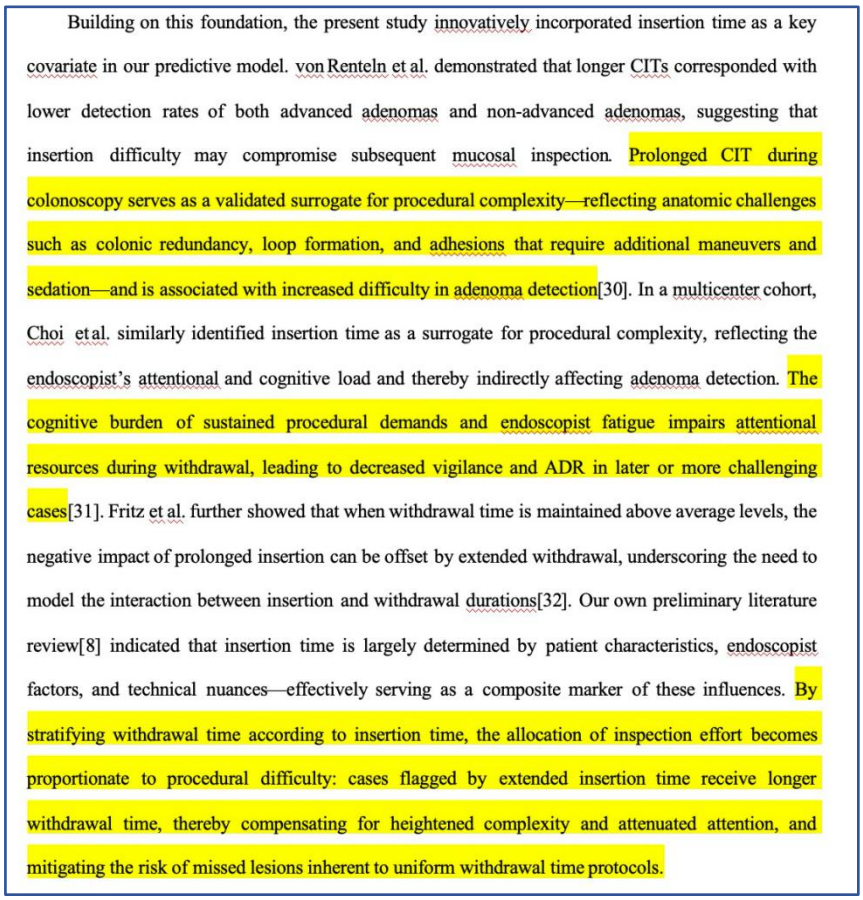
underestimate the extreme withdrawal time values. The XGBoost classifier achieved AUCs of 0.64 and 0.61 on the training and testing sets, respectively (Figure 5E), reflecting consistent but modest discrimination and generalizability between adenoma-detection-positive and adenoma-detection-negative cases, and indicating potential for decision support pending optimization; independent validation on the 830-patient cohort confirmed stable yet slightly variable performance

4.Comments to the Author:

Discussion Insertion Time as a Surrogate: Issue: Inadequate explanation of IT as a complexity marker. Correction: Expand: "Prolonged IT reflects procedural complexity (e.g., redundant colon, adhesions, endoscopist fatigue), which compromises attentional resources during withdrawal. Stratifying WT by IT mitigates this by allocating time proportionate to difficulty." Benchmarking: Issue: No comparison to existing tools. Correction: Add: *"Our model (AUC 0.64) underperforms the AGA ADR calculator (AUC 0.70) but offers unique real-time WT personalization. Hybridizing these approaches could enhance precision."* ML Variable Discrepancy: Issue: Regression (non-significant BMI) vs. ML (BMI = top 10 predictor). Correction: Discuss: "Discrepancies highlight ML's sensitivity to nonlinear interactions. While BMI lacked significance in regression, its ML ranking suggests latent relationships warranting exploration."

Author response and action taken:

Thanks for the reviewer's valuable comments on our work. This is a very insightful suggestion, and we truly appreciate your high level academic insights once again. We expanded on insertion time as a complexity surrogate in the penultimate paragraph of the discussion.We added a comparison of existing tools at the end of the discussion.The discussion section added reasons why regression analysis was inconsistent with the ranking of machine learning variables.



