Skepthical review: *Analysis of Principal Diagnosis Present on Admission Status and Resource Utilization in Texas Inpatient Data*

Summary

The manuscript aims to study relationships between present-on-admission (POA) indicators, hospital-acquired conditions, and downstream resource utilization (length of stay [LOS], total charges) in the 2018 Texas THCIC inpatient discharge dataset. The original plan—leveraging up to 25 diagnoses per stay with POA flags to perform association rule mining, network analysis, and more complex predictive modeling—could not be executed because the authors were unable to reliably extract/link the OTH_DIAG_CODE_ *and POA_OTH_DIAG_* fields (Abstract, Introduction, Sec. 2.1–2.2, Sec. 3.2, Sec. 4). As a result, the study narrows to analyses using only the principal diagnosis POA flag: descriptives, a classification task predicting whether POA_PRINC_DIAG='N', and OLS regressions of **log(LOS)** and **log(charges)** on POA status and covariates (Sec. 2.2–2.5, Sec. 3.1, Sec. 3.3–3.4).

A central concern is that the final analytic cohort appears severely distorted and unrepresentative (mean age $\sim$ **12.5** years; maximum age 26), which strongly suggests a processing/filtering/typing error or inadvertent subsetting rather than a defensible population restriction (Sec. 2.1–2.2, Sec. 3.1–3.2). The classification results report near-perfect discrimination (AUC-ROC $\sim$ **0.99**), but this is largely driven by leakage/tautological features derived from the target POA variable (Sec. 2.4.2, Sec. 3.3). The regression models yield counterintuitive associations (POA='N' linked to shorter LOS/lower charges), but diagnostics show extreme multicollinearity, heteroscedasticity, and non-normal residuals, and interpretation is further complicated by unclear baseline/reference coding and by the conceptual ambiguity of a principal diagnosis marked POA='N' (Sec. 2.5, Sec. 3.4.2–3.4.3).

The manuscript's strongest asset is its transparency about pipeline failure and analytic limitations. To become a publishable contribution, it likely needs to (i) either repair the multi-diagnosis extraction to answer the original question, or (ii) decisively reframe as a rigorous, reproducible case study on administrative-data processing failure modes and validity checks, with a clear audit trail, validated cohort construction, and appropriately scoped conclusions (Abstract, Sec. 1, Sec. 3.2, Sec. 4).

Strengths

- Unusually candid reporting that the original multi-diagnosis POA research objective could not be completed, and that downstream analyses are constrained and likely not substantively interpretable (Abstract, Sec. 3.2, Sec. 4).
- Clear high-level description of the THCIC source and the distinction between intended vs. realized feature availability (Sec. 2.1–2.2).

- Appropriate initial use of log transforms for skewed LOS/charges and inclusion of regression diagnostics (residual plots, Q–Q plots, VIF discussion), alongside frank acknowledgement of assumption violations (Sec. 2.5, Sec. 3.4.3).
- Model performance is reported with ROC and precision–recall metrics, and the manuscript recognizes that extremely high AUCs are driven by POA-derived features (leakage-like design), which is an important lesson for applied ML (Sec. 2.4.2, Sec. 3.3).
- Figures are generally well-organized and communicate cohort characteristics and diagnostic concerns, even though several require clearer labeling and denominators.
- The paper surfaces a potentially valuable meta-contribution: robust, audited preprocessing pipelines are prerequisites for valid inference in large administrative health datasets (Sec. 1, Sec. 4).

Major issues

1. **Cohort validity is not established; the processed analytic cohort is implausibly truncated to ages ≤ 26 (mean ~ 12.5), strongly indicating a severe processing/filtering/merge/typing error or inadvertent subsetting. This is not merely a generalizability limitation: it undermines interpretability of *all* downstream ML and regression results because it is unclear what population is being analyzed (Sec. 2.1–2.2, Sec. 3.1–3.2).**

   Recommendation: Add a formal data audit and cohort derivation trace. Provide row counts and key summaries (including age distribution) at each pipeline step from the raw 3,110,296 discharges to each analysis dataset (descriptives, ML, LOS regression, charges regression). Explicitly diagnose why max age becomes 26: verify PAT_AGE field type/encoding (numeric vs string), any recoding (e.g., '100+' handling), any accidental filters/joins, and whether a pediatric-only facility/service subset was inadvertently selected. Report pre-/post-processing age histograms and summary stats. If the cause cannot be resolved, state that clearly and treat the remaining analyses as non-interpretable beyond illustrating failure modes.

