Skepthical review: *3.2. Modeling the emergence of new conditions*

Summary

This manuscript analyzes approximately **3.11**M Texas **2018** inpatient discharge abstracts to approximate within-stay morbidity dynamics using Present on Admission (POA) flags: "initial morbidity" is defined from diagnoses with POA$=' Y'$, while "emergent conditions" are defined from diagnoses with POA$=' N'$ (Sec. 2.2.4). The stated aims are (i) to predict whether any emergent condition occurs during the hospitalization (HAD_EMERGENT_CONDITION) using admission-time information (Sec. 2.3.1, Sec. 3.2; Table 1), and (ii) to quantify the incremental association of emergent conditions with resource utilization (LOS; log total charges) by comparing baseline (admission features) vs full (baseline + emergent flags/counts) models (Sec. 2.4.2, Sec. 3.3; Table 2). The paper documents extensive preprocessing (Sec. 2.2–2.6) and reports: (a) essentially perfect performance for emergent-condition prediction (AUC/F1 $\approx$ **1.0** across models/subgroups; Sec. 3.2, Sec. 3.4.1; Figures 15–16), strongly indicating target leakage/circularity; and (b) modest out-of-sample fit for LOS/charges using baseline features (up to $\sim$ **0.32** R^2 for LOS and $\sim$ **0.57** for log-charges; Sec. 3.3), with negligible or slightly negative incremental R^2 when adding coarse emergent-condition indicators (binary/counts). While the manuscript is commendably candid that the perfect prediction invalidates SHAP explanations and hospital O/E comparisons (Sec. 3.2, Sec. 3.4.2, Sec. 3.5), the current implementation leaves the main predictive objective unresolved and the clinical interpretation of "emergent conditions" weak—particularly because POA$=' N'$ diagnoses appear overwhelmingly dominated by ICD-10 Chapter O (pregnancy/childbirth) codes (Sec. 3.1), raising concerns that the target is not a coherent proxy for hospital-acquired complications and may drive the resource-utilization conclusions.

Strengths

- Large, real-world statewide dataset ($>$ **3.1**M discharges) with clear overview of the source and preprocessing pipeline (Sec. 2.1–2.2, Sec. 3.1).
- Transparent description of key feature-engineering steps (POA-based diagnosis partitioning, missingness indicators, LOS handling, log transform for charges) and multiple model families/metrics (Sec. 2.2–2.4, Sec. 2.6, Sec. 3.2–3.3; Tables 1–2).
- Appropriately flags the perfect/near-perfect emergent-condition prediction as leakage/circularity and warns against clinical inference from SHAP/hospital comparisons derived from those models (Sec. 3.2, Sec. 3.5, Conclusions).
- The baseline vs full model comparison for LOS/charges is a sensible framing for assessing incremental predictive value (Sec. 2.4.2, Sec. 3.3), and the negative finding (minimal gain from coarse emergent features) is potentially important if endpoint definitions are refined.

- Includes subgroup and hospital-oriented exploratory analyses (Sec. 3.4), which—once leakage is resolved and endpoints are clinically coherent—could become a useful lens on coding heterogeneity and case-mix.

Major issues

1. **Emergent-condition prediction (HAD_EMERGENT_CONDITION) shows perfect/near-perfect discrimination (AUC-ROC/AUC-PR/F1 ≈ 1.0) across models and subgroups (Sec. 3.2; Table 1; Figures 15–16), indicating severe target leakage or circular construction.** Although acknowledged, the manuscript does not isolate which predictors encode post-admission/discharge-only information (e.g., diagnosis-group features potentially built from all diagnoses, POA-missingness artifacts, downstream fields), nor does it provide a leakage-free reformulation—leaving a primary stated objective unresolved (Sec. 2.3.1, Sec. 3.2, Sec. 3.5, Conclusions).

