Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

DOI: 10.37943/JXDB7524

Valeriya Kunichik
Master's Degree, Department of Software
kunichikval@gmail.com, orcid.org/0009-0004-8580-7731
Akhmet Baitursynuly Kostanay Regional University, Kazakhstan
Olga Salykova
Candidate of Technical Sciences, Associate Professor, Department of Software
salykova0603@gmail.com, orcid.org/0000-0002-8681-4552
Akhmet Baitursynuly Kostanay Regional University, Kazakhstan

EXPLORATORY DATA ANALYSIS AS A FOUNDATION FOR TRUSTWORTHY CLINICAL MACHINE LEARNING

Abstract: Exploratory data analysis is frequently treated as a preliminary or purely descriptive step in clinical machine learning studies. However, many methodological failures of clinical prediction models arise not from algorithm selection, but from unrecognized properties of the data-generating process, including temporal non-stationarity, non-random missingness, and systematic documentation effects. These issues are particularly critical in early clinical triage settings, where risk assessment must be performed at hospital admission using limited and heterogeneous clinical information. In this study, we examine exploratory data analysis as a dedicated methodological stage that precedes and constrains downstream machine learning design in early triage–level mortality risk prediction. Using a large-scale national registry of hospitalized patients with confirmed COVID-19 infection, we conduct a structured exploratory analysis of admission-level clinical data, focusing on temporal drift in outcome prevalence and patient case-mix, feature-level missingness mechanisms, documentation bias in symptom recording, and structural redundancy among clinical indicators. The results demonstrate pronounced temporal non-stationarity across pandemic phases, indicating that commonly used random data splitting strategies may introduce methodological bias even prior to model development. Missingness patterns are shown to be structured and outcome-dependent rather than random, reflecting both clinical severity and organizational processes. In addition, several apparent univariate associations between symptoms and outcomes are attributable to documentation bias rather than underlying clinical effects. These findings provide concrete diagnostic evidence that exploratory analysis directly constrains valid choices related to validation design, handling and representation of missing information, feature specification, and interpretation boundaries. The findings suggest that exploratory data analysis should be regarded as a fundamental methodological stage supporting reliable and trustworthy clinical machine learning in early-stage decision-making settings.

Keywords: exploratory data analysis; early clinical triage; temporal drift; missing data mechanisms; documentation bias; data-centric machine learning; methodological reliability

Introduction

Recent advances in machine learning and data-driven analytics have accelerated the adoption of predictive models in healthcare, particularly for risk stratification and clinical decision support [1]. The growing availability of electronic health records and large-scale clinical registries has enabled the development of machine learning–based approaches for prognosis, triage, and outcome prediction across a wide range of clinical conditions. In infectious disease settings such as the COVID-19 pandemic, these approaches have been widely explored to support early identification of patients at increased risk of adverse outcomes under severe time and information constraints.

Received: 30.03.2026 Accepted: 17.07.2026 Published: 30.06.2026

DOI: 10.37943/JXDB7524

Despite substantial methodological progress, many clinical machine learning models exhibit limited reliability when deployed in real-world settings [2]. While model overfitting and algorithm selection are often cited as primary challenges, accumulating evidence indicates that a substantial proportion of failures originate earlier in the analytical pipeline, at the stage of data understanding and problem formulation. Clinical datasets are shaped by complex data-generating processes, including temporal non-stationarity, non-random missingness, heterogeneous documentation practices, and shifting patient case-mix, which fundamentally distinguish them from standard benchmark tabular datasets [3].

Exploratory data analysis (EDA) is intended to address these complexities by enabling systematic inspection of data structure, distributions, and potential sources of bias prior to model development. However, in many published clinical machine learning studies, EDA is treated as a secondary or purely descriptive step, often limited to basic summary statistics or illustrative visualizations. As a result, critical data properties that directly affect the validity of model training, validation, and interpretation may remain unrecognized [4]. Temporal drift in outcome prevalence, structured missingness mechanisms, and documentation bias can invalidate common methodological choices, such as random data splitting, naïve imputation strategies, or univariate interpretation of feature effects, even before any predictive model is trained [5].

These limitations are particularly consequential in early clinical triage settings, where risk assessment must be performed at hospital admission using limited, heterogeneous, and frequently incomplete clinical information. Methodological errors introduced at the exploratory stage may propagate downstream, affecting validation design, probability calibration, interpretability, and ultimately clinical decision-making. Treating exploratory data analysis as an informal or optional step therefore undermines the methodological soundness of subsequent machine learning pipelines, regardless of their apparent discriminative performance [6-7].

In this study, we examine exploratory data analysis as a dedicated methodological stage in clinical machine learning rather than as a preliminary visualization exercise. Using a large-scale national registry of hospitalized patients with confirmed COVID-19 infection, we conduct a structured exploratory analysis of early-stage admission data. The analysis focuses on temporal non-stationarity, feature-level missingness mechanisms, documentation bias in symptom recording, and structural redundancy among clinical indicators. The objective of this work is not to develop or evaluate predictive models, but to demonstrate, within the context of early triage-level mortality risk prediction, how disciplined exploratory analysis constrains valid methodological choices in downstream machine learning pipelines. By explicitly linking exploratory findings to implications for validation design, handling of missing information, feature specification, and interpretation boundaries, this study formalizes exploratory data analysis as a diagnostic methodological foundation for reliable clinical machine learning in early-stage decision-making settings.

