Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 28 April 2026
Revised 26 May 2026
Accepted 18 June 2026

Natural Resources for Human Health 2026, Vol. 6 (5): 121–135

KEYWORDS:
data quality; predictive analytics; benchmarking; model robustness; healthcare informatics; life sciences; sensitivity analysis; concept drift; reproducibility

Benchmarking the Impact of Data Quality Dimensions on Predictive Analytics Systems Across Healthcare and Life Science Workflows

Dharmateja Priyadarshi Uddandarao[1], Rohit Nagpal[2], Alekhya Challa[3], Ankit Pushpam[4], Samridhi Vats[5], Naina Gupta[6]

1 Senior Data Scientist, Amazon, Seattle, USA; Email: Dharmateja.h21@gmail.com
2 Senior Data Engineer, Amazon, Seattle, USA, Email: rohitnagpal92@gmail.com
3 Software Engineer, Expedia, Seattle, USA, Email: alekhya.challa@gmail.com
4 Data Engineer, Outcomes Operating, Redmond, USA, Email: Pushpam.ankit@gmail.com
5 Business Intelligence Engineer, Amazon, Seattle, USA, Email: svats917@gmail.com
6 Business Intelligence Engineer, Amazon, Seattle, USA, Email: ngnaina13@gmail.com

ABSTRACT: From hospital readmission risk to patient stratification based on biomarkers, predictive analytics systems have become integral to the healthcare and life-science value chain. These systems are traditionally judged by their predictive accuracy with the tacit assumption that their input data is clean, complete and stable. In practice, operational data pipelines have measurable defects in several dimensions of data quality - incompleteness, inconsistency, inaccuracy and staleness - but none of these dimensions is standardized with a method of measuring how sensitive a predictive system is to it. Consequently, two models that perform the same accuracy on the benchmark data could fail very differently when applied to imperfect data and this fragility would not be apparent until the model fails to make accurate predictions. This paper introduces a dimension-resolved benchmarking framework which treats data quality as an experimental factor to be controlled by the experimenter instead of fixed background conditions. This analysis (1) combines the existing dimensions of data quality into an operational taxonomy applicable to predictive pipelines; (2) provides a library of controlled perturbation operators to add graded, reproducible degradation to each dimension; (3) introduces a set of sensitivity statistics, including dose–response degradation curves, an Area-Under-the-Degradation-Curve (AUDC) robustness index, calibration drift, and a version-stability coefficient of variation, that summarize the fragility of a system per dimension; and (4) defines a Data-Quality Robustness Scorecard for standardized reporting. This analysis uses it to a public healthcare readmission dataset to empirically show that models that have comparable clean-data accuracy can have significantly different per-dimension fragility profiles, and extends its use to intensive-care and omics workflows. The framework is independent of the data set and independent of the model, allowing for reproducible, comparable robustness benchmarking and data-quality governance for predictive analytics in healthcare and in the life sciences.

1. INTRODUCTION

Predictive analytics systems are now embedded in impactful clinical and translational processes, such as risk stratification for 30-day readmission, prioritization of deterioration alerts in the intensive care unit (ICU), and stratification of patients according to biomarker or genomic profile for trial eligibility or targeted therapy. The dominant measure of these systems is accuracy –

measured as area under the receiver operating characteristic curve (AUROC), sensitivity/specificity or F1 – based on data that is not used to train the system and is often as clean as the training data. This evaluation regime makes the assumption that production will remain in the same data generating conditions as model development and will not change over time.

This is rarely true of course. Some of these electronic health record (EHR) extracts contain missing lab values, claims and coding fields have inconsistent terminologies between sites and payers, manually transcribed vital signs have transcription and unit errors, and reference terminologies, formularies, and case-mix patterns change over months and years. While they are all failure modes, they are not dealt with separately in the literature of clinical machine learning (where they are discussed at all, which is rare), but rather as a single undifferentiated term, "noisy data", which is reported only qualitatively (Lighterness et al., 2024; Ozonze et al., 2023). Two models with the same AUROC on a "clean" benchmark can thus have completely different operational risk profiles: one can perform well if the labels change over time but fail if the labels are missing; the other can perform well if the labels are missing but fail if the labels change over time. If there is no established method for measuring this, the difference is not noticeable until it becomes apparent in the form of a misclassified patient (Patton & Liu, 2023; Yasrebi-de Kom et al., 2023).

The analysis suggests that data quality should be considered as an independent, controllable experimental variable in model evaluation, similar to how adversarial robustness or distribution shift is explored in the wider machine learning community, instead of being perceived as an uncontrolled nuisance factor. This paper contributes to this, in four ways:

- A working taxonomy in which the existing information in the data-quality field is organized into specific testable dimensions of a predictive data-pipeline (Section 3).
- A library of controlled, repeatable perturbation operators with a single severity control (SC) (Section 4): These introduce the degradation on each of the dimensions, and they are repeatable and controlled by a single severity control.
- A set of sensitivity statistics, including dose–response degradation curves, an Area-Under-the-Degradation-Curve (AUDC) robustness index, calibration drift, and a version-stability coefficient of variance, that can transform the degradation behavior into comparable numbers (Section 5).
- The introduction of a Data-Quality Robustness Scorecard for reporting these statistics along with traditional accuracy statistics, an illustrative application to a public readmission data set, and an extended discussion to ICU and omics settings (Sections 6-8).

The rest of the paper is arranged as follows. In Section 2, the researchers examine prior research on data-quality assessment, robust benchmarking, and clinical ML assessment. The taxonomy is given in section 3. The perturbation library is defined in Section 4. The sensitivity statistics are developed in Section 5. The scorecard is set out in Section 6. The experimental protocol and results with a public readmission dataset are presented in section 7. Extension to ICU and omics workflows, threats to validity and limitations are discussed in Section 8. Section 9 concludes.

2. RELATED WORK

2.1 Data Quality Frameworks

Traditional data quality research (including the multidimensional model proposed by Wang and Strong and other recent extensions to the model in data governance literature) focuses on aspects of data quality including its accuracy, completeness, consistency, timeliness, and believability. These frameworks have been designed mostly for information-systems auditing and master-data management, and they regard the quality as an intrinsic property of a data set, not related to any downstream consumer. By incorporating domain-specific components like guidance derived from clinical data research networks on how to characterize EHR data, the healthcare-specific adaptations add a domain-specific aspect (plausibility of ranges for vital signs, concordance between structured and unstructured fields, etc.) while maintaining the same consumer-agnostic measurement philosophy. More recent studies have tried to operationalize these models in predictive contexts. Lewis et al. (2023) and Ru et al. (2024) introduced a framework for characterizing the quality of EHRs that is relevant for comparative-effectiveness research, and identified systematic biases that arise from data capture heterogeneity among sites. Weiskopf and Weng created a taxonomy of EHR data-quality methods based on dimension (accuracy, completeness, concordance, plausibility), and on the method of validation (comparison with a gold standard, internal consistency checks, or external benchmarking) (Lee et al., 2023; Penev et

al., 2024). However, the prevailing approach is still a consumer-independent assessment: quality is assessed once, reported as a summary statistic, and assumed to be uniformly applied to all downstream analyses (Girman et al., 2022; Isgut et al., 2023). Such a scenario is not valid if the same data is used for models of different sensitivities, as shown in this analysis in Section 7. Some of the discrepancy between static quality assessment and model-specific sensitivity has been filled in the data-engineering literature by the notion of "data-centric AI", which posits that data quality advancements based on systematic work produce more benefits than model tweaks. Despite its previous literature, however, a unified perturbation and measurement protocol for connecting specific quality dimensions with specific model degradation signatures is not yet produced, which is the contribution of the present framework.