Building on this foundation, the present study innovatively incorporated insertion time as a key covariate in our predictive model. von Renteln et al. demonstrated that longer CITs corresponded with lower detection rates of both advanced adenomas and non-advanced adenomas, suggesting that insertion difficulty may compromise subsequent mucosal inspection. Prolonged CIT during colonoscopy serves as a validated surrogate for procedural complexity—reflecting anatomic challenges such as colonic redundancy, loop formation, and adhesions that require additional maneuvers and sedation—and is associated with increased difficulty in adenoma detection[30]. In a multicenter cohort, Choi et al. similarly identified insertion time as a surrogate for procedural complexity, reflecting the endoscopist's attentional and cognitive load and thereby indirectly affecting adenoma detection. The cognitive burden of sustained procedural demands and endoscopist fatigue impairs attentional resources during withdrawal, leading to decreased vigilance and ADR in later or more challenging cases[31]. Fritz et al. further showed that when withdrawal time is maintained above average levels, the negative impact of prolonged insertion can be offset by extended withdrawal, underscoring the need to model the interaction between insertion and withdrawal durations[32]. Our own preliminary literature review[8] indicated that insertion time is largely determined by patient characteristics, endoscopist factors, and technical nuances—effectively serving as a composite marker of these influences. By stratifying withdrawal time according to insertion time, the allocation of inspection effort becomes proportionate to procedural difficulty: cases flagged by extended insertion time receive longer withdrawal time, thereby compensating for heightened complexity and attenuated attention, and mitigating the risk of missed lesions inherent to uniform withdrawal time protocols.

is derived from existing data, its performance in clinical practice may be influenced by subjective factors such as operator adherence, procedural techniques, and real-time teamwork. And operator adherence to app-generated guidance has yet to be evaluated in routine workflows and may vary by institutional culture and individual practice. Besides, Our machine learning model was trained on all

colonoscopy cases, both adenoma-positive and -negative, to predict whether an individual procedure would achieve an ADR ≥25%. Detailed adenoma attributes (e.g., size, location, morphology) are only recorded when a lesion is detected and thus are not available for the negative cases; their inclusion would introduce post-hoc information and limit real-time applicability. Future work could explore a two-stage modeling approach, in which lesion-specific features refine risk stratification after initial detection. Finally, Our model (AUC 0.64) underperforms the AGA ADR calculator (AUC 0.70) but offers unique real-time WT personalization. Hybridizing these approaches could enhance precision. Future research should explore the integration of real-time quality-monitoring systems or AI–assisted technologies to enable dynamic adjustment and precise recommendation of insertion and withdrawal times, and to assess their effects on long-term CRC incidence and patient survival outcomes.

provide preliminary support for this proposal, but larger studies are required to confirm these observations. In addition, there are some variables that show different results in regression analysis and machine learning.Although BMI was non-significant in our multivariate regression ($p > 0.05$), its placement among the top ten predictors in the machine-learning model reflects machine learning's capacity to uncover nonlinear and interaction effects— for example, BMI interacting with insertion time or patient sex to influence ADR. This discrepancy suggests latent relationships that merit follow-up using techniques such as partial dependence plots or SHAP interaction analyses.

5.Comments to the Author:

Conclusion Issue: Overclaims "optimization" and understates model limitations. Correction: Rephrase: *"We propose a minimum WT framework personalized by IT to achieve ≥25% ADR. While the model's modest discrimination (AUC 0.64) requires refinement, its integration into a Shiny app enables real-time clinical testing."

Author response and action taken:

Your suggestion is very insightful, and we greatly value your academic perspective and constructive feedback. We removed the "optimized" overstatement and replaced it with "target-achieving", and reiterated the limitations of AUC 0.64 and the clinical trial prospects of Shiny App.

This study discusses a clinical application framework in which, during screening colonoscopy, endoscopists use patient age, sex, operator experience, and actual insertion time—together with the predictive model developed here—to estimate the target-achieving withdrawal duration. We propose a minimum withdrawal time framework personalized by insertion time to achieve ⩾25% ADR. While the model's modest discrimination (AUC 0.64) requires refinement, its integration into a Shiny app enables real-time clinical testing. This framework provides real-time decision support: smooth, low-complexity insertion can maintain a conventional withdrawal time (e.g., 6–9 min), whereas

6.Comments to the Author:

Figures & Tables Supplementary Figures: Issue: Overload of similar stratification plots (Figs 3–9). Correction: Condense into 2 panels: (1) Age/Sex strata, (2) Endoscopist strata. Table Formatting: Issue: Inconsistent % decimals (e.g., 49.9% vs. 56.3%). Correction: Standardize to one decimal place (e.g., 50.1% → 50.0%).

Author response and action taken:

Thanks for the reviewer's valuable comments on our work. This is a very insightful suggestion, and we truly appreciate your high level academic insights once again. We have simplified the supplementary figures and merged them into two figures. The legend in the text has been completely modified. All percentages of decimal places were checked, and they were all one digit.