Methods and Materials

Data source and cohort construction

This study was based on publicly available anonymized data from the Brazilian national SIVEP-Gripe registry, a mandatory surveillance system for hospitalized cases of severe acute respiratory infections that served as the primary national registry for COVID-19 hospitalizations during the pandemic [8]. In-hospital mortality was defined as a binary outcome. The analytical cohort comprised approximately 2.53 million hospital admissions recorded between 2020 and 2022. All records were harmonized into a unified feature schema to ensure consistency across calendar years and to enable systematic longitudinal assessment of data properties. Overall in-hospital mortality prevalence was 0.284, with substantial variation across calendar years, reflecting temporal changes in patient case-mix and pandemic dynamics [9]. Only fully anonymized secondary data were analyzed. No personally identifiable information was accessed or processed. As the analysis was based exclusively on publicly available anonymized data, additional ethical approval and informed consent were not required.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

The feature set was deliberately restricted to admission-level clinical information available at the time of hospital presentation, corresponding to an early clinical triage setting. Included variables comprised basic demographic characteristics, documented symptoms, and recorded comorbidities. Laboratory measurements, imaging findings, and in-hospital treatment information were excluded. This restriction ensures that the exploratory analysis reflects data-generating processes relevant to prospective early decision-making rather than retrospective outcome characterization. Early-triage clinical data are characterized by heterogeneous information completeness, variable documentation practices, and severity-dependent symptom recording. These properties give rise to non-random missingness, documentation bias, and temporal instability, which impose explicit methodological constraints that must be addressed prior to any downstream machine learning design.

Exploratory analysis framework

Exploratory data analysis was conducted as a structured methodological stage preceding any model development and explicitly separated from predictive optimization. The objective was to diagnose properties of the data-generating process that explicitly constrain permissible modeling, validation, and interpretation choices in clinical machine learning.

The framework comprised three complementary components. First, temporal analysis was performed to assess non-stationarity in outcome prevalence, patient characteristics, and feature distributions across calendar years. Second, missingness analysis was conducted at both feature and temporal levels to characterize incomplete documentation and identify non-random and outcome-dependent missingness patterns. Third, structural feature analysis examined correlation patterns, redundancy, and clinically coherent groupings among symptoms and comorbidities. All analyses relied exclusively on summary statistics, stratified visualizations, and distributional comparisons without fitting predictive models or optimizing feature effects. Leakage-safe aggregation was applied for temporal summaries, and no outcome-informed feature selection was performed. The framework was designed to expose systematic data properties that may remain obscured when analysis begins directly with model training.

Temporal perspective and non-stationarity assessment

A temporal perspective was incorporated into the exploratory analysis to explicitly assess non-stationarity inherent in large-scale clinical registries collected over extended periods. The study spanned multiple pandemic phases characterized by evolving epidemiological conditions, clinical practices, and patient case-mix, which necessitated formal evaluation of temporal stability prior to any modeling assumptions [10]. Temporal analysis was conducted by stratifying the cohort by calendar year and aggregating observations at a weekly level. Outcome prevalence, key demographic characteristics, and selected admission-level clinical indicators were examined over time to identify systematic distributional shifts. To visualize temporal dynamics while preserving chronological ordering and avoiding information leakage, expanding means and lagged moving averages were applied within each temporal segment.

To quantify temporal drift, distributional comparisons between predefined calendar periods were performed using the Population Stability Index [11]. PSI was applied to outcome prevalence and selected covariates to assess changes in distribution across years. The purpose of this analysis was not causal attribution, but identification of methodological risks associated with treating temporally evolving data as stationary. Evidence of temporal drift was used to delineate the limitations of random data splitting and to motivate temporally ordered validation strategies in downstream clinical machine learning pipelines.

Missing data mechanisms and documentation effects

Missing data patterns were analyzed as an integral component of the exploratory framework rather than treated as a preprocessing artifact. In early clinical triage data, missingness reflects systematic clinical and organizational processes and therefore cannot be assumed to occur at random [12]. Missingness was examined at both feature and temporal levels to characterize heterogeneity in documentation completeness. Distinct patterns were observed between symptom

DOI: 10.37943/JXDB7524

indicators and comorbidity variables, reflecting differences in recording mechanisms at hospital admission. Consequently, missing values across feature groups may encode different types of information and carry distinct methodological implications.

Automatic imputation and recoding of missing values were deliberately avoided during the exploratory stage to prevent conflation of undocumented information with true clinical absence. Instead, missingness was analyzed as a structured and potentially informative signal, including its association with outcome prevalence and temporal context [13].

Documentation effects were examined as a related source of bias. Apparent univariate associations between selected clinical indicators and outcomes may arise from differential recording practices rather than underlying clinical relationships. Identification of such patterns at the exploratory stage constrains valid univariate interpretation and delimits permissible downstream modeling and interpretability choices.

Interpretation principles and scope of analysis

The exploratory analyses in this study were designed to characterize properties of the data-generating process rather than to estimate predictive effects or infer causal relationships. Findings are interpreted at the level of data structure, documentation patterns, and temporal stability, without attributing observed associations to underlying clinical mechanisms. The scope of the analysis is deliberately restricted to identifying methodological constraints that delimit valid downstream machine learning choices. Exploratory findings inform decisions related to validation design, feature specification, handling of missing information, and interpretation boundaries. No predictive models were fitted, and no performance estimates or risk thresholds were derived within the scope of this study, rather than addressing strategies for mitigating dataset shift or preserving model performance, which have been extensively reviewed elsewhere [14].

By explicitly defining these interpretive boundaries, the analysis positions exploratory data analysis as a diagnostic methodological stage that precedes and constrains model development. This framing aligns the Methods directly with the Results, underscoring the central argument that reliable clinical machine learning begins with disciplined examination of data properties rather than with model training.