   Recommendation: Add an explicit leakage audit and a corrected modeling attempt. Concretely: (a) In Sec. 2.3.1 (or an appendix), enumerate every predictor used in the classifier (including one-hot diagnosis-group indicators, counts, missingness flags, POA-uncertainty flags, any physician/hospital identifiers, and any discharge disposition–like fields), tagging each with its provenance and whether it is available at admission. (b) Verify in code (ideally with unit-test style checks) that any diagnosis-derived features for the emergence model are computed exclusively from POA$=' Y'$ diagnoses (and not from the union of all diagnoses due to a merge/mapping step). (c) Re-run the emergence task with a strictly admission-time feature set (demographics; admission type/source; payer; principal diagnosis and comorbidities derived only from POA $=' Y'$; counts of POA$=' Y'$ diagnoses). (d) Use evaluation splits that help detect facility-level/coding artifacts (e.g., hospital-held-out split and/or temporal split across months) and report performance there. If leakage-free prediction is not feasible from discharge abstracts alone, explicitly reframe the emergence-prediction objective (Sec. 3.2, Conclusions) as a cautionary demonstration rather than a substantive predictive contribution.

2. **The operational definition of "emergent condition" as any POA$=' N'$ diagnosis is clinically heterogeneous and appears overwhelmingly dominated by obstetric (ICD-10 Chapter O) codes (Sec. 2.2.4, Sec. 3.1).** This raises a 'bigger picture' concern: the target may primarily capture routine delivery-related coding rules rather than hospital-acquired morbidity, undermining interpretability of both the emergence task and the conclusion that emergent morbidity adds negligible information for LOS/charges (Sec. 3.3, Sec. 3.5, Conclusions).

   Recommendation: Refine and stratify the endpoint in Sec. 2.2.4 and re-run the core analyses. Minimum set of revisions: (a) Separate obstetric vs non-obstetric analyses (e.g., exclude Chapter O diagnoses and/or identify delivery-related encounters via

principal diagnosis/procedure where possible) and report prevalence for each stratum in Sec. 3.1. (b) Define one or more clinically coherent complication endpoints (e.g., CMS HAC-like groups, PSI-like complications, or a curated list of complication CCSR groups where POA is intended to distinguish complications from comorbidity) and repeat the emergence and utilization analyses on these endpoints. (c) Report how conclusions (incremental R^2; effect directions) change when excluding or separately modeling pregnancy-related stays, and discuss explicitly in Sec. 3.5/Conclusions how Chapter O dominance affects the headline results.

3. **Interpretation of POA$='N'$ as 'developed during hospitalization' and language implying 'impact' on LOS/charges is not well supported by the data-generating process: POA is a documentation/coding flag without onset time; POA$='N'$ can reflect delayed recognition, coding variation, or POA uncertainty rather than true in-hospital acquisition (Sec. 1, Sec. 2.2.4, Sec. 3.3, Sec. 3.5, Conclusions).** This creates a risk of causal overinterpretation and, for utilization models, post-treatment adjustment confusion (emergent diagnoses are recorded at discharge and are downstream of care intensity/LOS/charges).

   Recommendation: Tighten the inferential framing throughout. (a) In the Introduction and Sec. 2.2.4, explicitly state that POA$='N'$ does not provide a timestamp and is not equivalent to 'hospital-acquired' without additional clinical validation. (b) In Sec. 3.3 and Conclusions, replace causal language ('impact', 'contribution', 'drives') with association/prediction language, and clearly label models that include emergent features as 'hindsight'/discharge-informed rather than admission-time models. (c) If you want an 'incremental impact' interpretation, either (i) redesign around a causal estimand (e.g., mediation-aware framing) or (ii) restrict to complication definitions with stronger face validity (see Major Issue 2) and emphasize descriptive associations only.

4. **Hospital and subgroup comparisons based on the emergent-condition prediction model—especially hospital observed-to-expected (O/E) ratios and subgroup AUC$=1.0$ results—are not interpretable given leakage and likely heterogeneity in coding completeness/intensity across hospitals (Sec. 3.4.1–3.4.2; Figures 15–16).** Even if leakage is fixed, hospital comparisons will remain highly sensitive to POA coding practices and 'non-informative POA' handling.