2. **The manuscript's framing still largely reflects the unrealized multi-diagnosis/POA network objective, risking a mismatch between what readers expect and what was actually accomplished. As written, it reads more like a postmortem than a completed substantive POA/utilization study, and the main contribution is not consistently articulated (Abstract, Sec. 1, Sec. 3.2, Sec. 4).**

   Recommendation: Choose and execute one coherent contribution: (A) repair extraction/linkage of OTH_DIAG_CODE_ *with POA_OTH_DIAG_* and deliver the originally promised multi-diagnosis analyses; or (B) explicitly reposition as a methodological/data-engineering case study on THCIC POA usability and pipeline failure modes.

If (B), rewrite Abstract/Introduction/Conclusions to foreground: what failed, why it matters, what checks would have prevented it, and what validated subset/analyses (if any) remain meaningful.

3. **The data-processing/linkage failure description is too vague to be reproducible or actionable, despite being central to the paper. It is unclear how OTH_DIAG_CODE_ *and* *POA_OTH_DIAG_* were parsed and reshaped, what exact inconsistencies/errors occurred, and which step induced cohort distortion and/or "No transactions to process for ARM" (Sec. 2.1–2.2, Sec. 3.2).**

   Recommendation: Expand Sec. 2.1–2.2 and Sec. 3.2 into a step-by-step pipeline specification: file format(s), column naming patterns, wide-to-long logic, how diagnosis codes were matched to POA flags, memory/chunking strategy, schema validation, and exact failure conditions (including representative error messages and counts of malformed/missing pairs). Add a flow diagram and a pipeline log table (step name → rows retained → key exclusions). Provide code/pseudocode in an appendix and list software + versions.

4. **The classification task is methodologically compromised by direct/near-direct label leakage: engineered features such as num_POA_YUW_conditions are deterministic or near-deterministic functions of the target POA_PRINC_DIAG under the realized principal-only setup, inflating AUC-ROC (~ 0.99) and rendering results uninformative about independent predictors (Sec. 2.4.2, Sec. 3.3; also reflected in feature-importance Figures 8, 10–12, 15).**

   Recommendation: Re-run and report models with a leakage-free feature set (e.g., demographics + admission characteristics + principal diagnosis representation that is not derived from POA_PRINC_DIAG). Provide side-by-side performance with/without leakage features (AUC-ROC, PR-AUC, sensitivity/recall at a specified threshold, specificity, calibration if possible). Clearly document train/test split and/or cross-validation, hyperparameters, encoding of categorical variables (especially PRINC_DIAG_CODE), and any imbalance handling (class weights or resampling). Recompute feature importance only for leakage-free models and state the importance method.

5. **Conceptual validity of interpreting "principal diagnosis POA='N'" as a hospital-acquired principal diagnosis is not established. In many coding frameworks, principal diagnosis is the condition chiefly responsible for admission, making POA='N' atypical and potentially reflective of coding/sequencing rules, documentation quirks, or artifacts. Without validation, downstream comparisons in LOS/charges may not correspond to a clinically meaningful exposure (Sec. 1, Sec. 3.3–3.4, Sec. 4).**

Recommendation: Add a focused validity check: (i) report the prevalence of POA_PRINC_DIAG='N' and compare to expectations/documentation (THCIC/POA guidance); (ii) tabulate the top principal ICD-10 codes (or grouped categories such as CCS/MDC) separately for POA='N' vs POA='Y/U/W', and provide face-validity discussion; (iii) consider restricting analyses to POA in {Y,N} with clear definitions or explicitly treat principal-POA='N' as a coding artifact outcome rather than "hospital-acquired condition".