Results

Evidence of temporal drift and non-stationarity

Exploratory analysis revealed pronounced temporal non-stationarity in both outcome prevalence and patient characteristics across the study period. As shown in Figure 1, in-hospital mortality varied substantially over calendar time, with elevated levels during early pandemic phases and a sustained decline in later periods. This pattern indicates a non-stationary outcome distribution rather than random temporal fluctuation.

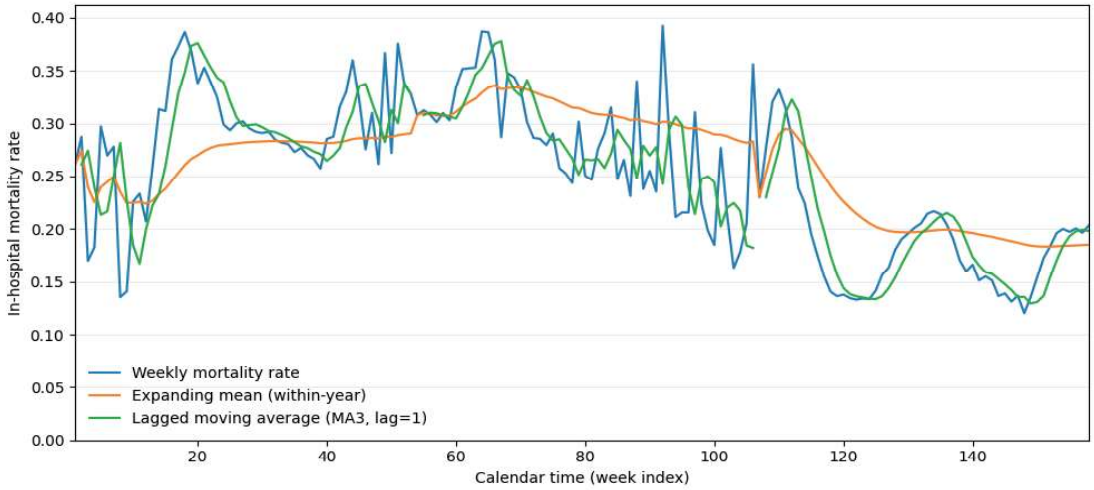

Figure 1. Temporal non-stationarity of in-hospital mortality

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

Weekly aggregation demonstrated structured temporal dynamics rather than short-term noise. Leakage-safe smoothing using expanding means (within-year) and lagged moving averages revealed gradually evolving trends in mortality risk, indicating strong temporal dependence between adjacent periods. Such behavior violates the assumption of identically distributed observations that underlies random data splitting and standard cross-validation strategies.

Temporal drift was not confined to outcomes alone. Changes in patient case-mix were evident across pandemic waves, particularly in the age structure of hospitalized patients. As illustrated in Figure 2, later periods were characterized by a relative shift toward younger age groups and altered distributional shapes compared with earlier years. These shifts reflect systematic changes in the population entering the hospital rather than sampling variability.

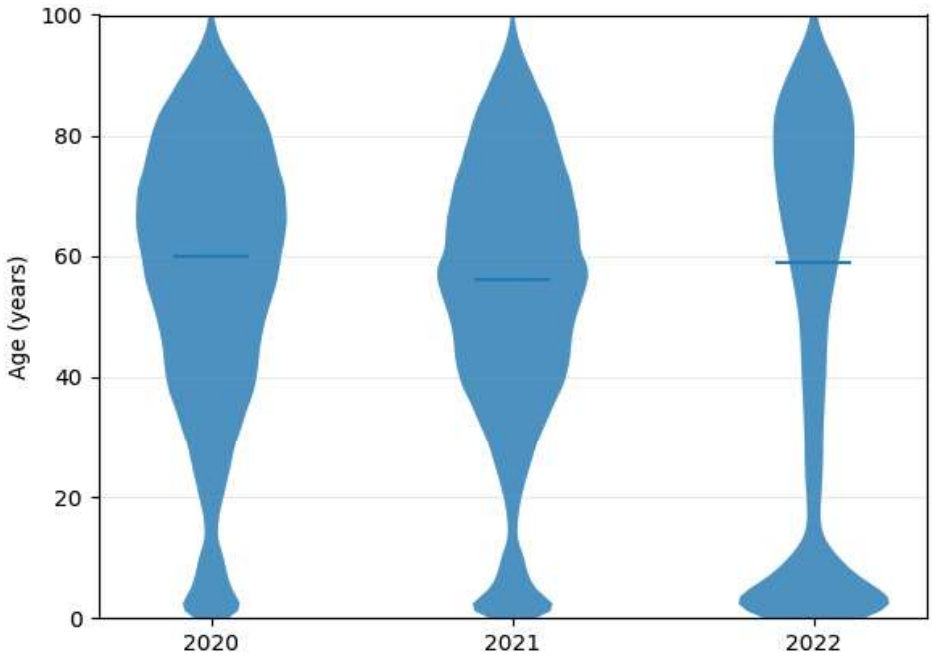

Figure 2. Age distribution across calendar years

Quantitative distributional comparisons between predefined temporal segments confirmed the visual findings. Population Stability Index estimates in Table 1 indicated minimal drift between 2020 and 2021 but moderate to severe shifts between 2021 and 2022, most notably for age (PSI = 0.5792). Outcome prevalence and representative clinical indicators (e.g., dyspnea) also exhibited measurable temporal divergence. Together, these results demonstrate the presence of joint covariate and label drift across calendar years.

Table 1. Population Stability Index based assessment of temporal drift across key variables.

Variable	2020→2021	2021→2022
In-hospital mortality	0.0002	0.0381
Age	0.0328	0.5792
Dyspnea (representative symptom)	0.0057	0.0354

In combination, these findings show that treating the dataset as temporally homogeneous constitutes a methodological error prior to any model development. Observed non-stationarity directly motivates temporally ordered validation strategies and underscores the necessity of explicit temporal diagnostics as a core component of exploratory data analysis in clinical machine learning.