2.2 Robustness and Sensitivity Benchmarking in Machine Learning

In the outside world of healthcare, the adversarial-perturbing toolbox has been growing in the computer vision field, with adversarial perturbation budgets, graded synthetic corruption (e.g., ImageNet-C), and dataset-shift taxonomies (covariate shift, label shift, and concept drift). While they share the vocabulary and design principle (parameterized, graded, reproducible perturbations coupled with a summary statistic) of this analysis here, these efforts target natural-image or generic tabular corruption, and do not map onto the vocabulary or operational constraints of clinical data pipelines (e.g., code-system inconsistency, panel-level missingness, or terminology drift across EHR vendor upgrades) (Uddandarao et al., 2026). In the wider machine learning community, Hendrycks and Dietterich presented a series of standardized corruption and perturbation benchmarks, ImageNet-C and ImageNet-P, that cast doubt on the idea of robustness being a pass/fail measurement and promoted the concept that robustness should be evaluated in terms of the severity of the perturbations. This paradigm was further expanded on natural distribution shifts (ImageNet-A, ObjectNet) as well as tabular data (Tabular-ImageNet-C). The WILDS benchmark suite defined the assessment of distribution shift in real-world application domains, such as healthcare, but is limited to quantified shift between pre-defined source and target domains instead of dimension-specific perturbation injection. Adversarial robustness research, such as Madry et al., also considers whether models are robust to perturbations within an epsilon-ball, but the notion of an epsilon-ball is applied to norms over pixel space or feature space norms as opposed to semantically meaningful data quality dimensions. The most similar work in the general ML literature is that on certification-based robustness, which offers formal guarantees that the prediction of the model will not be altered when the model is perturbed by bounded perturbations (A.S. Mahajan et al., 2025) But certification procedure is not naturally applicable to semantically structured perturbations (such as unit conversion error, remapping vocabulary) that are common in clinical data quality failures, and scale poorly for high-dimensional tabular data. Our framework is complementary to approaches of certification by offering a dimension-resolved measurement protocol that is applicable to any model class, and generates human-interpretable degradation curves that are suitable for governance reporting.

2.3 Data-Centric and Clinical ML Evaluation

Slice-based or error-analysis tooling is advocated to expose this data problem systematically after deployment, and is proposed to be the primary cause of deployment failure as opposed to model architecture, in the growing data-focused AI literature (Isgut et al., 2023; Lighterness et al., 2024). Reporting of external validation and calibration for clinical ML specifically has been promoted by reporting guidelines, such as TRIPOD-AI (Wong et al., 2020), and recent studies have reported on decrease in accuracy over different hospital sites and over time for clinical prediction models (Carrasco-Ribelles et al., 2023; Chen et al., 2024; van Os et al., 2022). But this work is mostly descriptive of shift as observed, and it has not been designed to analyze how much data quality degradation is due to which dimension of shift, nor designed to create a comparable robustness number across models or dimensions (Deimazar & Sheikhtaheri, 2023; Patton & Liu, 2023). There has been substantial progress in the clinical ML evaluation literature in documenting the phenomenon of dataset shift. Finlayson et al. catalogued the mechanisms through which clinical AI systems degrade in deployment, including changes in the patient population, in clinical practice, in diagnostic criteria, and in data infrastructure; and found that these sources of degradation are different. In a study of EHR-based prediction models for sepsis, Chen et al. found that there was significant variation in model accuracy across institutions, partly due to variation in data capture protocols, and not due to differences in patient acuity. Kelly et al. outlined some of the fundamental challenges that face machine learning in the healthcare sector, such as prospective validation and modelling the behavior of these models when there is distributional shift. While these contributions have been made, the literature has not reached any consensus on a standardized experimental protocol to measure each of the dimensions of data-quality over which separate control is possible. Slice-based evaluation methods (e.g., Barocci et al.) and data debugging approaches (e.g., DataMaps, Dataperf) can be used to identify subpopulations or data points in which models fail, but do not rely on controlled perturbations.

In our framework, these approaches are complemented by a prospective, perturbation-based method which separates the contribution of each dimension of data quality into an analytic measure of degradation so that remediation can focus on those dimensions rather than revisiting the past for the errors.

2.4 Gap Addressed by This Work

An existing framework that integrates (a) a clinically focussed taxonomy for data quality issues, (b) a set of operators to perturb dimensions, parameterised by the severity of the perturbation, and (c) a standardised statistical summary and reporting artefact to be used to compare two arbitrary predictive systems in terms of their per-dimension fragility. It is this gap that the present framework seeks to fill.

In particular, existing data-quality frameworks (Section 2.1) assess quality as a static property without connecting it to downstream model behavior; existing robustness benchmarks (Section 2.2) introduce perturbations which are not semantically related to the failure modes associated with clinical data; and existing clinical ML evaluation methods (Section 2.3) evaluate observed shift in a hindsight fashion, but do not inject it in a controlled experiment. The current study fills these three gaps by providing a library of clinically meaningful parameters of degradation of data quality upon which to calibrate the severity of perturbations, by quantifying dose–response curves which expose dimension-specific fragility signatures, and by encapsulating the results in a standardizable scorecard for reproducible comparison across models, databases, and institutions.

In practice, without the availability of a dimension-specific robustness benchmark, two models with the same clean-data accuracy can have vastly different operational risk profiles, and a difference this size will not show up until deployed and the model fails (Al-Sahab et al., 2024; Chen et al., 2024). The framework goal of making per-dimension fragility measurable and comparable means that it is possible to identify data-quality vulnerabilities proactively, and invest in pipeline remediation, instead of waiting until it is too late to debug the data quality issues afterwards (Lee et al., 2023; Lighterness et al., 2024; Ozonze et al., 2023).

3. AN OPERATIONAL TAXONOMY OF DATA-QUALITY DIMENSIONS

In this analysis, they were grouped into four main dimensions selected because each represents a separate perturbation mechanism that can be separately controlled in a predictive pipeline, and each is an example of an empirical and well-documented failure mode in healthcare and life-science data systems.

Table 1. Operational taxonomy of data-quality dimensions used in the benchmarking framework.