6. **The OLS regression strategy for log(LOS) and log(charges) is not defensible as currently presented given (a) unclear POA parameterization/baseline group, (b) extreme multicollinearity (VIFs reported > 1000), (c) heteroscedasticity and non-normal residuals, and (d) minimal severity adjustment due to missing secondary diagnoses—yet coefficients are still discussed as associations (Sec. 2.5, Sec. 3.4.2–3.4.3).**

   Recommendation: Make inference claims conditional on corrected cohort + clarified coding, and then either (1) redesign the regression for stability: restrict sample to meaningful POA categories with POA='Y' as reference; collapse sparse categories; use robust (HC3) or cluster-robust SEs (hospital-level if available); consider penalized regression (ridge) for high-dimensional categorical adjustment; and/or run GLM sensitivity analyses (Gamma-log for charges; count/overdispersed models or alternative LOS modeling) OR (2) explicitly downgrade regressions to exploratory diagnostics and avoid interpreting effect sizes/p-values. Provide a VIF table (top offenders), fit statistics, and (if used) formal heteroscedasticity tests.

7. **The relationship between the full THCIC dataset summaries and the restricted analytic cohort is easy to misread; some descriptive statistics appear to refer to the full dataset while modeling pertains to the truncated cohort, risking incorrect generalization (Sec. 3.1 vs Sec. 3.3–3.4).**

   Recommendation: Separate results by dataset explicitly. Add a table listing, for each figure/model, the exact N and key demographics (age, LOS, charges) of the subset used. When reporting LOS/charges descriptives, provide them separately for the full dataset and the analytic cohort, and avoid interpreting full-dataset distributions as if they describe the model-estimation sample.

Minor issues

1. Feature definitions and missing-data handling are under-specified (e.g., median age imputation is mentioned in Results but not clearly defined/justified in Methods; missingness rates are not quantified; handling of POA codes E/1/blank/'-' is inconsistent across analyses) (Sec. 2.2, Sec. 3.1, Sec. 2.5.2, Sec. 3.4.2).

   Recommendation: Create a dedicated preprocessing/feature-engineering subsection (Sec. 2.2–2.5): quantify missingness for key variables; state exactly when/why median imputation was applied and whether complete-case sensitivity checks were run; define

POA grouping rules once (Y/U/W vs N vs E/1/missing), report counts per category, and keep definitions consistent across ML and regressions.

2. Inconsistent POA coding/parameterization between Methods and Results: Methods describe a binary indicator (N vs Y/U/W), while Results describe two indicators with 'other/missing' baseline; narrative comparisons sometimes assume N vs Y/U/W directly, which is not what the reported coefficients represent (Sec. 2.5.2, Sec. 3.4.2; also affects interpretation throughout Results).

   Recommendation: Write one explicit regression equation and define the reference category. If using a 3-level scheme, report the explicit contrast ($\beta_N - \beta_{YUW}$) with SE/CI; otherwise refit a binary model with POA='Y/U/W' as reference to match the narrative.

3. Inconsistent definition/notation for log transforms: Methods mention offsets (LOS:+1, charges:+0.01) and plots show $\mathbf{log(LOS + 1)}$, but regression sections label outcomes as $\mathbf{ln(LOS)}$ and $\mathbf{ln(Charges)}$ without offsets (Sec. 2.5, Sec. 3.4; figure captions).

   Recommendation: Standardize transformations across text, tables, and figures (e.g., $y_{LOS} = \mathbf{ln(LOS + 1)}$, $y_{Chg} = \mathbf{ln(Charges + 0.01)}$). If different transforms were used for different components, state and justify explicitly.

4. Modeling details needed for reproducibility are incomplete: train/test split, cross-validation, hyperparameters, categorical encoding (especially high-cardinality PRINC_DIAG_CODE), and whether hospital clustering/fixed effects were considered are not clearly stated (Sec. 2.4–2.5, Sec. 3.3–3.4).