DOI: 10.37943/JXDB7524

Feature-level missingness patterns across time

Exploratory analysis revealed substantial and structured feature-level missingness across the cohort, with clear differences both between feature groups and across calendar years as highlighted in Figure 3. The heatmap demonstrates that missingness is neither uniform across variables nor temporally stable, indicating systematic documentation patterns rather than random data loss.

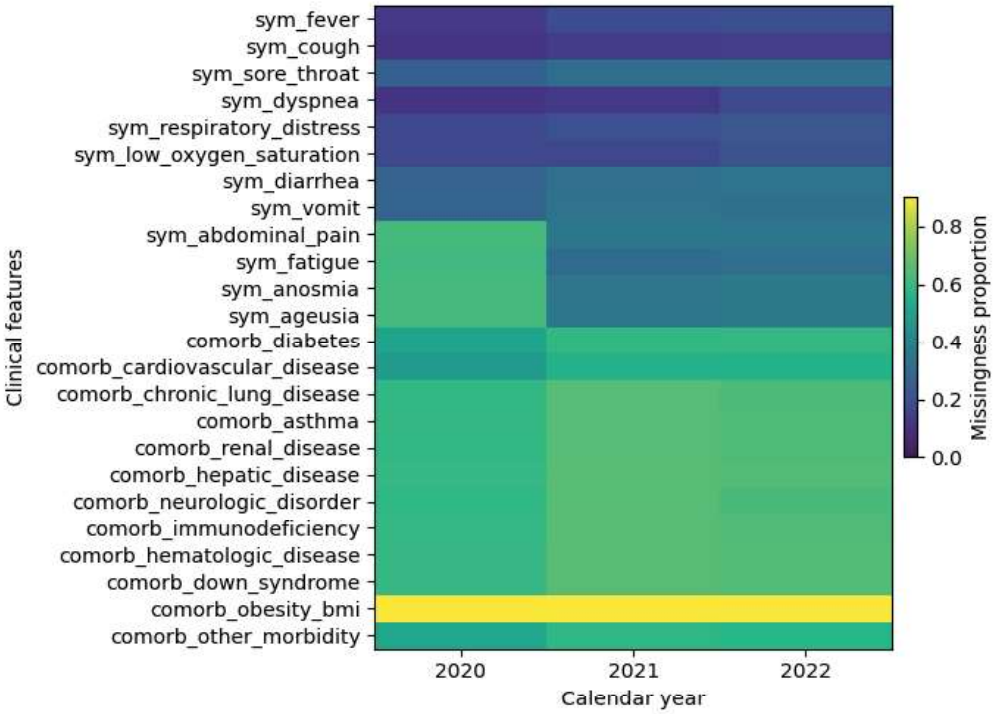

Figure 3. Feature-level missingness across calendar years

Admission-level symptoms exhibit heterogeneous completeness. Core respiratory indicators (fever, cough, dyspnea, respiratory distress, and oxygen desaturation) show consistently low missingness across all years, whereas non-respiratory symptoms, such as fatigue, gastrointestinal complaints, and sensory symptoms display markedly higher and more variable missingness. This contrast is directly visible in Figure 3 and reflects selective documentation at the time of admission rather than stochastic omission.

In contrast, comorbidity variables are characterized by uniformly elevated missingness across years. Most comorbid conditions exhibit missingness levels substantially higher than those of admission symptoms, with limited temporal improvement. Extreme cases, such as obesity defined by body mass index, show near-complete missingness throughout the study period, indicating structural absence of reliable documentation rather than true absence in the patient population. Temporal stratification further shows that missingness patterns evolve over time. For several features, documentation completeness differs between early and later pandemic phases, demonstrating that missingness itself is non-stationary. These shifts co-occur with observed temporal drift in outcomes and patient case-mix, indicating that documentation processes evolve alongside broader system-level changes.

Observed patterns are inconsistent with assumptions of missing completely at random. Features with high clinical relevance often exhibit both high prevalence and high missingness, indicating selective recording. Treating missing values as absence or applying naïve imputation would therefore conflate undocumented information with true clinical negatives.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

Taken together, Figure 3 demonstrates that feature-level missingness in early-stage clinical data constitutes a structured and time-dependent property of the data-generating process. Ignoring this structure introduces methodological risk, particularly for probability calibration and interpretation of feature effects, and necessitates explicit, missingness-aware exploratory analysis prior to model development.

Missingness as an informative signal

Exploratory analysis demonstrated that missingness itself carries clinically and methodologically relevant information. Figure 4 illustrates that in-hospital mortality differs systematically between records with documented values and those with missing entries for several admission-level features, indicating outcome-dependent missingness rather than random data loss.