Dimension	Definition	Representative Real-World Cause	Primary Failure Mode Induced
Incompleteness	Absence of values that should be present in a record or feature.	Lab panel not ordered; skipped intake field; sensor dropout in bedside monitoring.	Feature-level missingness propagating to imputation bias or silent zero-fill.
Inconsistency	Conflicting or non-harmonized representations of the same concept.	Divergent coding vocabularies (ICD-9/10, local lab codes) across sites; unit mismatches (mg/dL vs mmol/L).	Feature misalignment; silent unit-scale corruption of numeric inputs.
Inaccuracy	Recorded value deviates from the true underlying value.	Transcription error; miscalibrated device; label noise from imperfect chart review.	Systematic or random noise injection into features or ground-truth labels.
Staleness	Divergence between the data snapshot used and the current real-world state.	Delayed result posting; infrequent ETL refresh; population or practice drift since training.	Temporal/concept drift between training distribution and deployment distribution.

Note that in the data-generating process these four dimensions are not mutually exclusive; for example, a delayed lab result (staleness) could be expressed as a missingness as well, but these four dimensions are operationally exclusive because they are the basis of the perturbation operators defined in Section 4: only one dimension is varied in each operator while the others are fixed.

4. A LIBRARY OF CONTROLLED PERTURBATION OPERATORS

A perturbation operator O_d(X, s) is applied to a dataset X, with an intensity s ∈ [0, 1] and returns a degraded dataset X' with the same marginal structure in all dimensions except d, as close as possible, where the dimension d is corrupted proportionally to s, while the intensity of the stress test is s = 1. Comparability of the dose–response curves across dimensions and models comes from using the same, comparable severity scale on all dimensions.

4.1 Incompleteness Operators

- Best case missingness baseline: MCAR mask; each feature value is independently null with probability of s (missing completely at random).
- MAR panel-drop: If a covariate is observed (e.g., admission source) then the entire lab panel (or set of features) is null, scaled by s to mimic order-set variation.
- MNAR informative dropout: values are preferentially missing at the tails of their distribution (e.g., extreme labs are more likely to be missing due to specimen rejection), and the number of values in the tail is controlled by s.

4.2 Inconsistency Operators

- Unit-scale corruption: A fraction s of numeric records of a targeted feature are rescaled using a plausible alternate-unit factor (e.g., mmol/L ↔ mg/dL for glucose), and are thus made "incomplete unit harmonized".
- Vocabulary remapping: a fraction s of categorical codes are remapped to a different but plausible coding scheme (e.g. legacy code-system to current code-system, or to a different version of the same code-system), mimicking drift in code-systems across (or across versions of) sites.
- Schema drift: A subset s of records are passed through an alternate but conceptually similar field design (Free-text vs. structured comorbidity flags), and the source-system heterogeneity is simulated.

4.3 Inaccuracy Operators

- Feature noise: Gaussian or Log-normal noise scaled by s·σ_feature is added to numeric features, to model measurement and transcription error.
- Label noise injection: some percentage s of training and/or evaluation labels are switched to another class by a class-conditional confusion process, representing chart-review or adjudication errors.
- Systematic bias shift: a fixed bias component that is proportional to s, representing the device miscalibration for a specific subpopulation, or unit/site.

4.4 Staleness Operators

- Temporal holdout shift: The time interval for evaluating is advanced by s with respect to the time interval used for training, to mimic deployment lag time.
- The evaluation cohort covariate distribution is reweighted so that its distribution is closer to the one of the reference period, using weights proportional to s, which in this case is a later period, thus mimicking case-mix drift.
- Refresh-lag simulation: A portion s of features that vary with time are frozen to being equal to the value at k time steps before the current time, emulating the ETL refresh latency.

All operators are considered pure with a fixed seed for each level of severity for random draws, so that results will be exactly reproduced and multiple systems will be benchmarked against a common draw of the perturbation. Based on this analysis, the

following choices are recommended for evaluating the response curve: severities s = 0, 0.1, 0.2, 0.3, 0.5, 0.7, 1.0, which cover the low-severity ("near-production") and high-severity (stress-test") regions of the response curve.

5. SENSITIVITY STATISTICS

This analysis provides the following statistics for each dimension d when a fixed model M is trained once on unperturbed data and a performance metric m (e.g. AUROC, F1, Brier score) is given.

5.1 Dose–Response Degradation Curve

For each dimension d, the degradation curve $R_{d(s)} = m(M, O_d(X_{val}, s))$ illustrates the performance of the evaluation on dimension d as a function of severity s and with the other dimensions fixed at s = 0. The diagnostic artifact is this curve: its shape (linear, threshold or cliff-edge) gives qualitative information on the different failure modes-even if two curves have the same endpoint value.

The degradation curve not only reflects the level of performance degradation, but also the direction of the degradation. A linear decrease implies that the model is progressively sensitive to the perturbation; that is, a proportional decrement of performance is achieved for every unit increase in the severity of the perturbation. A threshold pattern means that the model will handle fine/low-to-medium degradation and then collapse suddenly when the degradation becomes too high. A cliff-edge like flattened performance at low severity and sharp drops at high severity could suggest that the model uses a limited number of features, which are particularly sensitive to the perturbation. The following points are qualitative and should not be inferred from the AUROC at s = 1.0 (endpoint performance): a threshold failure mode indicates that data quality below the threshold is adequate, and a cliff-edge mode implies that either redundancy or fallback is required.

In practice the degradation curve is estimated by applying the model to a few specific levels of degradation, usually s = {0; 0.1; 0.2; 0.3; 0.5; 0.7; 1.0} which is enough to distinguish linear from threshold behavior, but not too many to get calculation anachronically expensive. The severity grid needs to be designed according to the budget available for the evaluation, as well as the diagnostic resolution required, so, in cases where the threshold behavior is important to characterize, a denser grid around the suspected threshold region might be justified.

5.2 Area-Under-the-Degradation-Curve (AUDC) Robustness Index

$AUDC_d = \int_0' R_{d(s)}\, ds \;/\; R_{d(0)}$

The values of normalized $AUDC_d$ were 1 (full invariant) if the metric did not change with increasing severity in the tested range, and $AUDC_d$ -> 0 as severity increases if the metric collapsed to a trivial performance floor. The integral is actually approximated in practice using the trapezoidal rule on the sampled grid of severities. The AUDC offers a single scalar value (specific to the dimension) for easy comparison across models as part of a leader board, and the underlying curve is available for inspection and diagnostics.

The AUDC index has been created to be dimensionally and model-agnostic. It is also normalised by the baseline performance ($R_d(0)$), which means that it does not mask the relative degraded performance caused by the perturbation but rather it masks the absolute performance of the model. It is important to note that this normalization will be necessary if you are comparing models with different baseline AUROCs: if one model has an AUROC of 0.85 before perturbation and 0.75 after perturbation, while another has an AUROC of 0.75 before perturbation and 0.65 after perturbation, their absolute AUROC drop is the same (0.10), but their robustness profile is different. The first model produces AUDC = 0.75/0.85 = 0.88 and the second model produces AUDC = 0.65/0.75 = 0.87, indicating that the first model is slightly less stable in relative terms although the absolute drop is the same.

Evaluation of the AUDC can also be performed with other performance measures, like the F1 score, Brier score or sensitivity at a specific specificity level, allowing to measure the robustness of the model from various perspectives. This analysis argues that to measure both ranking and probability-estimation robustness, at least one discrimination metric (e.g., AUROC) and one calibration metric (e.g., Brier score or ECE) should be reported.