   Recommendation: Revise Sec. 3.4 to avoid misleading performance/performance-comparison implications. (a) Remove or quarantine (clearly labeled as invalid artifacts) any O/E ratios and subgroup AUC summaries derived from the leaky model; do not interpret them as hospital performance or risk-adjusted comparisons. (b) If hospital-level results are retained after fixing leakage and endpoint definitions, add uncertainty intervals and consider hierarchical/shrinkage approaches; explicitly discuss coding completeness as a confounder (and, where feasible, report POA completeness metrics by hospital and their correlation with HAD_EMERGENT_CONDITION). (c) Consider switching hospital-level reporting to leakage-free descriptive rates (with minimal adjustment) framed as coding/case-mix exploration, not quality measurement.

5. **Key resource-utilization conclusions rely on very coarse emergent-condition features (binary presence and counts of POA=′ N' diagnoses) and on models with modest fit and heteroscedastic residuals (especially for LOS; Sec. 3.3).** The manuscript does not quantify uncertainty around the incremental R^2 changes (often tiny deltas) nor provide effect sizes on original scales, limiting the strength and interpretability of the 'negligible incremental value' conclusion (Sec. 3.3, Sec. 3.5, Conclusions).

   Recommendation: Strengthen Sec. 3.3 with effect sizes and uncertainty. (a) Report out-of-sample deltas explicitly (e.g., $+0.0004$ in R^2) and include bootstrap or cross-validation confidence intervals for baseline vs full model performance. (b) Provide interpretable effect sizes: for linear models, coefficients/SEs/CIs for emergent features and translate to days (LOS) and dollars (charges) where possible; for tree models, provide partial dependence or SHAP marginal effects with uncertainty (or at least stratified averages). (c) Explore richer but still interpretable emergent features tied to clinically coherent subsets (Major Issue 2), and consider interactions with baseline severity/service line. (d) Consider alternative outcome models more appropriate for skew/heteroscedasticity (e.g., log-LOS, negative binomial for LOS; Gamma/Tweedie for charges; quantile regression) and report whether conclusions are robust.

6. **Very high reported missingness for PAT_AGE ($\sim 59.3\%$) and its handling via midpoint mapping/median imputation plus missingness indicator (Sec. 2.2.2, Sec. 3.1) raises a 'bigger picture' data-quality concern: this may reflect masked/redacted age coding (e.g., neonates/very old) or an extraction/mapping bug.** If systematic, it can distort subgroup analyses, fairness interpretations, and any age gradients in LOS/charges and emergent condition rates.

   Recommendation: Validate and document the age field and missingness mechanism. (a) Cross-check PUDF documentation and show a pre-/post-mapping table of age code distributions (Sec. 2.2.2, Sec. 3.1). (b) Distinguish truly missing vs masked/interval-coded age categories; consider modeling age as categorical bands (including a 'masked/unknown' category) rather than midpoint+imputation when missingness is structural. (c) Add sensitivity analyses restricting to records with non-missing/usable age and report whether the main utilization conclusions (incremental value of emergent features; R^2 patterns) change.

Minor issues

1. Reproducibility/validation details are incomplete: exact feature sets per model (especially diagnosis-derived features), encoding decisions, hyperparameter grids/selection criteria, and data-splitting protocol (hold-out vs CV; stratification; any grouping by hospital) are not fully specified (Sec. 2.3–2.4, Sec. 2.6).

Recommendation: Add a compact reproducibility appendix: (a) full feature list per task/model family with encodings; (b) hyperparameter search spaces and selection metrics; (c) precise split/CV design including any stratification/grouping; and (d) software/library versions (including the exact CCSR version/mapping).

2. Handling of POA categories other than Y/N (e.g., U/W/E mapped to "Non-Informative POA") may materially affect both prevalence and model behavior, but is not evaluated as a sensitivity analysis (Sec. 2.2.4).