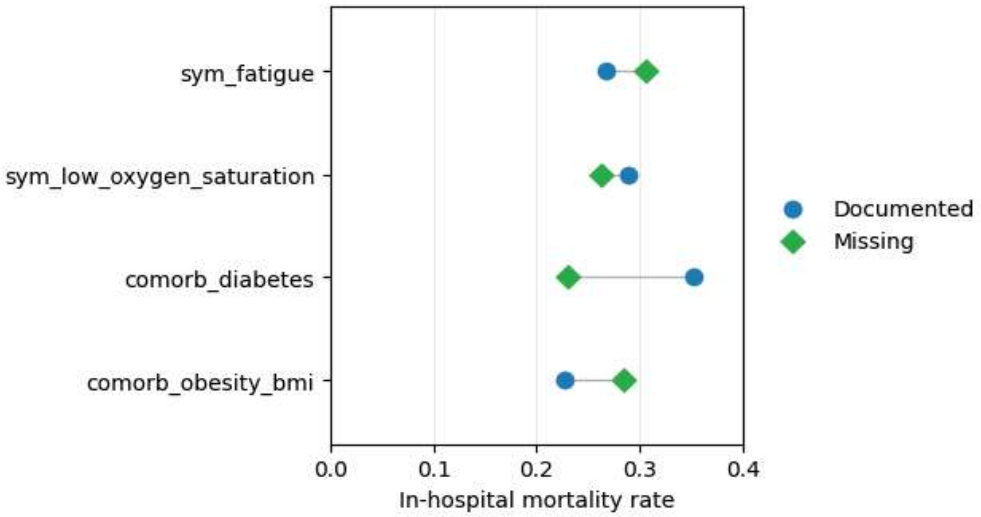

Figure 4. In-hospital mortality stratified by feature documentation status

For all illustrated variables, missing documentation is associated with higher observed mortality compared with documented cases. This pattern is evident for both symptoms and comorbidities, including fatigue, low oxygen saturation, diabetes, and obesity defined by body mass index. The consistent direction and magnitude of these differences indicate that missingness encodes information related to patient severity or clinical context at presentation. Importantly, these contrasts are observable without fitting predictive models or optimizing feature effects. The analysis relies exclusively on stratified outcome summaries, demonstrating that the informational content of missingness is detectable at the exploratory stage. Interpreting missing values as implicit negatives or applying naïve imputation would therefore collapse systematically different risk strata into a single category.

Outcome-dependent missingness has direct methodological implications for probability estimation. When missingness indicators correlate with outcome prevalence, ignoring them introduces systematic bias into predicted risks, particularly in early clinical triage settings where decision thresholds depend on absolute risk values rather than rank-based discrimination.

Overall, Figure 4 shows that missingness in early-stage clinical data functions as a proxy for latent clinical and organizational factors. Explicit representation of missingness structure is therefore necessary to preserve calibration and to delimit valid interpretation of feature effects in downstream clinical machine learning analyses.

Documentation bias in symptom recording

Exploratory analysis revealed systematic documentation bias affecting observed univariate associations between recorded symptoms and in-hospital mortality. Several symptoms commonly associated with viral infection exhibited counterintuitive patterns, whereby documented presence was associated with lower observed mortality compared with undocumented or missing entries.

DOI: 10.37943/JXDB7524

These patterns indicate differential documentation rather than protective clinical effects. Symptom recording at admission is not independent of patient condition: records with more complete symptom profiles are disproportionately represented among lower-risk patients, whereas severe or rapidly deteriorating cases exhibit abbreviated documentation. As a result, documented presence of non-critical symptoms becomes selectively enriched in lower-risk strata.

Documentation bias is further reinforced by structured missingness. Symptoms with low overall documentation frequency frequently show strong divergence in outcome prevalence between documented and undocumented records, demonstrating that undocumented entries cannot be interpreted as symptom absence. Without explicit consideration of documentation mechanisms, univariate symptom effects are unstable and potentially misleading.

Importantly, these effects are detectable at the exploratory stage, prior to any model fitting or feature selection. Ignoring documentation bias propagates structural artifacts into downstream interpretability analyses, constraining the validity of symptom-level conclusions.

Collectively, these findings demonstrate that documentation bias constitutes a structural component of the data-generating process. Explicit accounting for documentation mechanisms is therefore necessary to delimit valid interpretation of feature–outcome associations in clinical machine learning.

Feature structure, redundancy, and clinical grouping

Exploratory structural analysis revealed pronounced redundancy and coherent grouping among admission-level features. As shown in Figure 5, correlation analysis indicates that symptoms and comorbidities form structured feature blocks rather than independent indicators, demonstrating that the input feature space is inherently organized.

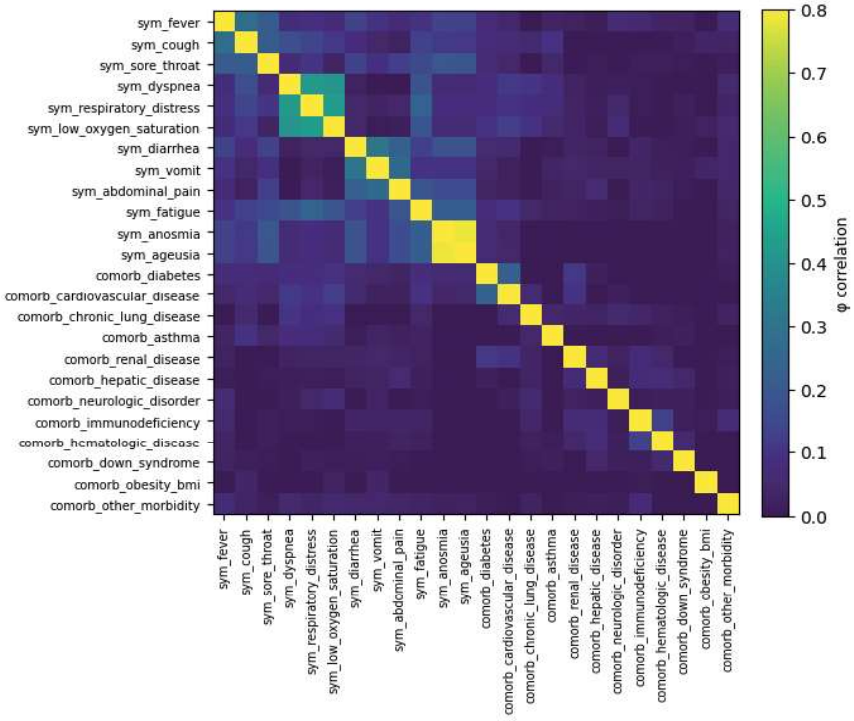

Figure 5. Correlation structure of binary clinical features

A distinct cluster of respiratory-related indicators, including dyspnea, respiratory distress, and low oxygen saturation, exhibits strong internal correlations, forming a tightly coupled block associated with acute disease severity. Non-respiratory symptoms show weaker but non-negligible correlations, consistent with partially overlapping clinical meaning and shared documentation

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

patterns. These block-level structures are directly visible in the correlation matrix as contiguous regions of elevated correlation.