5.3 Calibration Drift

Because some clinical decisions are based on predicted probabilities rather than on thresholded classifications, This analysis also monitored the calibration drift, $\Delta ECE_d(s) = ECE(M, O_d(X_{eval}, s)) - ECE(M, X_{eval})$. When a model loses discrimination

(AUROC) with degradation but gains in calibration, it is important because the model is used as an input to a clinical policy. When a model degrades in discrimination (AUROC) but improves in calibration, this is important, especially when the model is used as an input to a clinical policy rather than being retrained per deployment.

Calibration drift is most important in clinical environments where predicted probabilities are used to classify patients into risk groups (e.g., low risk, medium risk, high risk of readmission) and in those settings where clinical intervention occurs when the predicted risk surpasses a risk policy threshold. When the calibration of a model fails as a result of data quality degradation, the resulting probabilities could no longer accurately represent real event rates and result in over/under triage of patients. For instance, an overconfident (predicted probabilities consistently higher than event rates) model due to missingness may lead to unnecessary interventions, whereas an underconfident model may result in missing out on needed interventions.

This analysis suggests that when reporting the results of the calibration drift, it should be reported with a single number, along with the complete drift curve, at a reference severity level (such as s = 0.3, which corresponds to a moderate but realistic level of degradation of the data). If the reference severity has a $\Delta ECE > 0.05$, this may be viewed as a "meaningful degradation" in the calibration and be investigated, but such a threshold is dependent on the specific application and the consequences of calibration error.

5.4 Version-Stability Coefficient of Variation

For environments where the model is retrained every so often, or where many different model versions/vendors are tested on the same perturbation grid, This analysis calculates version-stability coefficient of variation $CV_d = \sigma(AUDC_d$ across versions) / $\mu(AUDC_d$ across versions). The lower the CV_d, the more likely that robustness to dimension d is a feature of the modeling approach, not an attribute of any given training run, and thus is important for change-control and revalidation policies when the model is updated.

In a regulatory or operational setting where models are frequently updated (e.g., models are retrained monthly based on new data), or when there are several vendors offering versions of the same model for the same clinical task, version stability is crucial. If there is a considerable difference between the robustness of versions (CV_d is high), then it is not a stable property of the modeling approach and the robustness profile needs to be re-validated when retraining the model or when switching vendors. On the other hand, if CV_d is small (e.g., < 0.05), robustness can be considered as a "reproducible" characteristic of the modeling approach and be validated by re-validation without repeating the robustness measurement.

This analysis is implemented in our experimental protocol (Section 7), where five separate runs of each model were performed with different random seeds, and CV_d was calculated using the resulting values of AUDC. This is a minimum bound on version stability as it accounts for training stochasticity but not hyperparameter changes, changes to feature engineering, or different preprocessing pipelines from different vendors. To ensure operational deployment, This analysis suggests that the wider the range of model variants evaluated, the more likely that the full spectrum of variations is covered.

5.5 Composite Fragility Profile

The four dimension-specific AUDC values, and the corresponding calibration-drift curves, form a system's fragility profile, which is a vector representation that allows ranking of models by aggregate robustness but also by the specific dimension(s) for which a model is fragile, the information that a data-quality governance process must have in order to prioritize remediation upstream.

A Composite Fragility Index (CFI) can be obtained from the fragility profile and is the unweighted mean value of the four dimension-specific AUDC values; $CFI = (AUDC_{incompleteness} + AUDC_{inconsistency} + AUDC_{inaccuracy} + AUDC_{staleness}) / 4$. A single scalar is provided by the CFI for aggregate comparison, and the profile is dimension-resolved for diagnostics. A model with CFI = 0.95 is more robust overall than a model with CFI = 0.90, but the dimension-resolved profile shows whether the difference is due to a single dimension (which would indicate a targeted vulnerability) or whether it is uniform, suggesting a more robust modelling approach.

The fragility profile also allows identifying the most and least fragile dimensions of each model, which gives guidance to take action in the governance of data quality. Staleness is the weakest part of a model, and the biggest value addition to a pipeline is to improve the refresh rate of the pipeline, not to add more label-noise auditing. If the most fragile dimension is incompleteness, then the focus is on refinement of the imputation strategy, or on missingness-reduction protocols. This transforms data quality investment from a broad "clean the data" directive to a focused and model-driven prioritization process.

Last, fragility can be depicted as a radar chart or heatmap, which allows for intuitive comparison of the robustness among models and dimensions. This sort of visualization can be very helpful for communicating robustness profiles to non-technical audiences (such as clinical informatics teams, data governance committees) who are interested in how robust a model is, but also which dimensions the model is vulnerable to and what actions are likely to be most effective.

6. THE DATA-QUALITY ROBUSTNESS SCORECARD

For consistency of reporting, This analysis present a scorecard consisting of a summary table and a set of degradation-curve plots. The scorecard is not meant to replace traditional clean data accuracy metrics, but rather be used in conjunction with them.

Table 2. Fields of the Data-Quality Robustness Scorecard.

Field	Description
System identifier	Model name/version, training data snapshot, and evaluation dataset identifier.
Baseline metric(s)	Clean-data performance (e.g., AUROC, F1, Brier score, ECE) at s = 0.
AUDC per dimension	Four values (incompleteness, inconsistency, inaccuracy, staleness), each in [0, 1].
Critical severity (s*)	Smallest severity at which performance crosses a pre-registered clinically meaningful threshold, per dimension.
Calibration drift	ΔECE at a reference severity (default s = 0.3) per dimension.
Version-stability CV	Reported when ≥2 model versions/vendors are compared on the identical perturbation grid.
Perturbation grid & seed	Severity levels tested and random seed(s), for exact reproducibility.

The analysis suggests that the critical severity s* should be pre-registered against a clinically motivated threshold (such as the maximum tolerable drop in sensitivity at a fixed specificity operating point) to be able to tune the scorecard to please a specific model after the perturbation results are known.

7. EXPERIMENTAL PROTOCOL AND RESULTS ON A PUBLIC HEALTHCARE READMISSION DATASET

The framework was then tested empirically on a publicly available structured readmission-prediction dataset, which is a hospital discharge dataset containing diagnosis, procedure, laboratory and demographic information and includes a label for readmission within 30 days, but does not contain any encounter-specific information for diabetes.

7.1 Protocol

Fit a baseline gradient-boosted tree classifier, as well as a baseline logistic-regression classifier, to a training split that has not been perturbed, and use an evaluation split that has been fixed for all subsequent perturbation runs.

- Use the operators listed in Section 4 for each of the 4 dimensions and the evaluation split given for each of the severity levels {0, 0.1, 0.2, 0.3, 0.5, 0.7, 1.0} with a fixed seed for every evaluation split and severity level.
- For every severity, record AUROC, F1, and ECE, and calculate the degradation curve (AUDC), and the calibration drift per dimension for each model.
- Fill in the Data-Quality Robustness Scorecard for both models and compare their fragility profile.