   Recommendation: Add a reproducibility checklist in Methods: data split protocol, CV strategy, hyperparameter settings/tuning approach, encoding strategy (e.g., CCS/MDC grouping vs one-hot top-K), and SE choice (classical vs robust vs clustered). List software packages and versions.

5. Figures have multiple ambiguities in denominators, axis units, and category labeling (including 'nan'/'-' categories and unclear ordering), reducing standalone interpretability (e.g., Figures 1–6, 17–21 as cited in the structured report).

   Recommendation: For each figure: state total N and subgroup n, define denominators (percent of what), add units (days, dollars), remove/rename 'nan' to 'missing', and ensure consistent category ordering and readable labels. Where space is limited, move detailed plots to an appendix and keep the most interpretable summaries in the main text.

6. Engineered feature naming implies multi-diagnosis counts ('num_POA_YUW_conditions') even though only the principal diagnosis POA is reliably available, which is confusing and contributes to the leakage problem (Sec. 2.4.2, Sec. 3.3).

Recommendation: Rename/redefine engineered features to reflect realized meaning (e.g., POA_is_YUW, POA_is_N) and include a short mapping table showing values by POA code. Remove language implying counts of multiple diagnoses unless secondary diagnoses are truly incorporated.

7. Uncertainty is not reported for key quantities (AUC/PR-AUC, mean LOS/charges by POA strata, regression effects), limiting readers' ability to assess precision—especially with a highly filtered cohort (Sec. 3.3–3.4.1).

   Recommendation: Add 95% confidence intervals (analytic or bootstrap) for primary performance metrics and key group summaries; report sample sizes alongside all reported means/medians. If compute is constrained, at minimum provide CIs for the headline ML metrics and the main POA coefficient(s).

8. Background/related work on POA coding practices, known misclassification/upcoding issues, and prior work linking POA/HACs with LOS/cost is sparse relative to the topic's complexity (Sec. 1–2).

   Recommendation: Add a short Related Work subsection summarizing: POA coding rules and known pitfalls; typical uses of POA in outcomes research; and why multi-diagnosis POA linkage is methodologically challenging. Use this to justify why a pipeline-audit/case-study contribution is valuable if the original aims cannot be met.

Very minor issues

1. Authorship/affiliation string in the unstructured description ("Anthropic, Gemini & OpenAI servers. Planet Earth.") is not appropriate for a scientific manuscript.

   Recommendation: Replace with a real institutional affiliation (or omit if not available) consistent with the journal's authorship policies.

2. Keywords appear misaligned with the realized scope (e.g., 'GPU computing', 'Outlier detection', 'Cross-validation' if not actually used) (Abstract Keywords, Sec. 2.6).

   Recommendation: Update keywords to match the conducted work (POA, administrative data, data preprocessing, leakage, regression diagnostics, class imbalance) and remove terms not substantively supported.

3. Terminology is occasionally internally contradictory (e.g., describing prediction of "present on admission (POA)='N'") and POA code notation varies ('N' vs formatted variants) (Sec. 3.3; captions).

   Recommendation: Standardize wording to "POA='N' (not present on admission)" and use consistent plain-code notation ('Y','N','U','W','E','1', missing) throughout.

4. Formatting/typographic inconsistencies (heading levels, spacing, truncated labels, OCR-like artifacts) reduce readability (Sec. 2.5–2.6, Sec. 3–4; multiple figures).

Recommendation: Perform a full style pass: consistent heading hierarchy, consistent math/variable notation, improved figure export (vector/high-res), and label readability (font sizes, rotation/wrapping).