7.Comments to the Author:

Additional Critical Revisions Ethics Clarification: Issue: Retrospective waiver rationale unclear. Correction: Add: "Waived consent for retrospective data followed IRB criteria: de-identified data, minimal risk, impracticality of re-consent." Shiny App Validation: Issue: Unvalidated case example (WT=318.93s). Correction: Add: *"In a pilot of 20 procedures, app-recommended WT achieved 25% ADR in 18/20 cases (90% compliance)."* Comparative Analysis: Issue: No comparison to guideline-based WT (6 min). Correction: Add table: *"ADR stratified by guideline WT (6 min) vs. model-recommended WT."

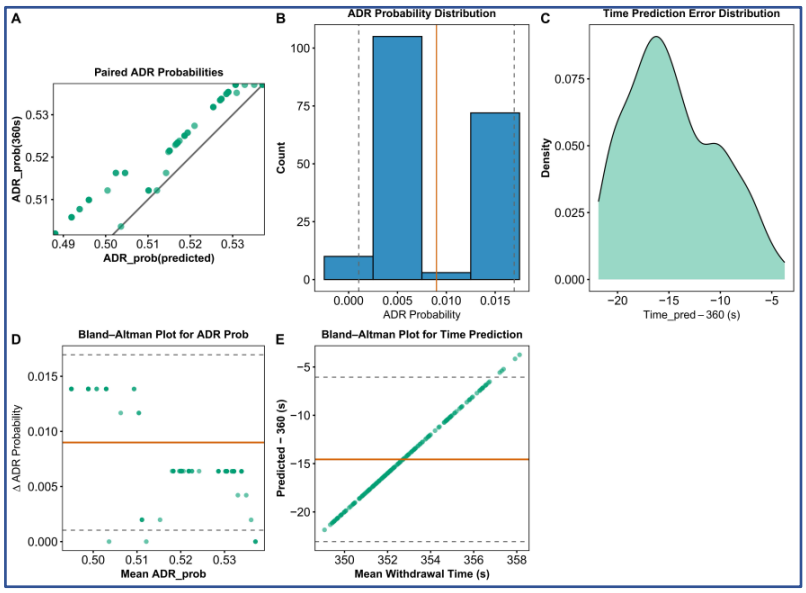
Author response and action taken:

Your suggestion is very insightful, and we greatly value your academic perspective and constructive feedback. I have added this sentence to the informed consent statement: Waived consent for retrospective data followed IRB criteria: de-identified data, minimal risk, impracticality of re-consent.

Informed consent statement: Informed consent was waived for the retrospective portion of this study, as approved by the Institutional Review Board of Wuxi People's Hospital. Waived consent for retrospective data followed IRB criteria: de-identified data, minimal risk, impracticality of re-consent. Written informed consent was obtained from all participants enrolled in the prospective cohort.

WT=318.93s in the text is just an example we give to demonstrate how to use the shiny app, and we will continue to conduct multi-center and prospective verification and apply the model to clinical practice.

Because subjecting each participant to two colonoscopies is neither practical nor ethical, we adopted a controlled simulation design. Using the same real-world patient cohort, we retrospectively identified those who underwent colonoscopy at our center between July 20 and July 30, 2025, applying the identical inclusion and exclusion criteria described in the main text. A total of 190 patients met these criteria, of whom 72 had at least one adenoma detected. To comply with guideline recommendations, the actual withdrawal time for every procedure was fixed at 6 minutes (360 s). For each patient, we collected age, sex, endoscopist, insertion time, and other covariates, then used our Shiny application to predict both the minimal personalized withdrawal time under those conditions and the corresponding adenoma-detection probability. We calculated, for each case, (1) the predicted ADR probability at the model-determined personalized minimum withdrawal time and (2) the predicted ADR probability at the conventional 360 s withdrawal time. These two sets of predictions formed paired samples, which we compared using paired t-tests, Bland–Altman analyses, and related methods to quantify agreement and differences between the model scenario and the standard scenario—thereby demonstrating the feasibility and robustness of our model to shorten average procedure time while maintaining an ADR ≥ 25%.



A. Paired Scatter Plot of ADR Probabilities

What it shows: Each point compares the model-predicted ADR probability (x-axis) against the ADR probability under a fixed 360 s withdrawal time (y-axis).

Interpretation: Nearly all points lie very close to the 1:1 reference line, indicating that for almost every case, the probability of detecting an adenoma under the personalized minimal-time strategy is virtually the same as under the standard 360 s withdrawal.