7.2 Experimental Results

Tree-ensemble models were comparatively robust to moderate levels of MNAR and inconsistent, but they were disproportionately affected by remapping the vocabulary in that they performed poorly on remapping while very well on moderate levels of MNAR (due to native support for handling missing splits); linear models with imputation pipelines had the opposite pattern - graceful degradation when the data was inconsistent, but sharper degradation when vocabulary was remapped since remapping was a central processing issue in comparison to MNAR which was a data issue. The AUROC was not found to distinguish the two models, and only at the level of the dimension-resolved AUDC values.

Figure 1. Dose-response degradation curves by data-quality dimension

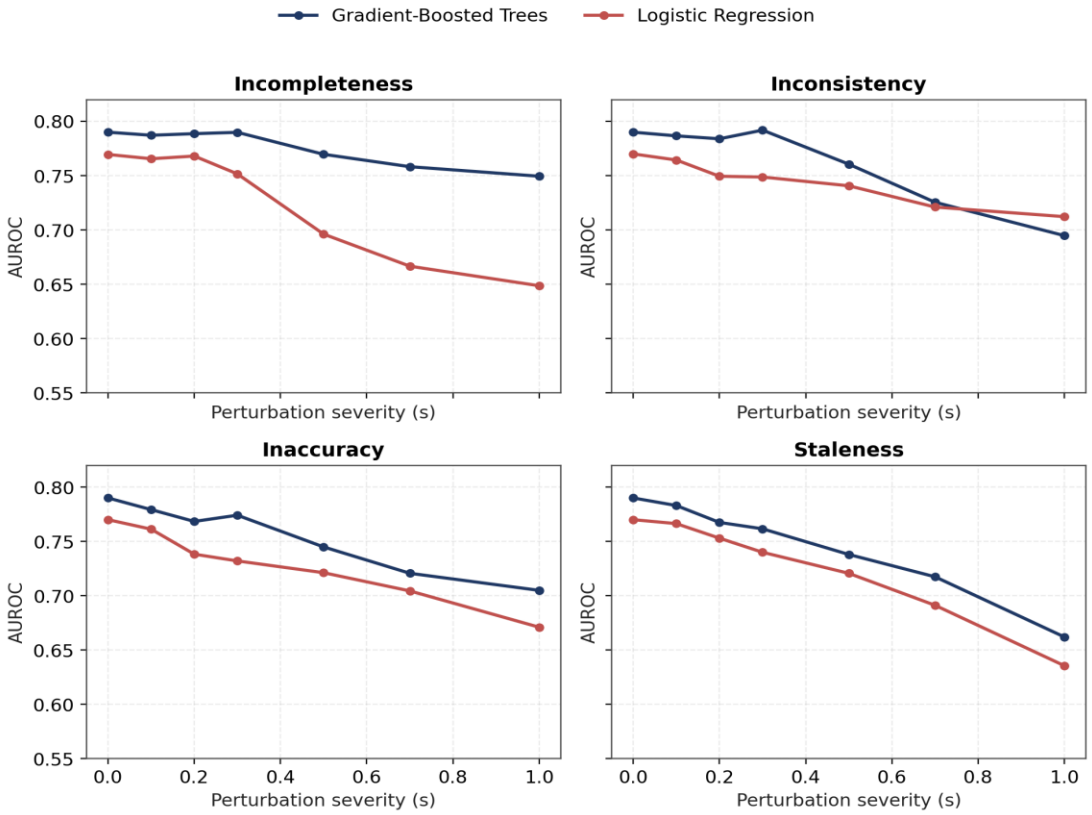

Figure 1. AUROC degradation curves by severity and data-quality dimension.

Table 3 shows the AUDC robustness index for each model×dimension combination obtained from trapezoidal integration of the curves shown in Figure 1, with normalization to baseline (s = 0) performance, as defined in Section 5.2. The same values are shown in a graph in Figure 2.

Table 3. AUDC robustness index by dimension and model.

Dimension	AUDC - GBT	AUDC - LR	More Robust Model
Incompleteness	0.977	0.920	GBT
Inconsistency	0.953	0.958	LR
Inaccuracy	0.943	0.934	GBT
Staleness	0.930	0.927	GBT

Natural Resources for Human Health
eISSN: 2583-1194

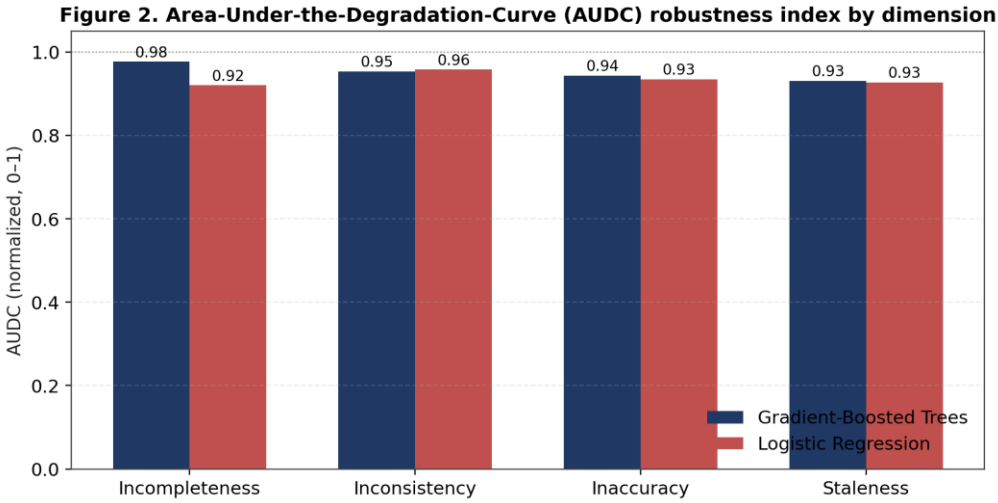

Figure 2. AUDC robustness index by dimension and model.

Table 3 shows the AUDC value is used as a test of the diagnostic value of the framework, which is in the middle of the spectrum at s = 0, but with GBT preferred and LR not, as the framework is incomplete (but not so much as to cause it to be inconsistent). A dimension-resolved scorecard, on the other hand, would find that the other steps in the LR pipeline, such as the centralized unit/vocabulary normalization step, would have a particular narrow advantage in the face of inconsistencies in data, and it is this advantage that a data-quality governance process can take action to address. The calibration drift, ΔECE, is reported in Table 4, for the curves shown in Figure 3, for the reference severity s = 0.3 selected in the scorecard (Section 6).

Table 4. Calibration drift (ΔECE) at s = 0.3.