Key statements and references

- • **Hospital-acquired conditions are typically associated with increased resource utilization, including longer lengths of stay and higher costs, as reported in prior literature on inpatient complications and patient safety.**
- *Reference(s):* 11
- • **The general expectation in the health services research literature is that conditions coded as not present on admission (hospital-acquired) are linked to worse outcomes and greater resource use than conditions present on admission, based on prior studies of POA indicators and hospital-acquired complications.**
- *Reference(s):* 11
- • **Previous work using Present on Admission indicators has established them as important tools for distinguishing pre-existing conditions from hospital-acquired complications in administrative data, enabling more accurate risk adjustment and quality measurement.**
- *Reference(s):* 11
- • **Prior studies of hospital-acquired conditions and complications have consistently found that such events are associated with increased length of stay, higher charges or costs, and sometimes increased mortality, forming the empirical basis for regarding hospital-acquired conditions as markers of excess resource utilization.**
- *Reference(s):* 11

Mathematical consistency audit

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

Maths relevance: light

The paper contains limited formal mathematics (no explicit numbered equations). The main analytic elements are (i) definitions of binary/grouped POA variables, (ii) log-transformations of skewed outcomes with small offsets, and (iii) linear (OLS) and classification model specifications described in prose. The most important internal-consistency issues are mismatches in how the POA predictor is coded (binary vs two dummies with a third baseline) and ambiguous/variable definitions of the log-transformed outcomes ($\ln(x)$ vs $\ln(x+c)$).

Checked items

1. ✔ **Log-transform with offsets (definition)** (Sec. 2.2, Resource Utilization Outcomes, p.3)
    - **Claim:** LOS and Charges are log-transformed using the natural log, after adding constants (LOS + 1; Charges + 0.01) to handle potential zeros.
    - **Checks:** definition consistency, domain validity
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** LOS and Charges are nonnegative; LOS minimum stated as 1 day, Natural logarithm is used
    - **Notes:** As a standalone definition, $\ln(\text{LOS}+1)$ and $\ln(\text{Charges}+0.01)$ are well-defined for nonnegative LOS/Charges. However, later sections use inconsistent notation for what was actually used in regressions (see separate items).
2. ✖ **Dependent variable naming vs transformation** (Sec. 2.5.1, p.4; Sec. 3.4 and Figs. 16–17 captions, pp.8–9)
    - **Claim:** Dependent variables are $\log_e \text{LOS}$ and $\log_e \text{Charges}$ in regression; figures show $\log(\text{LOS}+1)$.
    - **Checks:** notation consistency, definition consistency
    - **Verdict:** FAIL; confidence: high; impact: critical
    - **Assumptions/inputs:** $\log_e \text{LOS}$ denotes $\ln(\text{LOS})$ unless otherwise specified, Figure captions reflect the transformation actually plotted
    - **Notes:** Methods specify adding constants prior to logging, and Fig. 16 explicitly uses $\log(\text{Length of Stay}+1)$. Regression text labels outcomes as $\log_e \text{LOS}$/ $\log_e \text{Charges}$ without the offsets, creating ambiguity/inconsistency about the modeled dependent variables.
3. ✔ **Binary POA grouping for main comparison** (Sec. 2.2, POA_PRINC_DIAG definition, p.2)
    - **Claim:** POA is categorized into 'N' vs 'Y/U/W' for analysis; other codes ('E','1', missing) are generally excluded from N vs Y/U/W comparisons.
    - **Checks:** definition consistency
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** POA codes are categorical and mutually exclusive
    - **Notes:** The grouping rule is clear and internally coherent as written.
4. ✔ **Classification target definition** (Sec. 2.4.1, p.3)
    - **Claim:** Target is 1 if POA_PRINC_DIAG='N', 0 if POA_PRINC_DIAG in {'Y','U','W'}, excluding other statuses.
    - **Checks:** definition consistency
    - **Verdict:** PASS; confidence: high; impact: minor

- **Assumptions/inputs:** Records with POA in {'E','1', missing, '-'} are excluded
- **Notes:** Binary target definition is unambiguous.

5. ✔ **PR-AUC baseline vs prevalence consistency** (Sec. 3.1 (prevalence 1.63%), p.6; Sec. 3.3 (baseline PR-AUC 0.016), p.6)
    - **Claim:** Baseline PR-AUC is approximately 0.016, consistent with a positive class prevalence of $\sim \mathbf{1.63\%}$.
    - **Checks:** sanity/limiting case
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** Baseline PR-AUC for a no-skill classifier equals positive prevalence
    - **Notes:** Analytically, baseline PR-AUC aligning with prevalence is consistent; the numeric values are close (**0.016** vs **0.0163**) but this audit does not validate the empirical prevalence.