   Recommendation: Report the frequency of each POA category and run sensitivities: treat ambiguous POA codes as a separate category, and/or include them as 'unknown' rather than excluding; quantify the impact on emergent prevalence and on utilization associations.

3. Outlier handling is described qualitatively without counts/percentages affected (LOS winsorized at **365** days; charges corrections/transform) and with some internal inconsistency about charge preprocessing order (Sec. 2.2.3).

   Recommendation: In Sec. 2.2.3, report how many records are capped/trimmed/modified at each step; justify thresholds (e.g., **365** days); and provide a sensitivity analysis (e.g., exclude extreme LOS rather than cap; alternative caps). For charges, specify a single consistent pipeline including treatment of zeros/negatives and the order relative to \log(\text{TOTAL_CHARGES}+1).

4. Figures and subgroup plots (e.g., Figures 17–21) often lack sample sizes, uncertainty intervals, and clear statements of whether metrics are from CV or a held-out test set; some scatter/residual plots are likely unreadable at $N \sim$millions due to overplotting (Sec. 3.3–3.4).

   Recommendation: Add n per subgroup and CIs (bootstrap or across folds) to subgroup metrics; clearly state the evaluation protocol in each caption; use hexbin/density plots or downsampling for residual/scatter figures; and standardize axes/units/transform labels.

5. Engagement with related literature on POA usage, hospital-acquired conditions (CMS HAC/PSI), complication indices, and administrative-data risk adjustment is limited, weakening positioning and making it harder to interpret divergences from established approaches (Introduction, Sec. 4).

   Recommendation: Add a brief Related Work subsection (Intro or Methods) summarizing POA-based approaches and HAC/PSI-style endpoints; in Sec. 4, explicitly contrast your endpoint definition and discharge-abstract design with those frameworks and explain how this may yield Chapter O dominance and/or limited incremental predictive value.

6. Ethics/fairness implications are underdeveloped given analyses by race/ethnicity/payer and known variation in coding and charges (Sec. 3.4, Sec. 4).

Recommendation: Add a limitations/ethics paragraph (Sec. 3.5 or Sec. 4) addressing differential POA coding and billing practices, risks of misuse for hospital profiling/reimbursement (especially with leakage), and basic subgroup diagnostics you did/did not perform (e.g., calibration/error rates).

7. Template/presentation artifacts undermine professionalism and indexing: irrelevant keywords (e.g., "Computational astronomy") and non-academic affiliation text appear to be template leftovers (Abstract/front matter).

   Recommendation: Replace template artifacts with appropriate healthcare/ML keywords and correct affiliations to actual institutions (or remove if anonymized submission).

Very minor issues

1. Notation/terminology inconsistencies (e.g., LENGTH_OF_STAY vs LENGTH_STAY; mixed R-squared notation; unspecified log base for LOG_TOTAL_CHARGES) reduce clarity (Sec. 3.3; Table 2; throughout).

   Recommendation: Standardize variable names and mathematical notation throughout; explicitly state the log base (e.g., natural log) for LOG_TOTAL_CHARGES.

2. Typos and formatting glitches (split words, truncated text such as "OPERATING_PHYSICIAN_UNIF_ID (44.0", inconsistent section-heading markdown) hinder readability (Sec. 1, Sec. 2.2.1, Sec. 3.1–3.4).

   Recommendation: Proofread and clean formatting; ensure all truncated phrases are fixed and headings are consistent.

3. Figure captions can be generic/repetitive and sometimes reference unlabeled features; accessibility could be improved (font sizes, color choices).

   Recommendation: Make captions figure-specific and ensure labels match the plot; increase resolution/font size and use colorblind-safe palettes consistently.

4. Some qualitative claims could be made directly checkable by stating the computed values (e.g., 'nearly all' Chapter O emergent cases) (Sec. 3.1).

   Recommendation: Replace qualitative terms with explicit percentages and denominators (e.g., $\mathbf{96.7\%\ (49,030/50,681)}$), and report deltas numerically when emphasizing changes (e.g., R^2 $\mathbf{+0.0004}$).