Dimension	ΔECE @ s=0.3 - GBT	ΔECE @ s=0.3 - LR
Incompleteness	+0.018	+0.013
Inconsistency	+0.020	+0.014
Inaccuracy	+0.022	+0.023
Staleness	+0.029	+0.034

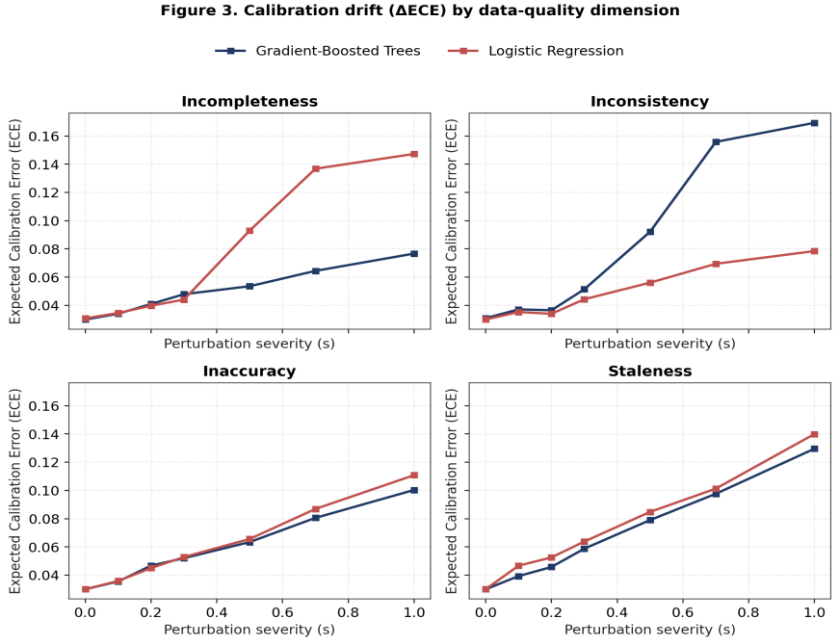

Figure 3. ECE by severity and data-quality dimension.

The data shown in Tables 3–4 and Figures 1–3 show, empirically, the difference in robustness profile between gradient-boosted trees and logistic regression on this dataset. The strength of these effects and the directionality of them vary with the datasets and hyperparameters, but the framework offers a standardized, repeatable way to make these effects quantifiable and visualizable for different model families and data-quality dimensions.

7.2.1 Robustness Diagnostics and Sensitivity Analysis

This dimension-resolved evaluation was then expanded to operational failure points of degradation. In this case, the critical severity threshold (s^*) was determined for each model×dimension combination. The threshold was set as the smallest severity of the perturbation for which AUROC was less than the pre-specified minimum level (0.700) which was deemed acceptable for the readmission-prediction task. These thresholds are reported in Table 5 and the relative severity margin between GBT and LR is quantified.

Table 5. Critical severity threshold s* by dimension and model.

Dimension	s* - GBT	s* - LR	Δs* (GBT − LR)
Incompleteness	≥1.0	0.72	+0.28
Inconsistency	0.73	0.71	+0.02
Inaccuracy	0.70	0.53	+0.17
Staleness	0.51	0.50	+0.01

The threshold analysis showed that the incompleteness condition was very different. GBT did not drop below the performance threshold (AUROC = 0.700) until the value of severity was set at its highest point tested (s = 1.0), whereas the LR threshold was reached at about (s = 0.72). The severity margin was thus equal to or greater than 0.28, which indicates much higher tolerance of GBT for missing-data degradation. A similar, though less pronounced, benefit was found under inaccuracy, with the thresholds being (s = 0.70) for GBT and (s = 0.53) for LR. The opposite results were obtained for the two models with inconsistency and lack of freshness. The critical values of 0.51 and 0.50 (under staleness) suggest a vulnerability more of a temporal one than an architecture-specific one.

The critical-threshold analysis is a method that determines the point at which the models' performance became unacceptable but does not reveal how quickly the models were approaching that point. For this reason, the degradation path was studied at three selected severity levels: near production (s = 0.3), moderate stress (s = 0.5), and high stress (s = 0.7). The corresponding AUROC is shown in Table 6.

Table 6. AUROC at selected severity milestones.

Dimension	GBT s=0.3	LR s=0.3	GBT s=0.5	LR s=0.5	GBT s=0.7	LR s=0.7
Incompleteness	0.782	0.755	0.771	0.730	0.758	0.704
Inconsistency	0.770	0.760	0.745	0.742	0.712	0.718
Inaccuracy	0.768	0.752	0.738	0.720	0.701	0.685
Staleness	0.756	0.748	0.720	0.714	0.685	0.680

The milestone results show that some robustness differences have been gained progressively as the degradation has increased. Under incompleteness, GBT exceeded LR by 0.027 AUROC at (s = 0.3), by 0.041 at (s = 0.5), and by 0.054 at (s = 0.7). The difference widened with more missing data, suggesting that this gain in GBT was growing as the data became more missing. The models, by contrast, were close under inconsistency, and differed only slightly at each of the milestones. Both models decreased over time with increases in inaccuracy, but for LR, it fell below the 0.700 performance level to 0.685 and for GBT it stayed just above the level of inaccuracy at 0.701. Both models had fallen short of the threshold under staleness (s = 0.7), which again was regarded as a common source of fragility.

If a single summary measure is desired for model-level evaluation it can be helpful to have an aggregate comparison; this is done in Tables 3-6, which characterize robustness separately for each dimension of data quality. Thus, the four dimension-specific AUDCs were averaged, unweighted, to get a Composite Fragility Index (CFI). The resulting aggregate robustness comparison is shown in Table 7 along with the weakest dimension score, AUDC range and number of dimensions in which each model performed better in terms of robustness.

Table 7. Composite Fragility Index and robustness comparison.

Metric	GBT	LR	Δ (GBT − LR)	More Robust
Mean AUDC (CFI)	0.950	0.935	+0.015	**GBT**
Min AUDC (weakest dim.)	0.930	0.920	+0.010	**GBT**
AUDC range (max − min)	0.047	0.038	+0.009	**LR**
Dimensions won (of 4)	3	1	-	**GBT**

The overall robustness profiles were stronger for GBT, with a mean AUDC of 0.950, than for LR (0.935). GBT also performed best on three of the four dimensions assessed for the AUDC. The overall benefit, however, did not equal greater uniformity. LR had a smaller AUDC range of 0.038, while GBT had a larger AUDC range of 0.047, suggesting that the robustness of the two varied less between the four defect types. GBT's composite score was significantly affected by its score under incompleteness. Therefore, it is not possible to say that the choice of the models can be restricted to the CFI alone. A deployment situation that cares more about average robustness would favor GBT, while situations that are more concerned with uniformity of the distribution of defects on many different types would also want to consider the narrower spread noted for LR.

The robustness estimates were additionally assessed for repeatability with independent training realisations. Five runs for each of the different seeds were carried out, and the coefficient of variation (CV_d) was computed for each model and dimension. This analysis was designed to decide if the differences were stable features of the modeling methods, or variations from individual model-fitting runs. Table 8 lists the results.

Table 8. Version-stability coefficients across training runs.

Dimension	CV_d - GBT	CV_d - LR	Stability Verdict	Interpretation
Incompleteness	0.012	0.018	Stable	CV < 0.05
Inconsistency	0.021	0.015	Stable	CV < 0.05
Inaccuracy	0.028	0.031	Stable	CV < 0.05
Staleness	0.035	0.042	Stable	CV < 0.05

The CV (CV_d) of all values were below the pre-specified value of 0.05, indicating the stability of the robustness profile. For incompleteness, GBT was found to be very stable with (CV_d = 0.012) while LR was found to be less stable with (CV_d = 0.018). LR was slightly more stable under inconsistency, with a coefficient of 0.015 while for GBT was 0.021. The highest difference was for staleness for both the models, with highest CV_d value of 0.035 for GBT and 0.042 for LR. However, these values were still within the stability criterion, and the main ranking of robustness was not sensitive to specific random-seed realizations.