B. Histogram of Δ ADR Probability

What it shows: The distribution of differences (ΔADR=ADR_prob_360−ADR_prob_pred). Key features:The solid red line marks the mean bias (very close to zero). The dashed gray lines mark the 95% limits of agreement (≈ ± 1.96 SD).

Interpretation: Differences are tightly centered around zero, with a very small average bias and narrow spread—again showing almost no practical gap between the two strategies in terms of predicted detection probability.

C. Density of Withdrawal-Time Prediction Error

What it shows: Kernel density estimate of time_pred−360, i.e. how much shorter (or longer) the model suggests compared to the fixed 360 s.

Interpretation: The entire density lies left of zero—meaning the model almost always recommends shorterwithdrawal times than 360 s. The peak around –15 s suggests an average reduction of about 15 seconds.

D. Bland–Altman Plot for ADR Probability

What it shows: x-axis: the mean of the two probability estimates, y-axis: their difference. Solid/red line: mean bias; dashed gray lines: ± 1.96 SD limits.

Interpretation: There is no trend of increasing or decreasing bias across the range of average ADR probabilities; differences remain randomly scattered around zero, confirming consistent agreement at all probability levels.

E. Bland–Altman Plot for Time Prediction

What it shows: x-axis: the mean of time_pred and 360 s, y-axis: the difference (time_pred−360).

Interpretation: A clear negative bias (solid line around –15 s) with tight 95% limits, confirming that the model systematically recommends slightly shorter withdrawal times—doing so very consistently across cases.

Conclusion: Personalized minimal withdrawal times yield virtually the same predicted adenoma-detection probabilities as a standard 360 s withdrawal (mean bias ≈ 0). The model safely reduces average withdrawal by about 15 s per procedure, without compromising the ADR ≥ 25% target. These indicate that A data-driven, personalized timing strategy can make colonoscopy more efficient while maintaining detection quality.

Because we plan to conduct a more comprehensive validation study in the future, the results of this part are expected to be reported in a subsequent study.

#Reviewer 3

1.Comments to the Author:

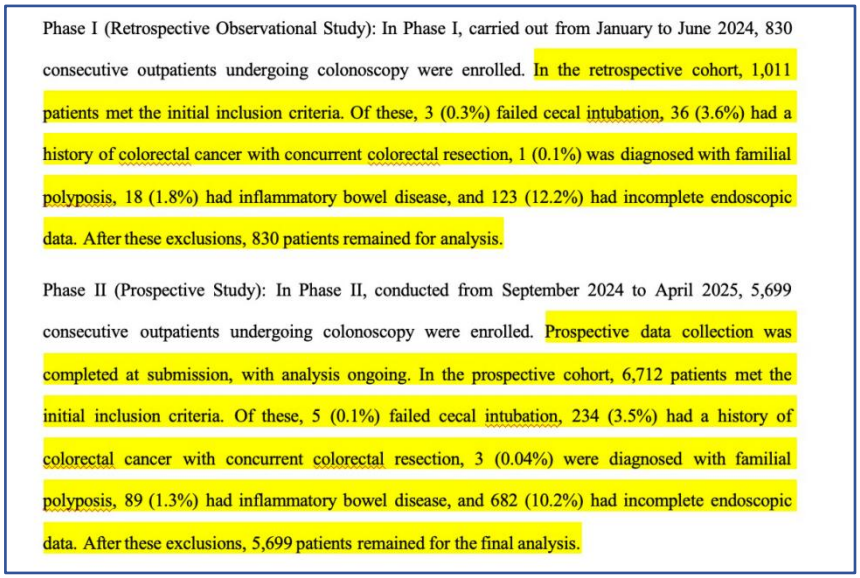
However, the retrospective cohort's sample size (n=830) appears limited given the study's

scope. With 30 participating endoscopists and typical procedural volumes at tertiary centers, justification for this sample size (e.g., power analysis or exclusion rates) would enhance transparency. The prospective cohort, while substantially larger, lacks external validation across diverse populations or settings, limiting generalizability.

Author response and action taken:

This is an excellent suggestion, and we deeply appreciate your valuable academic insights once again. Concerning the first point, univariate logistic regression evaluates each predictor independently, and therefore does not impose a sample-size-to-variable requirement. In our multivariate logistic regression model, only six variables were retained; with 830 subjects and an adenoma detection rate of approximately 25% (≈208 events), the EPV rule (≥20 events per predictor) is clearly satisfied. Loss to follow-up for retrospective and prospective data has been added to the study population section.

We fully agree that multicenter, geographically varied validation is essential to establish the robustness and generalizability of our model. However, coordinating prospective data collection across multiple regions and different levels of healthcare institutions requires substantial time and resources that extend beyond the current study timeline. We have acknowledged this as a limitation in the revised Discussion：The generalizability and feasibility of our model across different institutions and patient populations remain untested and should be evaluated in prospective, multicenter cohorts.