Key statements and references

- • **Based on Present on Admission indicators in the 2018 Texas Hospital Inpatient Discharge Public Use Data File, $50,681$ of $3,110,296$ inpatient discharge records (1.63%) had at least one diagnosis coded as POA $=' N'$,**

indicating an emergent condition as defined in this study, while 98.37% of records had no such diagnosis.

- *Reference(s):* (none)

- **• Using Logistic Regression, Random Forest, and XGBoost models to predict the binary outcome HAD_EMERGENT_CONDITION from admission-time features, all models achieved perfect or near-perfect performance (AUC-ROC = 1.0, AUC-PR = 1.0, F1-score = 1.0, and Brier Score as low as 0.000), which the authors interpret as strong evidence of data leakage or a circular target-feature definition that invalidates clinical interpretation of these prediction results.**

- *Reference(s):* (none)

- **• In regression models predicting Length of Stay from baseline (admission-only) features, the Linear Regression model achieved $R^2 = 0.1128$, the Random Forest model achieved $R^2 = 0.3166$, and the XGBoost model achieved $R^2 = 0.3215$, indicating that initial patient characteristics explain only a modest proportion of LOS variance in this dataset.**

- *Reference(s):* (none)

- **• In regression models predicting log-transformed total charges (LOG_TOTAL_CHARGES) from baseline features, Linear Regression achieved $R^2 = 0.5166$, Random Forest achieved $R^2 = 0.5633$, and XGBoost achieved $R^2 = 0.5699$, showing that admission-time characteristics explain a moderate proportion of variation in hospital charges.**

- *Reference(s):* (none)

- **• Adding emergent-condition features (HAD_EMERGENT_CONDITION and NUM_EMERGENT_DIAGNOSES) to the baseline predictors produced minimal or no improvement in explanatory power for resource utilization, with R^2 for LOS remaining unchanged for Linear Regression (0.1128) and Random Forest (0.3166) and increasing only from 0.3215 to 0.3219 for XGBoost, while for LOG_TOTAL_CHARGES R^2 was unchanged for Linear Regression (0.5166) and slightly decreased for Random Forest (0.5633 to 0.5605) and XGBoost (0.5699 to 0.5694).**

- *Reference(s):* (none)

- **• Across demographic subgroups (age, payer, race, ethnicity), XGBoost models predicting HAD_EMERGENT_CONDITION continued to yield AUC-ROC values of 1.0 in all groups meeting minimum size thresholds, indicating that the data leakage or circularity affecting the overall emergence prediction task is systemic rather than confined to specific patient subpopulations.**

- *Reference(s):* (none)

Mathematical consistency audit

This section audits **symbolic/analytic** mathematical consistency (algebra, derivations, dimensional/unit checks, definition consistency).

Maths relevance: light

The document is primarily methodological and descriptive (ML workflow, feature engineering, and reported metrics) with very few explicit mathematical expressions. The main explicit formula is the log transform for charges; most other mathematical content is definitional (binary flags, counts) or refers to standard metrics (AUC, Brier score, R^2) without providing formulas. Internal consistency is generally acceptable, with some ambiguities/inconsistencies in preprocessing descriptions and variable naming.