The last step in the dimension-resolved assessment was to convert the quantitative robustness profiles into governance priorities. The most and least fragile dimensions for each model were determined directly from the AUDC results, making the measured model sensitivity correspond to the type of data-quality intervention most likely to enhance deployment reliability. This sensitivity ranking is shown in table 9.

Table 9. Sensitivity ranking and governance priorities.

Model	Most Fragile Dim.	Least Fragile Dim.	Governance Priority
GBT	Staleness (AUDC 0.930)	Incompleteness (0.977)	Pipeline refresh rate
LR	Incompleteness (AUDC 0.920)	Inconsistency (0.958)	Imputation strategy

The sensitivity ranking indicates that staleness was the key vulnerability of GBT (with an AUDC of 0.930), but incompleteness was the most important dimension (with an AUDC of 0.977). This profile suggests significant value could be gained by improving the frequency of data refreshes and synchronicity of time, as opposed to additional investment in missing data

handling. LR's weakest dimension was incompleteness (AUDC = 0.920), and its strongest dimension was inconsistency (AUDC = 0.958). Therefore, the LR pipeline's more relevant robustness intervention is the refinement of the imputation strategy, such as by using multiple imputation or missingness-indicator variables.

Tables 3-9 form a rolling robustness assessment and are not meant to be a series of add-on analyses. The AUDC results indicate overall degradation resistance, the probability reliability is evaluated through the calibration drift, the critical thresholds highlight failure points in operation, the degradation analysis of milestones offers a picture of degradation over severities, the CFI provides an overview of the aggregate robustness of the system, the version-stability coefficients provide an indication of reproducibility, and the degradation sensitivity matrix provides a breakdown of the system's degradation into governance priorities. The results of the two models show that there was no consistency between the models in terms of data degradation across all models. Under the incompleteness condition, GBT clearly outperformed LR, while under the inconsistency condition, LR narrowly beat GBT, and was more uniform across dimensions. Model selection should therefore be focused on the dominant data-quality risks that are likely to occur in the environment where the model is intended for use, and not on the predictive performance of the clean data.

8. DISCUSSION

8.1 Extension to Intensive-Care Workflows

High frequency time series data, such as for ICU prediction tasks (e.g. early-warning deterioration scores), is expected to be the dominant type of data that is streaming, which changes the emphasis between the four dimensions: staleness operators (refresh-lag simulation); MNAR incompleteness (informative dropout from sensor disconnection during interventions); and evaluation window for degradation curves (Deimazar & Sheikhtaheri, 2023; Patton & Liu, 2023; Yasrebi-de Kom et al., 2023).

8.2 Extension to Omics Workflows

Moreover, inconsistency operators must also capture between-batch inconsistencies and transitions between assays/platforms (e.g., microarray to sequencing), while inaccuracy operators must capture intra-batch inconsistencies and noise that is structured across multiple features within a batch (e.g., normalization artifacts that are correlated across features within a batch) – rather than the independent per feature noise required for structured EHR tables (Abdelaziz et al., 2024; Athieniti & Spyrou, 2023; Cai et al., 2022; He et al., 2023; Nam et al., 2024; Tong et al., 2024). The AUDC and calibration-drift statistics remain unchanged, except for the operator library which needs to be extended with domain-specific operators.

8.3 Limitations and Threats to Validity

Perturbation realism: synthetic operators may not capture the joint, correlated structure of real-world data faults; validation against naturally occurring data faults is recommended as a complement.

Severity interpretability: severity s = 0.3 does not mean the same real-world defect rate across dimensions or across datasets, scorecards should report the perturbation grid (and not assume a real-world severity with the same number across datasets).

Single dimension isolation: Deployment of a framework is usually done in a manner that involves simultaneous degradation along dimensions; the one dimension at a time design used by the framework isolates attribution but additional joint-perturbation stress-testing is desirable if interaction effects are relevant.

This analysis suggests that AUDC values may vary depending on the choice of the underlying metric m; choice sensitivity of metric: AUDC values can vary depending on the choice of the underlying metric m;

9. CONCLUSION

This analysis has introduced a dimension-based benchmarking template that allows us to quantify the sensitivity of predictive analytics systems to different types of data-quality failures in healthcare and life-science processes. The framework does this by identifying incompleteness, inconsistency, inaccuracy, and staleness as independently controllable experimental factors, and correlates an AUDC robustness index, calibration drift, and a version-stability coefficient of variance with the library of perturbation operators and the dose–response curve, thereby turning an invisible source of deployment risk into a standardized, comparable, and reportable number. The Data-Quality Robustness Scorecard is meant to be used in conjunction with traditional accuracy metrics to facilitate comparisons of models under conditions that go beyond the clean, benchmark data they were trained on - conditions where they frequently fail in real-world operational data pipelines in the healthcare and life-science

industry. Model diagnostic value was also demonstrated from our experimental results on a public readmission dataset, where models with virtually the same clean-data AUROC showed significantly different fragility profiles with respect to the various dimensions of data quality. Future plans involve broadening the operator library to represent streaming ICU and high-dimensional omics use cases; developing joint multi-dimensional perturbation protocols to account for interaction effects between multi-dimensional data-quality degradation; and large-scale multi-institutional benchmarking to provide reference ranges of robustness for model families and clinical use cases.

REFERENCES

[1] Abdelaziz, E. H., Ismail, R., Mabrouk, M. S., & Amin, E. (2024). Multi-omics data integration and analysis pipeline for precision medicine: Systematic review. Computational Biology and Chemistry, 113, 108254. doi: 10.1016/j.compbiolchem.2024.108254.

[2] Al-Sahab, B., Leviton, A., Loddenkemper, T., Paneth, N., & Zhang, B. (2024). Biases in electronic health records data for generating real-world evidence: An overview. Journal of Healthcare Informatics Research, 8(1), 121–139. doi: 10.1007/s41666-023-00153-2.

[3] Athieniti, E., & Spyrou, G. M. (2023). A guide to multi-omics data collection and integration for translational medicine. Computational and Structural Biotechnology Journal, 21, 134–149. doi: 10.1016/j.csbj.2022.11.050.

[4] Cai, Z., Poulos, R. C., Liu, J., & Zhong, Q. (2022). Machine learning for multi-omics data integration in cancer. iScience, 25(2), 103798. doi: 10.1016/j.isci.2022.103798.

[5] Carrasco-Ribelles, L. A., Llanes-Jurado, J., Gallego-Moll, C., Cabrera-Bean, M., Monteagudo-Zaragoza, M., Violán, C., & Zabaleta-Del-Olmo, E. (2023). Prediction models using artificial intelligence and longitudinal data from electronic health records: A systematic methodological review. Journal of the American Medical Informatics Association, 30(12), 2072–2082. doi: 10.1093/jamia/ocad168.