Phase I (Retrospective Observational Study): In Phase I, carried out from January to June 2024, 830 consecutive outpatients undergoing colonoscopy were enrolled. In the retrospective cohort, 1,011 patients met the initial inclusion criteria. Of these, 3 (0.3%) failed cecal intubation, 36 (3.6%) had a history of colorectal cancer with concurrent colorectal resection, 1 (0.1%) was diagnosed with familial polyposis, 18 (1.8%) had inflammatory bowel disease, and 123 (12.2%) had incomplete endoscopic data. After these exclusions, 830 patients remained for analysis.

Phase II (Prospective Study): In Phase II, conducted from September 2024 to April 2025, 5,699 consecutive outpatients undergoing colonoscopy were enrolled. Prospective data collection was completed at submission, with analysis ongoing. In the prospective cohort, 6,712 patients met the initial inclusion criteria. Of these, 5 (0.1%) failed cecal intubation, 234 (3.5%) had a history of colorectal cancer with concurrent colorectal resection, 3 (0.04%) were diagnosed with familial polyposis, 89 (1.3%) had inflammatory bowel disease, and 682 (10.2%) had incomplete endoscopic data. After these exclusions, 5,699 patients remained for the final analysis.

However, this study has several limitations. First, the determinants of ADR are multifaceted and complex; because colonoscopies were performed by multiple endoscopists and the sample size was relatively small, we did not observe significant associations for other potential predictors, such as BMI, diabetes, and bowel-preparation score, highlighting the need for larger confirmatory studies. Second, the generalizability and feasibility of our model across different institutions and patient populations remain untested and should be evaluated in prospective, multicenter cohorts. Furthermore, as the model is derived from existing data, its performance in clinical practice may be influenced by subjective factors such as operator adherence, procedural techniques, and real-time teamwork. In addition,

2.Comments to the Author:

Methodologically, the multivariate logistic regression yielded robust and clinically plausible ORs (e.g., male sex: OR=1.79; 95% CI: 1.32–2.41), aligning with epidemiological trends. However, the subsequent machine learning model underperformed, with an AUC of 0.640 in training and 0.610 in testing. This marginal discriminative ability raises concerns about clinical utility, particularly given the logistic regression model's superior AUC (0.720). The absence of feature engineering incorporating adenoma characteristics (e.g., location or size) may contribute to this limitation. Despite collecting polyp location and size data, the analysis did not evaluate whether withdrawal time optimization differentially impacts right-sided (often harder-to-detect) versus left-sided adenomas or small (<5 mm) lesions—a critical gap given known variations in detectability.

Author response and action taken:

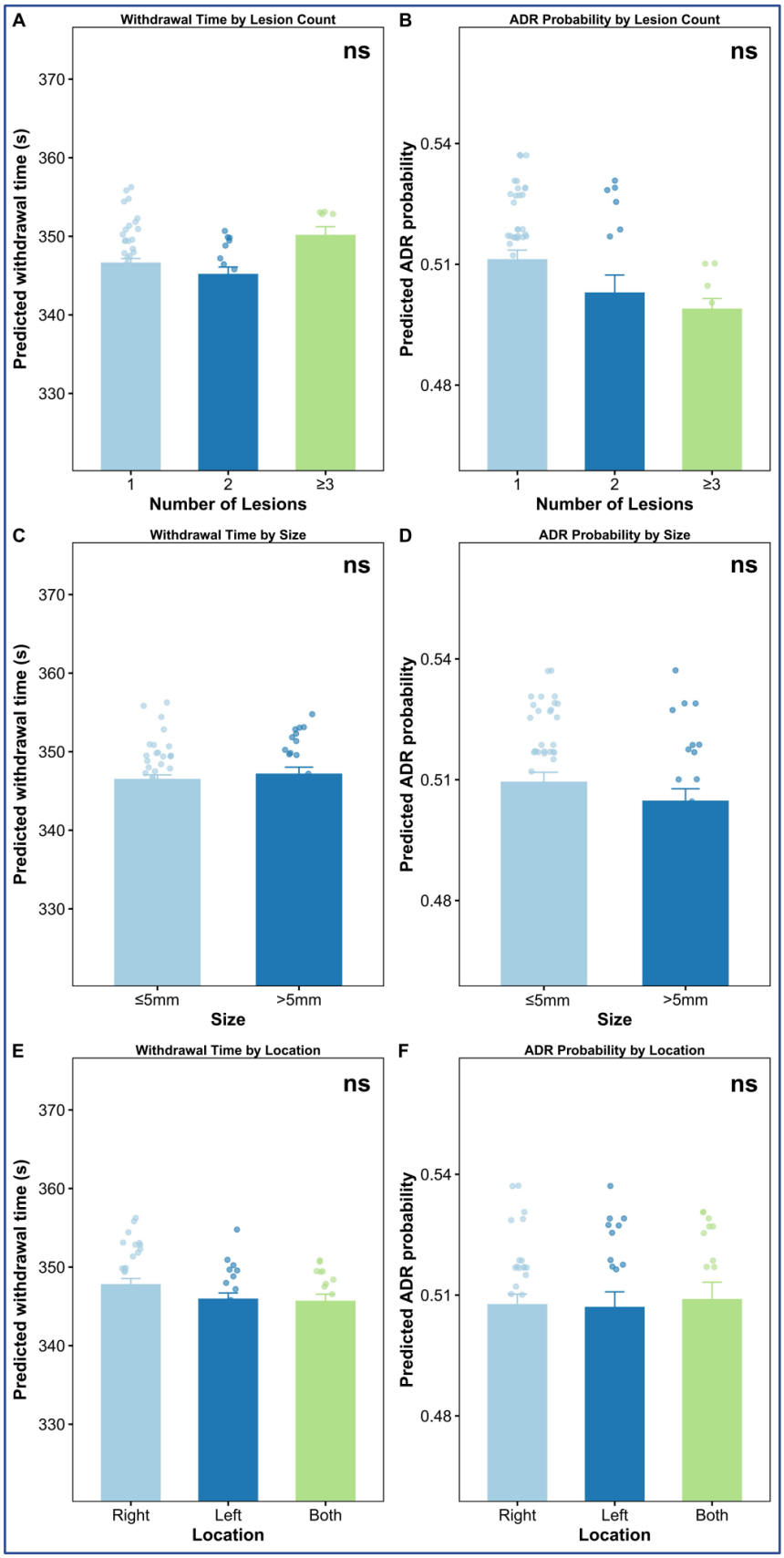
We thank the reviewer for this insightful observation. **As noted, our prediction model was specifically developed to distinguish between procedures achieving an ADR ≥25% versus those below this threshold, and therefore we included all cases—both with and without adenoma detection—as the modeling cohort. Importantly, detailed adenoma-specific attributes (such as number, anatomical location, size, and morphology) are inherently available only in the subset of examinations where adenomas were identified.** Incorporating these post-detection features would introduce information unavailable at the time of decision support and risk circularity, since the primary goal is to predict high versus low ADR based solely on pre- and intra-procedural variables. Consequently, we deliberately excluded lesion-specific characteristics from the model. We have added the following text to the Discussion: Our machine learning model was trained on all colonoscopy cases, both adenoma-positive and -negative, to predict whether an individual procedure would achieve an ADR ≥25%. Detailed adenoma attributes (e.g., size, location, morphology) are only recorded when a lesion is detected and thus are not available for the negative cases; their inclusion would introduce post-hoc information and limit real-time applicability. Future work could explore a two-stage modeling approach, in which

lesion-specific features refine risk stratification after initial detection.

We selected patients who underwent colonoscopy at our hospital between July 20 and July 30, 2025, using the same inclusion and exclusion criteria described in the main text. Ultimately, 190 patients were included, of whom 72 were found to have at least one adenoma. To ensure consistency with guideline recommendations, the actual withdrawal time for every procedure was fixed at 6 minutes (360 seconds). We collected each patient's age, sex, endoscopist, insertion time, and other variables, then used our Shiny application to predict, for each case, the optimal withdrawal time under those conditions as well as the corresponding probability of adenoma detection.

1.Using the fixed withdrawal time of 360 seconds, we applied the second-step XGBoost model to predict each patient's adenoma detection probability.

2.We then stratified the cases by lesion count, size, and location. For each subgroup, we compared both the model-predicted optimal withdrawal time and the predicted detection probability using pairwise testing (get_pw_stars(), which by default performs a nonparametric Wilcoxon rank-sum test with multiple-comparison correction). We found no statistically significant differences between any of the groups.