Checked items

1. ✔ **Emergent-condition binary target definition** (Sec. 2.2.4 and Sec. 2.3.1, pp. 4–5)
    - **Claim:** HAD_EMERGENT_CONDITION is a binary indicator equal to 1 iff NUM_EMERGENT_DIAGNOSES > 0, where emergent diagnoses are those with POA$=' N'$.
    - **Checks:** definition consistency, logical implication check
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** POA categories are correctly mapped/cleaned beforehand, NUM_EMERGENT_DIAGNOSES counts only diagnoses with POA$=' N'$ (excluding non-informative/missing POA as stated)
    - **Notes:** The definitions of emergent diagnoses (POA$=' N'$), their count, and the derived binary flag are mutually consistent.
2. ✔ **Initial-state definition via POA$=' Y'$** (Sec. 2.2.4, p. 4)
    - **Claim:** Initial diagnoses are those with POA$=' Y'$ and are used to construct initial-state features including NUM_INITIAL_DIAGNOSES and grouped-code indicators.
    - **Checks:** definition consistency, set partition/overlap check
    - **Verdict:** PASS; confidence: medium; impact: moderate
    - **Assumptions/inputs:** POA$=' Y'$ indicates present on admission as used in the paper, Diagnoses with 'Non-Informative POA' or 'Missing POA Data' are excluded from initial/emergent lists unless analyzed separately
    - **Notes:** As described, the initial set (POA$=' Y'$) and emergent set (POA$=' N'$) are disjoint, with other POA states excluded; this is internally consistent.

3. ✔ **Principal-diagnosis POA uncertainty flag** (Sec. 2.2.4 and Sec. 2.3.1, p. 4–5)
    - **Claim:** PRINC_POA_UNCERTAIN flags records where the principal diagnosis has 'Missing POA Data' or 'Non-Informative POA' and is used as a predictor.
    - **Checks:** definition consistency, feature-target separation sanity check (symbolic)
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** Principal-diagnosis POA categories include the mapped 'Non-Informative POA' and explicit missing category
    - **Notes:** The flag is consistent with the stated POA cleaning scheme and does not, by itself, create a mathematical inconsistency (though it could correlate with target empirically).
4. ✔ **LOS winsorization rule** (Sec. 2.2.3 and Sec. 3.1, pp. 3 and 6)
    - **Claim:** LENGTH_OF_STAY is capped at **365** for outlier management.
    - **Checks:** definition consistency, units/dimensional sanity
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** LOS is measured in days and is converted to integer prior to winsorization as stated
    - **Notes:** Capping at a fixed number of days is mathematically coherent and consistent across methods/results narration.
5. ✔ **Charge log-transform definition** (Sec. 2.2.3, p. 3)
    - **Claim:** LOG_TOTAL_CHARGES is defined as \log(\text{TOTAL_CHARGES} + 1).
    - **Checks:** algebraic validity, domain check
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** TOTAL_CHARGES is nonnegative after cleaning, or at least \text{TOTAL_CHARGES} + 1 remains positive for all modeled rows
    - **Notes:** The transform is algebraically valid and ensures the log argument is positive when TOTAL_CHARGES $\geq$ **0**.
6. ⚠ **Zero/negative charges handling vs stated log transform** (Sec. 2.2.3 and Sec. 3.1, pp. 3 and 6)
    - **Claim:** Negative charges are set to NaN; zero charges may be set to a small positive value (e.g., \1)***beforeapplying***\log(\text{TOTAL_CHARGES} + 1)$.
    - **Checks:** pipeline/operation-order consistency, definition consistency
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate

- **Assumptions/inputs:** The transform and any zero-replacement are applied consistently to all records used in modeling
- **Notes:** If zeros are replaced with $\1 and then \log(\text{TOTAL_CHARGES}+1) is applied, the stated transform no longer maps zero to 0$. The text also later mentions 'winsorizing extreme values' for charges despite earlier stating no winsorization unless clear error. The exact finalized preprocessing sequence is not fully specified.

7. ✔ **Baseline vs Full model comparison via R-squared** (Sec. 2.4.3, p. 5 and Sec. 3.3/Table 2, pp. 7–9)
    - **Claim:** Incremental explanatory power of emergent-condition features is assessed by comparing R^2 of baseline and full models.
    - **Checks:** logical consistency of metric-based comparison
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** R^2 is computed on comparable test folds/splits for baseline and full models
    - **Notes:** Using ΔR^2 between nested/non-nested feature sets is a coherent analytic criterion (symbolically). The paper does not provide formulas, but the described comparison is consistent.