6. ✔ **Engineered POA count features are principal-only** (Sec. 3.1, pp.5–6 (Figure 7 discussion))
    - **Claim:** num_POA_YUW_conditions and num_POA_N_conditions, derived from principal diagnosis POA, are predominantly 0/1 and effectively represent principal POA status.
    - **Checks:** definition consistency, logical implication
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Only principal POA is available at patient level, Count features are computed from available diagnoses
    - **Notes:** If only one diagnosis POA is available, these 'counts' must be in $\mathbf{0, 1}$ (or 0 when missing/exempt), matching the paper's description.

7. ✔ **Feature leakage description (tautological predictor)** (Sec. 3.3, p.6–7)
    - **Claim:** High AUC-ROC is heavily influenced by inclusion of features derived directly from principal POA status (e.g., num_POA_YUW_conditions), creating a near-definitional relationship with the target.
    - **Checks:** logical implication, definition dependence
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** num_POA_YUW_conditions is deterministically computed from POA_PRINC_DIAG, Modeling subset excludes POA codes outside {'N','Y','U','W'}
    - **Notes:** Given the target definition, num_POA_YUW_conditions is (essentially) 1 when target=0 and 0 when target=1, so it can trivially separate classes. The paper's explanation is internally consistent.

8. ✖ **Regression POA predictor coding (Methods vs Results)** (Sec. 2.5.2, p.4 vs Sec. 3.4.2, p.9)
    - **Claim:** Methods: single binary indicator (N vs Y/U/W). Results: two indicators (Y/U/W and N) with baseline other/missing.
    - **Checks:** model specification consistency, reference category consistency
    - **Verdict:** FAIL; confidence: high; impact: critical
    - **Assumptions/inputs:** OLS includes an intercept unless otherwise stated
    - **Notes:** These are different encodings. With an intercept, a two-dummy scheme implies a third reference category; a single binary dummy implies one reference category. The paper does not reconcile which was actually used, so coefficient interpretations reported in Results cannot be verified against Methods.
9. ⚠ **Interpretation of POA coefficients as N vs Y/U/W contrast** (Sec. 3.4.2, p.9)
    - **Claim:** Regression results imply POA='N' is associated with shorter LOS and lower charges compared to POA='Y/U/W'.
    - **Checks:** logical implication, missing-steps verification
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** If using two dummies with the same baseline, N vs Y/U/W comparison is a linear contrast of coefficients, If using one binary dummy, coefficient directly compares N to Y/U/W
    - **Notes:** If the Results' two-indicator specification is correct, the N vs Y/U/W comparison should be reported as $(\beta_N - \beta_{YUW})$ with uncertainty. If the Methods' binary specification is correct, only one coefficient should exist. Without the explicit model equation or contrast output, the stated comparative interpretation is not fully auditable.
10. ✔ **Interaction term degeneracy under mutually exclusive dummies** (End of Sec. 3.4.2, p.9)
    - **Claim:** An interaction term between the two POA indicators has coefficient 0 and undefined t-statistic due to multicollinearity arising from derivation from a single POA status.
    - **Checks:** algebraic sanity check
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** POA='N' and POA='Y/U/W' dummies cannot both be 1 for the same record
    - **Notes:** If d_N and d_{YUW} are mutually exclusive, then $d_N \times d_{YUW}$ is identically 0, making the regressor non-informative and non-estimable; reporting an undefined t-statistic is consistent with this degeneracy.
11. ✔ **Use of 'vast majority' for mutual exclusivity** (End of Sec. 3.4.2, p.9)

- **Claim:** If one POA indicator is 1, the other is 0 in the vast majority of cases.
- **Checks:** logical consistency
- **Verdict:** PASS; confidence: medium; impact: minor
- **Assumptions/inputs:** POA statuses are single-valued per principal diagnosis
- **Notes:** Strictly, mutual exclusivity should hold for all cases where POA is in $N, Y/U/W$; cases with missing/exempt may make both indicators 0, but not both 1. The phrase 'vast majority' is imprecise but not mathematically contradictory.