Comorbidity variables also demonstrate structured relationships. Despite higher overall missingness, comorbid conditions form correlated subgroups consistent with known multimorbidity patterns, while exhibiting comparatively weaker correlations with acute symptom clusters. This separation indicates that comorbidities encode a distinct but internally structured dimension of patient risk. These correlated feature blocks are observed consistently across the cohort and reflect a systematic property of the data-generating process. Treating highly correlated indicators as independent features inflates the apparent dimensionality of the input space, amplifies documentation effects, and obscures the interpretation of individual feature contributions.

These structural properties are identifiable at the exploratory stage, prior to any model fitting or feature selection. Figure 5 demonstrates that feature aggregation or grouped representations derived from observed correlations are methodologically warranted to constrain model complexity and preserve interpretability.

Taken together, the observed redundancy and clustering of admission-level features show that structural exploratory analysis is necessary to prevent conflation of correlated documentation artifacts with independent clinical signals in downstream clinical machine learning analyses.

EDA-driven methodological constraints and implications

Exploratory analyses of temporal structure, missingness patterns, documentation processes, and feature relationships demonstrate that early-stage clinical data impose explicit constraints on permissible machine learning design choices. These constraints arise prior to any model training and define the methodological boundaries within which reliable and interpretable analyses can be conducted.

Synthesis of exploratory findings, summarized in Table 2, shows that pronounced temporal drift in outcome prevalence and patient case-mix invalidates assumptions underlying random data splitting and leads to systematically optimistic validation results. Structured and outcome-dependent missingness demonstrates that missing values cannot be treated as random noise or implicit negatives without introducing bias into risk estimation. Documentation effects in symptom recording indicate that univariate feature–outcome associations may be structurally confounded by data collection mechanisms rather than reflecting underlying clinical effects. In addition, redundancy and coherent clustering among admission-level features show that the feature space is inherently structured rather than composed of independent indicators.

Table 2. Exploratory findings and corresponding methodological implications

Exploratory finding	Methodological risk if ignored	Implication for ML analysis
Temporal drift in outcome and case-mix	Optimistic validation and poor generalization	Temporally ordered data splitting
Non-random missingness	Miscalibration and biased risk estimates	Explicit representation of missingness
Outcome-dependent missingness	Distorted interpretation of feature effects	Avoid treating missing values as absence
Documentation effects	Spurious univariate associations	Context-aware interpretation of features
Feature redundancy and clustering	Inflated dimensionality and unstable estimates	Composite or grouped feature representations

Together, these observations indicate that key methodological risks originate at the data level and are detectable before model development through disciplined exploratory analysis. By explicitly linking exploratory findings to concrete methodological constraints, this section demonstrates that EDA functions as a diagnostic safeguard that constrains the space of valid

DOI: 10.37943/JXDB7524

downstream modeling decisions and enables anticipation of methodological errors prior to model training, validation, and interpretation.

Discussion

This study suggests that exploratory data analysis represents a foundational methodological stage in clinical machine learning rather than a preliminary descriptive step. By systematically examining temporal non-stationarity, missingness mechanisms, documentation effects, and feature structure in early-stage clinical data, the analysis shows that multiple sources of methodological risk can be identified prior to any predictive model development. These risks originate at the level of the data-generating process and cannot be mitigated by algorithm choice or model complexity alone [15].

A central finding is the presence of pronounced temporal drift in both outcome prevalence and patient case-mix across pandemic phases. Such non-stationarity directly violates the assumptions underlying random data splitting and standard cross-validation procedures. Importantly, this limitation arises independently of model type: even highly regularized linear models or ensemble-based approaches cannot compensate for validation strategies that ignore temporal structure. The observed drift therefore motivates temporally ordered evaluation as a prerequisite for reliable performance assessment and generalization in real-world clinical settings [16-17]. The results further demonstrate that missing data in early clinical triage contexts cannot be treated as random noise or a purely technical preprocessing issue. Feature-level and outcome-dependent missingness patterns reflect latent clinical severity, organizational constraints, and documentation practices. Treating missing values as implicit negatives or applying naïve imputation strategies risks distorting probability calibration and undermining interpretability. Identifying missingness as an informative signal at the exploratory stage thus provides a principled basis for explicit representation of missingness structure in downstream analyses [18-19].

Documentation effects emerged as a distinct and systematic source of methodological bias. Apparent univariate associations suggesting protective effects of certain symptoms were attributable to differential recording practices rather than underlying clinical mechanisms. In early triage settings, symptom documentation is conditioned by patient stability, clinical workload, and time pressure, leading to selective recording. Without explicit consideration of these processes, univariate feature interpretation and post hoc explainability analyses may produce misleading narratives, thereby eroding trust in deployed clinical machine learning systems [20].

Structural analysis revealed substantial redundancy and clinically coherent grouping among admission-level features, indicating that the feature space is inherently structured rather than composed of independent indicators. Ignoring this structure inflates effective dimensionality, amplifies documentation-related artifacts, and complicates interpretation. Exploratory identification of correlated feature blocks supports the use of aggregated or grouped representations that better align model inputs with meaningful clinical constructs and reduce instability in downstream modeling.