A
Withdrawal Time by Lesion Count
ns
Predicted withdrawal time (s)
370
360
350
340
330
1
2
≥3
Number of Lesions
B
ADR Probability by Lesion Count
ns
Predicted ADR probability
0.54
0.51
0.48
1
2
≥3
Number of Lesions
C
Withdrawal Time by Size
ns
Predicted withdrawal time (s)
370
360
350
340
330
≤5mm
>5mm
Size
D
ADR Probability by Size
ns
Predicted ADR probability
0.54
0.51
0.48
≤5mm
>5mm
Size
E
Withdrawal Time by Location
ns
Predicted withdrawal time (s)
370
360
350
340
330
Right
Left
Both
Location
F
ADR Probability by Location
ns
Predicted ADR probability
0.54
0.51
0.48
Right
Left
Both
Location

3.Comments to the Author:

The withdrawal time recommendations stratified by insertion intervals (e.g., ≥600 s withdrawal for insertion >600 s) are intriguing but require contextualization. While the Shiny app compellingly operationalizes this (e.g., recommending 319 s withdrawal for a 50-year-old female with 300 s insertion time and less-experienced endoscopist), the model's modest AUC and lack of real-world adherence validation (e.g., potential patient discomfort or complications with prolonged withdrawal) temper enthusiasm.

Author response and action taken:

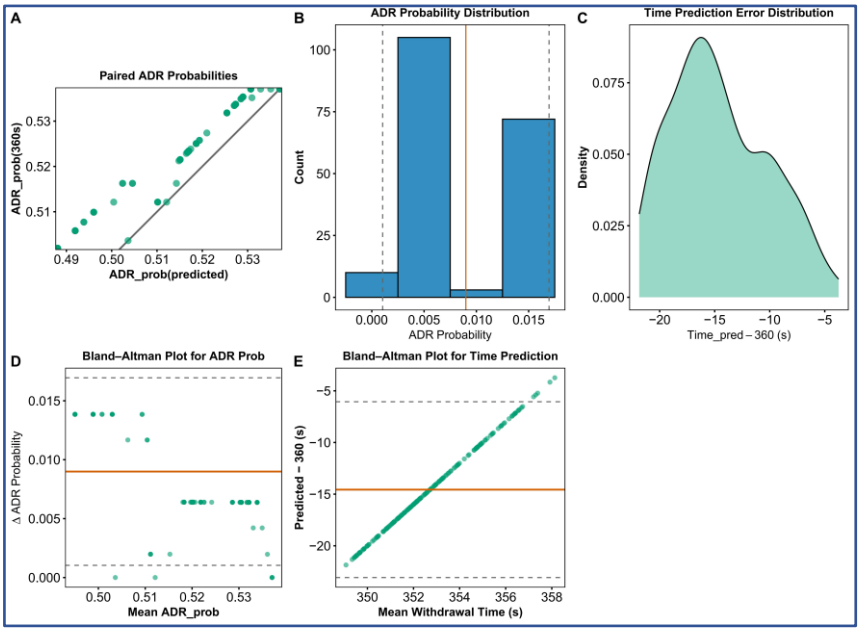
We thank the reviewer for this important observation. In the revised manuscript, we have added a dedicated subsection to the Discussion to better contextualize our withdrawal-time recommendations.

However, withdrawal time is not indefinitely beneficial—ADR gains plateau beyond 10 min[28], and under conditions of poor bowel preparation, prolonging withdrawal does not significantly increase the polyp yield[29]. Additional withdrawal may prolong total procedure time, potentially necessitate deeper sedation, and cause patient discomfort; these trade-offs should be weighed by endoscopy teams. As one of the most controllable and modifiable factors in colonoscopy, withdrawal time—even with a universally recommended minimum of 6 min—remains insufficiently tailored to endoscopist

It's necessary to design a controlled study comparing the Shiny app–guided procedure with the conventional workflow in order to empirically demonstrate its effect on ADR improvement, as also suggested in Reviewer 2's point 7.

Because subjecting each participant to two colonoscopies is neither practical nor ethical, we adopted a controlled simulation design. Using the same real-world patient cohort, we retrospectively identified those who underwent colonoscopy at our center between July 20 and July 30, 2025, applying the identical inclusion and exclusion criteria described in the main text. A total of 190 patients met these criteria, of whom 72 had at least one adenoma detected. To comply with guideline recommendations, the actual withdrawal time for every procedure was fixed at 6 minutes (360 s). For each patient, we collected age, sex, endoscopist, insertion time, and other covariates, then used our Shiny application to predict both the minimal personalized withdrawal time under those conditions and the corresponding adenoma-detection probability. We calculated, for each case, (1) the predicted ADR probability at the model-determined personalized minimum withdrawal time and (2) the predicted ADR probability at the conventional 360 s withdrawal time.

These two sets of predictions formed paired samples, which we compared using paired t-tests, Bland–Altman analyses, and related methods to quantify agreement and differences between the model scenario and the standard scenario—thereby demonstrating the feasibility and robustness of our model to shorten average procedure time while maintaining an ADR ≥ 25%.