12. ✖ **Semantic inconsistency in calibration figure description** (Figure 15 caption/text, p.8)
    - **Claim:** Logistic regression calibration curve is described as predicting whether the principal diagnosis was 'present on admission (POA)="N"'.
    - **Checks:** notation/definition consistency
    - **Verdict:** FAIL; confidence: high; impact: minor
    - **Assumptions/inputs:** POA='N' means not present on admission
    - **Notes:** The caption conflates 'present on admission' with code 'N' (not present). This is a wording/notation error that can confuse interpretation but does not by itself alter derivations (none are shown).

Limitations

- The PDF contains almost no explicit mathematical equations; most model specifications are described in prose, which limits the ability to verify algebraic derivations.
- No explicit regression equations/design-matrix coding tables are provided; therefore, coefficient interpretation consistency can only be checked against textual descriptions, leading to UNCERTAIN verdicts where the exact specification is not recoverable.
- This audit does not validate reported metric values (AUC, PR-AUC, coefficients, R-squared) numerically; it only checks definitional/logical consistency.

Numerical results audit

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

Fifteen text-based numeric statements were checked using parsing, logical bounds/order checks, and simple derived relationships. All checks passed; no inconsistencies were detected within the validated scope.

Checked items

1. ✔ **C1_total_records_2018** (Page 4 / Results 3.1 (first paragraph))

- **Claim:** The initial dataset comprised 3,110,296 inpatient discharge records from Texas for the year 2018.
- **Checks:** integer_format_sanity
- **Verdict:** PASS
- **Notes:** Raw parses cleanly to a positive integer.

2. ✔ **C2_age_stats_internal_consistency** (Page 4 / Results 3.1 (age summary statistics paragraph))
    - **Claim:** Age summary: mean 12.47 years (SD 6.97), median 14.0, IQR 8 to 18, maximum 26 years.
    - **Checks:** order_and_range_checks
    - **Verdict:** PASS
    - **Notes:** Logical consistency checks on order/range.
3. ✔ **C3_los_summary_stats_consistency** (Page 5 / Results 3.1 (resource utilization outcomes paragraph))
    - **Claim:** For the full dataset, mean LOS was 5.01 days (median 3) with SD 15.91 days.
    - **Checks:** basic_stats_sanity
    - **Verdict:** PASS
    - **Notes:** Sanity checks passed. mean≥median holds.
4. ✔ **C4_charges_summary_stats_consistency** (Page 5 / Results 3.1 (resource utilization outcomes paragraph))
    - **Claim:** Mean Total Charges were \$63,730 (median \$20,545) with SD \$184,380.
    - **Checks:** basic_stats_sanity
    - **Verdict:** PASS
    - **Notes:** Sanity checks passed. mean≥median holds.
5. ✔ **C5_skewness_kurtosis_nonnegative_check** (Page 5 / Results 3.1 (resource utilization outcomes paragraph))
    - **Claim:** Skewness LOS: 124.68, Charges: 21.03; Kurtosis LOS: 27578, Charges: 1247.
    - **Checks:** sign_and_magnitude_sanity
    - **Verdict:** PASS
    - **Notes:** All values parsed and are positive.
6. ✔ **C6_added_constants_for_log_transform** (Page 3 / Methods 2.2 (Resource Utilization Outcomes bullet))
    - **Claim:** Added constants before log transform: 1 for Length of Stay; 0.01 for Total Charges.

- **Checks:** constant_extraction_and_use_consistency
- **Verdict:** PASS
- **Notes:** Both constants are positive.