Taken together, these findings position exploratory data analysis as a diagnostic safeguard that constrains permissible methodological choices in clinical machine learning. A data-centric workflow grounded in disciplined exploratory analysis enables early identification of validation risks, interpretability limitations, and calibration challenges, rather than addressing these issues post hoc during model development.

To make these implications more explicit, the exploratory findings can be directly mapped to downstream machine learning design choices. Evidence of temporal non-stationarity indicates that model development should rely on temporally ordered validation rather than random data splitting, since performance must be assessed under realistic future-like conditions. Structured and outcome-dependent missingness suggests that missing values should not be treated as random noise or implicit clinical absence; instead, missingness-aware representations and calibration-sensitive modeling choices become necessary. Documentation bias in symptom recording limits the validity of naïve univariate interpretation and implies that downstream explainability analyses

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

must be interpreted in the context of data collection processes rather than as direct reflections of underlying clinical effects. Finally, feature redundancy and clinically coherent grouping support the use of grouped or composite representations that reduce instability and better align model inputs with meaningful clinical structure. In this sense, EDA does not merely describe the dataset, but defines the admissible methodological space for trustworthy downstream clinical ML.

The practical value of such exploratory diagnostics extends beyond methodological planning and can be directly linked to implementation of hospital ML systems. In real deployment settings, temporal drift diagnostics may be used to trigger model monitoring, recalibration, or retraining when incoming patient populations begin to diverge from development data. Analysis of missingness patterns can support admission-stage data quality monitoring by identifying which variables are systematically underdocumented and whether such underdocumentation is associated with elevated clinical risk. Detection of documentation bias can further inform interface design and data entry workflows, helping hospitals reduce systematic recording distortions that would otherwise propagate into model outputs. Thus, EDA-driven diagnostics can serve not only as a research-stage safeguard, but also as an operational component of trustworthy clinical ML implementation. Importantly, these mechanisms directly affect decision thresholds and risk communication in clinical workflows, reinforcing the need to integrate EDA-derived insights into real-world decision-support systems.

The present study deliberately focuses on exploratory diagnostics rather than predictive modeling. The identified properties of the data-generating process, including temporal non-stationarity, outcome-dependent missingness, documentation bias, and structural feature redundancy define methodological constraints that precede and shape any downstream clinical machine learning pipeline. These constraints directly inform validation design, feature representation, handling of missing information, probability calibration, and interpretation boundaries in applied settings [21-22]. Empirical model development and performance evaluation are therefore intentionally considered outside the scope of the current work. These findings align with a growing emphasis on data-centric approaches in clinical AI, suggesting that model trustworthiness is primarily shaped by upstream data properties rather than downstream algorithmic sophistication.

Several limitations should be acknowledged. The analysis was conducted using data from a single national registry and focused exclusively on admission-level clinical information. This limits direct empirical generalization of the observed distributions and documentation patterns to other countries, institutions, or disease contexts. At the same time, the central contribution of the study is methodological rather than population-specific: the identified data-generating mechanisms, including temporal drift, structured missingness, documentation bias, and feature redundancy, are characteristic of many real-world clinical datasets, although their specific manifestations may vary across institutions and healthcare systems. In addition, by design, this study refrains from model development in order to isolate exploratory effects. Future work may extend this framework by empirically evaluating how EDA-informed methodological choices influence downstream model performance, calibration, and clinical utility.

Despite these limitations, this study makes a clear methodological contribution by formalizing exploratory data analysis as an essential stage in clinical machine learning. The results underscore that reliable clinical ML begins not with algorithm selection or performance optimization, but with disciplined examination of the data-generating processes that shape model inputs. This principle is particularly critical in early clinical decision-making settings, where methodological errors introduced prior to model development may propagate downstream and directly affect patient care.

Conclusion

The findings of this study suggest that exploratory data analysis plays a fundamental role in defining the boundaries of valid modeling decisions in clinical machine learning. Systematic examination of temporal non-stationarity, missingness mechanisms, documentation effects, and

DOI: 10.37943/JXDB7524

feature structure shows that critical sources of methodological error can be identified prior to predictive model development.

The findings suggest that methodological failures in clinical machine learning may originate not only from algorithmic limitations, but also from unrecognized properties of the data-generating process. Temporal drift undermines naïve validation strategies, structured and outcome-dependent missingness distorts calibration and interpretation, documentation effects generate misleading univariate associations, and feature redundancy obscures clinically meaningful structure. Identifying these properties at the exploratory stage constrains the space of valid modeling, validation, and interpretation choices, thereby reducing the risk of biased evaluation and spurious conclusions.

By formalizing exploratory data analysis as a diagnostic safeguard that precedes model optimization, this work advances a data-centric perspective on trustworthy clinical machine learning. In practical terms, such a perspective supports the use of exploratory diagnostics as part of real-world hospital ML development, including temporal validation planning, monitoring of documentation quality, explicit handling of informative missingness, and safer interpretation of model inputs in deployment-oriented settings. Robust clinical ML systems, particularly in early decision-making settings, should therefore be grounded in disciplined examination of data-generating processes rather than post hoc correction of model behavior.