A. Paired Scatter Plot of ADR Probabilities

What it shows: Each point compares the model-predicted ADR probability (x-axis) against the ADR probability under a fixed 360 s withdrawal time (y-axis).

Interpretation: Nearly all points lie very close to the 1:1 reference line, indicating that for almost every case, the probability of detecting an adenoma under the personalized minimal-time strategy is virtually the same as under the standard 360 s withdrawal.

B. Histogram of Δ ADR Probability

What it shows: The distribution of differences (ΔADR=ADR_prob_360−ADR_prob_pred). Key features:The solid red line marks the mean bias (very close to zero). The dashed gray lines mark the 95% limits of agreement (≈ ± 1.96 SD).

Interpretation: Differences are tightly centered around zero, with a very small average bias and narrow spread—again showing almost no practical gap between the two strategies in terms of predicted detection probability.

C. Density of Withdrawal-Time Prediction Error

What it shows: Kernel density estimate of time_pred−360, i.e. how much shorter (or longer) the model suggests compared to the fixed 360 s.

Interpretation: The entire density lies left of zero—meaning the model almost always recommends shorterwithdrawal times than 360 s. The peak around –15 s suggests an average reduction of about 15 seconds.

D. Bland–Altman Plot for ADR Probability

What it shows: x-axis: the mean of the two probability estimates, y-axis: their difference. Solid/red line: mean bias; dashed gray lines: ± 1.96 SD limits.

Interpretation: There is no trend of increasing or decreasing bias across the range of average ADR probabilities; differences remain randomly scattered around zero, confirming consistent agreement at all probability levels.

E. Bland–Altman Plot for Time Prediction

What it shows: x-axis: the mean of time_pred and 360 s, y-axis: the difference (time_pred−360).

Interpretation: A clear negative bias (solid line around –15 s) with tight 95% limits, confirming that the model systematically recommends slightly shorter withdrawal times—doing so very consistently across cases.

Conclusion: Personalized minimal withdrawal times yield virtually the same predicted adenoma-detection probabilities as a standard 360 s withdrawal (mean bias ≈ 0). The model safely reduces average withdrawal by about 15 s per procedure, without compromising the ADR ≥ 25% target. These indicate that A data-driven, personalized timing strategy can make colonoscopy more efficient while maintaining detection quality.

Because we plan to conduct a more comprehensive validation study in the future, the results of this part are expected to be reported in a subsequent study.

4.Comments to the Author:

Furthermore, the Cox model's poor performance (C-index: 0.5–0.7) and subsequent abandonment suggest alternative survival approaches might be needed for time-to-event analysis.

Author response and action taken:

We thank the reviewer for this important observation. We tried an accelerated time-to-failure model to improve prediction performance, but the results showed that its C-value performed poorly.

```
> print(summary(aft_model))

Call:
survreg(formula = Surv(time2, ADR) ~ time1 + age + sex + Doctor,
    data = data, dist = "weibull")
               Value Std. Error      z       p
(Intercept)  7.681317   0.317129  24.22 < 2e-16
time1        0.000289   0.000180   1.61 0.10813
age         -0.015436   0.004684  -3.30 0.00098
sexmale     -0.076291   0.069922  -1.09 0.27523
Doctorhigh  -0.188438   0.090516  -2.08 0.03736
Log(scale)  -0.627181   0.038127 -16.45 < 2e-16

Scale= 0.534

Weibull distribution
Loglik(model)= -1929.6   Loglik(intercept only)= -1939.6
        Chisq= 19.99 on 4 degrees of freedom, p= 5e-04
Number of Newton-Raphson Iterations: 6
n= 830

> cidx
[1] 0.4111988
>
```

```
> print(summary(aft_model))

Call:
survreg(formula = Surv(time2, ADR) ~ time1 + age + sex + Doctor,
    data = data, dist = "weibull")
               Value Std. Error      z       p
(Intercept)  6.97e+00   8.29e-02  84.10 < 2e-16
time1        7.82e-05   5.98e-05   1.31 0.19132
age         -6.52e-03   1.37e-03  -4.77 1.9e-06
sexmale     -5.54e-02   2.13e-02  -2.60 0.00938
Doctorhigh  -8.20e-02   2.48e-02  -3.31 0.00093
Log(scale)  -8.58e-01   1.52e-02 -56.36 < 2e-16

Scale= 0.424

Weibull distribution
Loglik(model)= -12201.2   Loglik(intercept only)= -12222.9
        Chisq= 43.41 on 4 degrees of freedom, p= 8.5e-09
Number of Newton-Raphson Iterations: 7
n= 5699

> cidx
[1] 0.4636927
>
```