8. ✖ **Length-of-stay target naming consistency** (Table 2, p. 8; Sec. 2.4.1–2.4.2, pp. 4–5)
    - **Claim:** The LOS target is consistently the cleaned LENGTH_OF_STAY variable used across models.
    - **Checks:** notation consistency
    - **Verdict:** FAIL; confidence: high; impact: minor
    - **Assumptions/inputs:** LENGTH_STAY in Table 2 refers to the same cleaned variable as LENGTH_OF_STAY
    - **Notes:** Table 2 lists 'LENGTH_STAY' for the XGBoost rows while other rows and the methods use 'LENGTH_OF_STAY'. This is a notational inconsistency that should be corrected/clarified.

9. ⚠ **Hospital observed-to-expected (O/E) ratio definition** (Sec. 2.5.2 and Sec. 3.4.2, pp. 5 and 12–13)
    - **Claim:** A hospital's O/E ratio is computed as observed emergent-condition rate divided by an expected rate from the emergence model.
    - **Checks:** definition completeness, normalization/aggregation consistency
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Observed rate is computed per hospital on the same cohort used for expected rate, Expected rate is aggregated from model outputs in a well-defined way (e.g., mean predicted probability)

- **Notes:** The paper describes O/E qualitatively but does not provide explicit formulas for expected rate aggregation or handling of exclusions/missingness. This blocks full symbolic verification of the O/E construction.

Limitations

- The PDF contains very few explicit equations/derivation steps; most mathematical objects (AUC, Brier score, R^2, SHAP) are referenced without definitions or formulas, limiting symbolic verification to variable definitions and preprocessing transforms.
- No explicit modeling equations (e.g., linear regression specification, link functions, loss functions) are provided, so internal algebraic checks of model forms are not possible from the text alone.
- Several steps are described procedurally (data cleaning, winsorization, imputation) without a single consolidated pipeline/order-of-operations, which limits consistency checking to identifying ambiguities rather than proving equivalence.

Numerical results audit

This section audits **numerical/empirical** consistency: reported metrics, experimental design, baseline comparisons, statistical evidence, leakage risks, and reproducibility.

Arithmetic consistency checks largely pass for percentages derived from counts (emergent share; complement to 100%; implied non-emergent share; missing age share) and for several Table 2 R^2 comparisons. Two items remain UNCERTAIN because they rely on qualitative language without a defined numeric threshold or lacked an explicitly parsed expected delta despite the computed change being available.

Checked items

1. ✔ **C1_emergent_pct_from_counts** (Page 6, Section 3.1)
    - **Claim:** "$50,681$ records (1.63% of the total dataset) had at least one diagnosis coded as POA=$'N'$ …" with total analytical cohort "$3,110,296$ inpatient discharge records."
    - **Checks:** percentage_from_counts
    - **Verdict:** PASS
    - **Notes:** Computed percent $= 100 \times 50,681/3,110,296 = 1.629459...$, consistent with reported 1.63% within tolerance.
2. ✔ **C2_non_emergent_pct_complement** (Page 6, Section 3.1)
    - **Claim:** "The large majority of records (98.37%) did not have any such diagnosis." (complement of 1.63%).
    - **Checks:** percent_complement_to_100
    - **Verdict:** PASS
    - **Notes:** Reported percents sum to 100.00 exactly ($1.63 + 98.37$).