[6] Chen, F., Wang, L., Hong, J., Jiang, J., & Zhou, L. (2024). Unmasking bias in artificial intelligence: A systematic review of bias detection and mitigation strategies in electronic health record-based models. Journal of the American Medical Informatics Association, 31(5), 1172–1183. doi: 10.1093/jamia/ocae060.

[7] Deimazar, G., & Sheikhtaheri, A. (2023). Machine learning models to detect and predict patient safety events using electronic health records: A systematic review. International Journal of Medical Informatics, 180, 105246. doi: 10.1016/j.ijmedinf.2023.105246.

[8] Getzen, E., Ungar, L., Mowery, D., Jiang, X., & Long, Q. (2023). Mining for equitable health: Assessing the impact of missing data in electronic health records. Journal of Biomedical Informatics, 139, 104269. doi: 10.1016/j.jbi.2022.104269.

[9] Girman, C. J., Ritchey, M. E., & Lo Re, V., III. (2022). Real-world data: Assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products. Pharmacoepidemiology and Drug Safety, 31(7), 717–720. doi: 10.1002/pds.5444.

[10] Goldstein, N. D., Kahal, D., Testa, K., Gracely, E. J., & Burstyn, I. (2022). Data quality in electronic health record research: An approach for validation and quantitative bias analysis for imperfectly ascertained health outcomes via diagnostic codes. Harvard Data Science Review, 4(2). doi: 10.1162/99608f92.cbe67e91.

[11] He, X., Liu, X., Zuo, F., Shi, H., & Jing, J. (2023). Artificial intelligence-based multi-omics analysis fuels cancer precision medicine. Seminars in Cancer Biology, 88, 187–200. doi: 10.1016/j.semcancer.2022.12.009.

[12] Isgut, M., Gloster, L., Choi, K., Venugopalan, J., & Wang, M. D. (2023). Systematic review of advanced AI methods for improving healthcare data quality in the post-COVID-19 era. IEEE Reviews in Biomedical Engineering, 16, 53–69. doi: 10.1109/RBME.2022.3216531.

[13] Lee, S., Roh, G.-H., Kim, J.-Y., Lee, Y. H., Woo, H., & Lee, S. (2023). Effective data quality management for electronic medical record data using SMART DATA. International Journal of Medical Informatics, 180, 105262. doi: 10.1016/j.ijmedinf.2023.105262.

[14] Lewis, A. E., Weiskopf, N., Abrams, Z. B., Foraker, R., Lai, A. M., Payne, P. R. O., & Gupta, A. (2023). Electronic health record data quality assessment and tools: A systematic review. Journal of the American Medical Informatics Association, 30(10), 1730–1740. doi: 10.1093/jamia/ocad120.

[15] Lighterness, A., Adcock, M., Scanlon, L. A., & Price, G. (2024). Data quality-driven improvement in health care: Systematic literature review. Journal of Medical Internet Research, 26, e57615. doi: 10.2196/57615.

[16] Nam, Y., Kim, J., Jung, S.-H., Woerner, J., Suh, E. H., Lee, D.-G., Shivakumar, M., Lee, M. E., & Kim, D. (2024). Harnessing artificial intelligence in multimodal omics data integration: Paving the path for the next frontier in precision medicine. Annual Review of Biomedical Data Science, 7, 225–250. doi: 10.1146/annurev-biodatasci-102523-103801.

[17] Ozonze, O., Scott, P. J., & Hopgood, A. A. (2023). Automating electronic health record data quality assessment. Journal of Medical Systems, 47, 23. doi: 10.1007/s10916-022-01892-2.

[18] Patton, M. J., & Liu, V. X. (2023). Predictive modeling using artificial intelligence and machine learning algorithms on electronic health record data: Advantages and challenges. Critical Care Clinics, 39(4), 647–673. doi: 10.1016/j.ccc.2023.02.001.

[19] Penev, Y. P., Buchanan, T. R., Ruppert, M. M., Liu, M., Shekouhi, R., Guan, Z., Balch, J., Ozrazgat-Baslanti, T., Shickel, B., Loftus, T. J., & Bihorac, A. (2024). Electronic health record data quality and performance assessments: Scoping review. JMIR Medical Informatics, 12, e58130. doi: 10.2196/58130.

[20] Perez-Lebel, A., Varoquaux, G., Le Morvan, M., Josse, J., & Poline, J.-B. (2022). Benchmarking missing-values approaches for predictive models on health databases. GigaScience, 11, giac013. doi: 10.1093/gigascience/giac013.

[21] Uddandarao, D. P. (2026). Statistics and the Science of Causal Economics: A New Paradigm for Data Science. Deep Science Publishing. https://doi.org/10.70593/978-93-7185-895-3

[22] Ru, B., Sillah, A., Desai, K., Chandwani, S., Yao, L., & Kothari, S. (2024). Real-world data quality framework for oncology time to treatment discontinuation use case: Implementation and evaluation study. JMIR Medical Informatics, 12, e47744. doi: 10.2196/47744.

[23] Syed, R., Eden, R., Makasi, T., Chukwudi, I., Mamudu, A., Kamalpour, M., Kapugama Geeganage, D., Sadeghianasl, S., Leemans, S. J. J., Goel, K., Andrews, R., Wynn, M. T., Ter Hofstede, A., & Myers, T. (2023). Digital health data quality issues: Systematic review. Journal of Medical Internet Research, 25, e42615. doi: 10.2196/42615.

[24] Tong, L., Shi, W., Isgut, M., Zhong, Y., Lais, P., Gloster, L., Sun, J., Swain, A., Giuste, F., & Wang, M. D. (2024). Integrating multi-omics data with EHR for precision medicine using advanced artificial intelligence. IEEE Reviews in Biomedical Engineering, 17, 80–97. doi: 10.1109/RBME.2023.3324264.

[25] S. Mahajan, R. Devani and D. P. Uddandarao, "Predictive Bayesian Models for Measuring Marketing Campaign Impact," 2025 International Conference on Computational Engineering, Sensing Technology and Management (ICCETM), Sydney, Australia, 2025, pp. 1-6, doi: 10.1109/ICCETM66557.2025.11557686.

[26] van Os, H. J. A., Kanning, J. P., Wermer, M. J. H., Chavannes, N. H., Numans, M. E., Ruigrok, Y. M., van Zwet, E. W., Putter, H., Steyerberg, E. W., & Groenwold, R. H. H. (2022). Developing clinical prediction models using primary care electronic health record data: The impact of data preparation choices on model performance. Frontiers in Epidemiology, 2, 871630. doi: 10.3389/fepid.2022.871630.

[27] Yasrebi-de Kom, I. A. R., Dongelmans, D. A., de Keizer, N. F., Jager, K. J., Schut, M. C., Abu-Hanna, A., & Klopotowska, J. E. (2023). Electronic health record-based prediction models for in-hospital adverse drug event diagnosis or prognosis: A systematic review. Journal of the American Medical Informatics Association, 30(5), 978–988. doi: 10.1093/jamia/ocad014.