References

[1] Al-Nafjan, A., Aljuhani, A., Alshebel, A., Alharbi, A., Alshehri, A. (2025). Artificial intelligence in predictive healthcare: A systematic review. *Journal of Clinical Medicine*, 14(19), 6752. https://doi.org/10.3390/jcm14196752

[2] Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17, 195. https://doi.org/10.1186/s12916-019-1426-2

[3] Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., et al. (2020). A review of challenges and opportunities in machine learning for health. AMIA Joint Summits on Translational Science Proceedings, 2020, 191–200. https://pubmed.ncbi.nlm.nih.gov/32477638/

[4] Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., & Ranganath R. (2021). WILDS: A benchmark of in-the-wild distribution shifts. Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research, 139, 5637–5664. https://proceedings.mlr.press/v139/koh21a.html

[5] Subbaswamy, A., Saria, S. (2020). From development to deployment: Dataset shift, causality, and shift-stable models in health AI. Biostatistics, 21(2), 345–352. https://doi.org/10.1093/biostatistics/kxz041

[6] Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., et al. (2020). Prediction models for diagnosis and prognosis of COVID-19: Systematic review and critical appraisal. BMJ, 369, m1328. https://doi.org/10.1136/bmj.m1328

[7]. Rajwa, B., Naved, M. M. A., Adibuzzaman, M., Grama, A. Y., Khan, B. A., et al. (2024). Identification of predictive patient characteristics for assessing the probability of COVID-19 in-hospital mortality. *PLOS Digital Health*, 3(4), e0000327. https://doi.org/10.1371/journal.pdig.0000327

[8] Ranzani, O. T., Bastos, L. S. L., Gelli, J. G. M., Marchesi, J. F., Baião, F., Hamacher, S., & Bozza, F.A. (2021). Characterisation of the first 250 000 hospital admissions for COVID-19 in Brazil: a retrospective analysis of nationwide data. *The Lancet Respiratory Medicine*, 9(4), 407–418. https://doi.org/10.1016/S2213-2600(20)30560-9

[9] de Souza F.S.H., Hojo-Souza N.S., Batista B.D.O., da Silva C.M., Guidoni D.L. (2021). On the analysis of mortality risk factors for hospitalized COVID-19 patients in Brazil: Data from the SIVEP-Gripe database. *PLOS ONE*, 16(3), e0248580. https://doi.org/10.1371/journal.pone.0248580

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

[10] Guo, L. L., Pfohl, S. R., Fries, J., Johnson, A. E. W., Posada, J., Aftandilian, C., Shah, N., & Sung, L. (2022). Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Scientific Reports*, 12, 2726. https://doi.org/10.1038/s41598-022-06484-1

[11] Lu, S.-C., Song, W., Pfob, A., Gibbons, C. (2025). Assessing the representativeness of large medical data using population stability index. *BMC Medical Research Methodology*, 25, 44. https://doi.org/10.1186/s12874-025-02474-9

[12] Ren, W., Liu, Z., Wu, Y., Zhang, Z., Hong, S., Liu, H., & Group, O. B. O. T. M. D. I. E. H. R. (2024). Moving beyond medical statistics: A systematic review on missing data handling in electronic health records. Health Data Science, 2024, Article 0176. https://doi.org/10.34133/hds.0176

[13] Afkanpour, M., Hosseinzadeh, E., Tabesh, H. (2024). Identify the most appropriate imputation method for handling missing values in clinical structured datasets: A systematic review. *BMC Medical Research Methodology*, 24, 188. https://doi.org/10.1186/s12874-024-02310-6

[14] dos Santos Silva, G. F., Barcellos Filho, F. N., Wichmann, R. M., Costa da Silva Junior, F., Chiavegatto Filho, A. D. P. (2025). Strategies for detecting and mitigating dataset shift in machine learning for health predictions: A systematic review. *Journal of Biomedical Informatics*, 170, 104902. https://doi.org/10.1016/j.jbi.2025.104902

[15] Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., & Saria, S. (2021). The Clinician and Dataset Shift in Artificial Intelligence. *The New England Journal of Medicine*, 385(3), 283–286. https://doi.org/10.1056/NEJMc2104626

[16] Guo, L. L., Pfohl, S. R., Fries, J., Johnson, A. E. W., Posada, J., Aftandilian, C., Shah, N., & Sung, L. (2021). Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine. *Applied Clinical Informatics*, 12(4), 808–815. https://doi.org/10.1055/s-0041-1735184

[17] Vela, D., Sharp, A., Zhang, R., Nguyen, T., Hoang, A., & Pianykh, O. S. (2022). Temporal quality degradation in AI models. *Scientific Reports*, 12, 11654. https://doi.org/10.1038/s41598-022-15245-z

[18] Sperrin, M., Martin, G. P., Sisk, R., & Peek, N. (2020). Missing data should be handled differently for prediction than for description or causal explanation. *Journal of Clinical Epidemiology*, 125, 183–187. https://doi.org/10.1016/j.jclinepi.2020.03.028

[19] Nijman, S., Leeuwenberg, A., Beekers, I., Verkouter, I., Jacobs, J., Bots, M., Asselbergs, F., Moons, K., & Debray, T. (2022). Missing data is poorly handled and reported in prediction model studies using machine learning: A literature review. *Journal of Clinical Epidemiology*, 142, 218–229. https://doi.org/10.1016/j.jclinepi.2021.11.023

[20] Perets, O., Stagno, E., Yehuda, E. B., McNichol, M., Celi, L. A., Rappoport, N., & Dorotic, M. (2024). Inherent bias in electronic health records: A scoping review of sources of bias. *ACM Journal of Data and Information Quality*, 15(4), Article 35. https://doi.org/10.1145/3757924

[21] Rahmani, K., Thapa, R., Tsou, P., Chetty, S. C., Barnes, G., Lam, C., & Tso, C. F.(2023). Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction. *International Journal of Medical Informatics*, 173, 104930. https://doi.org/10.1016/j.ijmedinf.2022.104930

[22] Tamang, L., Bouadjenek, M. R., Dazeley, R., Aryal, S. (2025). Handling out-of-distribution data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 37(10), 5948–5966. https://doi.org/10.1109/TKDE.2025.3592614