3. ✔ **C3_non_emergent_count_from_total** (Page 6, Section 3.1)
    - **Claim:** Given total $3,110,296$ and emergent $50,681$, implied non-emergent count should be $3,059,615$; paper reports 98.37% non-emergent but not the count.
    - **Checks:** parts_sum_to_total
    - **Verdict:** PASS
    - **Notes:** Computed non-emergent count $= 3,110,296 - 50,681 = 3,059,615$ (exact). Implied percent $= 98.3705409\%$, consistent with reported 98.37% within tolerance.
4. ⚠ **C4_chapterO_near_total_emergent_share** (Page 6, Section 3.1)
    - **Claim:** "Chapter O … occurring in $49,030$ records, which is nearly the total number of records with any emergent condition." Total emergent records reported as $50,681$.
    - **Checks:** ratio_or_share_from_counts
    - **Verdict:** UNCERTAIN
    - **Notes:** Computed share $= 49,030/50,681 = 0.9674237$ (96.742%), remainder $= 1,651$ records. No explicit numeric criterion provided for "nearly."
5. ✔ **C5_age_missing_count_from_percent** (Page 6, Section 3.1)
    - **Claim:** "59.3% of records had missing original PAT_AGE values" with analytical cohort size $3,110,296$.
    - **Checks:** count_from_percentage
    - **Verdict:** PASS
    - **Notes:** Computed missing count $= 0.593 \times 3,110,296 = 1,844,405.528$; rounded to $1,844,406$ gives implied percent 59.300015%, matching 59.3% within tolerance.
6. ⚠ **C6_table2_xgb_los_delta_r2** (Page 8, Table 2; Page 9 text discussing change)
    - **Claim:** Table 2 reports XGBoost LENGTH_OF_STAY R^2 Baseline 0.3215 and Full 0.3219; text says "increased only marginally (from 0.3215 to 0.3219)".
    - **Checks:** difference_between_reported_numbers
    - **Verdict:** UNCERTAIN
    - **Notes:** Computed delta $= 0.3219 - 0.3215 = 0.00040000000000001146$. The arithmetic is available, but an explicit expected delta was not parsed for an exact equality check.
7. ✔ **C7_table2_rf_charges_delta_r2** (Page 8, Table 2; Page 9 text states slight decrease)

- **Claim:** Table 2 reports Random Forest LOG_TOTAL_CHARGES R^2 Baseline **0.5633** and Full **0.5605**; text says R^2 slightly decreased in the Full model.
- **Checks:** difference_between_reported_numbers
- **Verdict:** PASS
- **Notes:** Computed delta $= 0.5605 - 0.5633 = -0.0028$, matching the expected decrease.

8. ✔ **C8_table2_xgb_charges_delta_r2** (Page 8, Table 2; Page 9 text states slight decrease)
    - **Claim:** Table 2 reports XGBoost LOG_TOTAL_CHARGES R^2 Baseline **0.5699** and Full **0.5694**; text says R^2 slightly decreased in the Full model.
    - **Checks:** difference_between_reported_numbers
    - **Verdict:** PASS
    - **Notes:** Computed delta $= 0.5694 - 0.5699 = -0.0005$, matching the expected decrease.
9. ✔ **C9_table2_linear_models_equal_r2** (Page 8, Table 2)
    - **Claim:** Table 2 reports identical R^2 for Linear Regression Baseline vs Full for both LENGTH_OF_STAY (**0.1128** vs **0.1128**) and LOG_TOTAL_CHARGES (**0.5166** vs **0.5166**).
    - **Checks:** equality_of_repeated_constants
    - **Verdict:** PASS
    - **Notes:** Both pairs are exactly equal as printed (**0.1128** and **0.5166**).
10. ✔ **C10_oeratio_summary_internal** (Page 13, Section 3.4.2)
    - **Claim:** Hospital O/E ratios: mean $\approx$ **0.050** (SD= **0.092**), median **0.0028**, range **0** to max **1.256**; "**0** (for **200** hospitals)".
    - **Checks:** range_and_order_sanity
    - **Verdict:** PASS
    - **Notes:** Sanity checks passed: min $\leq$ median $\leq$ max; mean within [min,max]; SD $\geq$ **0**; zero_count is nonnegative and integer-like. This does not validate the statistics against underlying data.

Limitations

- Only parsed text from the provided PDF was used; no external datasets or internet lookups were used.
- Values shown only in figures (plots) were not extracted or checked because pixel-based extraction is out of scope.
- Many performance and descriptive statistics (means/SDs, AUCs, R2s beyond arithmetic comparisons) cannot be independently recomputed without underlying data or supplementary tables not included here.