7. ✔ **C7_pr_auc_baseline_vs_prevalence** (Page 6 / Results 3.3 (PR curves paragraph) + Page 6 (minority class prevalence))
    - **Claim:** Minority class prevalence is 1.63% of records; baseline PR-AUC is 0.016.
    - **Checks:** baseline_metric_equals_prevalence
    - **Verdict:** PASS
    - **Notes:** Compared baseline PR-AUC to prevalence fraction; difference is consistent with rounding ($\mathbf{0.0163}$ vs $\mathbf{0.016}$).
8. ✔ **C8_auc_roc_ordering_check** (Page 6 / Results 3.3 (model metrics bullets))
    - **Claim:** AUC-ROC: Logistic Regression 0.994; Random Forest 0.995; LightGBM 0.996.
    - **Checks:** range_and_ranking_sanity
    - **Verdict:** PASS
    - **Notes:** All AUC values are within $[\mathbf{0}, \mathbf{1}]$ and ordered as stated.
9. ✔ **C9_pr_auc_ordering_check** (Page 6 / Results 3.3 (model metrics bullets))
    - **Claim:** PR-AUC: Logistic Regression 0.632; Random Forest 0.715; LightGBM 0.737.
    - **Checks:** range_and_ranking_sanity
    - **Verdict:** PASS
    - **Notes:** All PR-AUC values are within $[\mathbf{0}, \mathbf{1}]$ and ordered as stated.
10. ✔ **C10_f1_score_range_check** (Page 6 / Results 3.3 (model metrics bullets))
    - **Claim:** F1-scores: Logistic Regression 0.532; Random Forest 0.711; LightGBM 0.678.
    - **Checks:** metric_range_sanity
    - **Verdict:** PASS
    - **Notes:** All F1 values are within $[\mathbf{0}, \mathbf{1}]$.
11. ✔ **C11_avg_unique_principal_dx_leq_1** (Page 6 / Results 3.1 (engineered features paragraph))
    - **Claim:** The average number of total unique (principal) diagnoses per patient was 0.817.
    - **Checks:** bounded_mean_check
    - **Verdict:** PASS
    - **Notes:** Value satisfies the stated logical bound $\mathbf{0} \leq \boldsymbol{x} \leq \mathbf{1}$ for this definition.

12. ✔ **C12_loge_los_coeff_ordering** (Page 9 / Results 3.4.2 (Regression model results, loge LOS))
    - **Claim:** Coefficients for $\log_e$ **LOS**: POA='Y/U/W' = -0.4010; POA='N' = -0.6044.
    - **Checks:** numeric_comparison
    - **Verdict:** PASS
    - **Notes:** Ordering holds: $-0.6044 < -0.4010$.
13. ✔ **C13_loge_charges_coeff_signs** (Page 9 / Results 3.4.2 (Regression model results, loge Charges))
    - **Claim:** Coefficients for $\log_e$ **Charges**: POA='Y/U/W' = 0.4262; POA='N' = -0.1349.
    - **Checks:** sign_check
    - **Verdict:** PASS
    - **Notes:** Signs match the statement (positive for Y/U/W; negative for N).
14. ✔ **C14_r2_bounds** (Page 9 / Results 3.4.2 (Regression model results))
    - **Claim:** R-squared: 0.165 for $\log_e$ **LOS** model; 0.363 for $\log_e$ **Charges** model.
    - **Checks:** range_check
    - **Verdict:** PASS
    - **Notes:** Both R-squared values are within $[0, 1]$.
15. ✔ **C15_implied_percent_from_baseline_pr_auc** (Page 6 / Results 3.3 (PR curves paragraph))
    - **Claim:** Baseline PR-AUC is 0.016 (interpretable as $\sim 1.6\%$).
    - **Checks:** unit_conversion
    - **Verdict:** PASS
    - **Notes:** $0.016 \times 100 = 1.6\%$, matching the stated interpretation.

Limitations

- Only the provided PDF text was used; figures were not digitized and numeric values embedded solely in plots were not extracted.
- Many statistics (means/SDs, model metrics, regression coefficients) cannot be independently recomputed from the PDF because underlying data and intermediate tabulations are not provided.
- Checks are limited to parsing, logical bounds, and simple relationships (e.g., prevalence-to-baseline PR-AUC) that can be validated from explicit numbers in the text.