Data Subcard: Evaluating Privacy, Fairness, Quality, and Protection in Tabular Data, as Part of the System Cards Framework

Tadesse K. Bahiru
Computational Biomedicine Lab
Department of Computer Science
University of Houston, Texas, USA
tbahiru@central.uh.edu

Carlos Ordonez
Data-Intensive Parallel Algorithms for AI
Department of Computer Science
University of Houston, Texas, USA
carlos@central.uh.edu

Ioannis A. Kakadiaris
Computational Biomedicine Lab
Department of Computer Science
University of Houston, Texas, USA
ikakadia@central.uh.edu

***Abstract*—Medical datasets play a crucial role in advancing healthcare research and supporting clinical decision-making. At the same time, the reliability of responsible and accountable AI systems is directly dependent on the integrity and transparency of the datasets on which they are built. The data subcard implements the System Cards framework's data assessment dimension to evaluate tabular medical datasets across four criteria: privacy, fairness, quality, and protection. It combines data-level profiling with optional model-based diagnostics, selected to fit each dataset, to assess completeness, duplication, outliers, demographic disparities, re-identification risk, and compliance readiness. Applied to the UCI Heart Disease and Diabetes Readmission datasets, the method flags privacy risks, fairness imbalances, quality defects, and protection gaps that warrant review before modeling. The data subcard produces quantitative scores and visual summaries, providing a structured and interpretable mechanism for dataset accountability within the System Cards framework.**

***Index Terms*—Data metrics, medical data, fairness, scorecard, privacy, compliance, data quality, responsible AI**

I. Introduction

Artificial intelligence now supports a wide range of clinical decisions, from diagnosing diseases [1], forecasting patient deterioration [2], to tailoring treatments based on individual characteristics [3]. In radiology, AI tools flag subtle abnormalities for review [4]. In intensive care, machine-learning models can raise sepsis alerts hours before conventional scores detect danger [5]. In hospital operations, predictive analytics streamline bed assignment and operating-room scheduling [6]. These systems analyze large volumes of patient data quickly, surfacing patterns that may otherwise go unnoticed and often help reduce diagnostic delays. As they mature, well-tested AI tools are becoming trusted clinical aids that enhance decision-making and workflow efficiency.

Still, the performance of such systems depends heavily on the quality and integrity of the data [7] behind them. Routine electronic health records often contain missing values, inconsistent coding, and duplicates, which weaken analytic integrity and introduce bias [8]. A widely used risk stratification tool, for example, deprioritized care for high-need black patients by relying on healthcare cost as a stand-in for illness severity [9]. Beyond technical concerns, regulatory constraints introduce further complexity. Regulatory frameworks such as HIPAA [10] and GDPR [11] impose strict requirements for consent, de-identification, and data stewardship. Recent enforcement actions demonstrate that inadequate documentation or unclear consent can hinder AI initiatives, regardless of their technical accuracy [12]. These challenges underscore the need for a unified audit method that evaluates dataset quality, fairness, privacy, and protection as interconnected dimensions rather than treating them as isolated checks.

II. Related Work

Building accountable AI systems in healthcare requires datasets that are well-documented, structurally sound, privacy-aware, and fair. Gebru et al. [13] introduced datasheets for datasets, a structured documentation format to promote transparency around dataset provenance, collection, and intended use. Bahiru et al. [14] used the System Cards framework [15] to evaluate an AI data development scorecard that audits dataset documentation and development practices. Sambasivan et al. [16] introduced the Data Readiness Report, which assesses class noise, feature correlations, outliers, and the lineage of data operations to support transparent documentation and reproducibility. The AIDRIN framework [17] integrates data quality metrics, Markov model–based privacy evaluation, fairness analysis, and compliance with FAIR principles to provide a multi-dimensional assessment of dataset readiness. Gupta et al. [18] proposed the Data Quality Toolkit, which analyzes structured datasets using metrics such as label purity, class overlap, outlier rate, and feature relevance, and produces lineage-aware diagnostic outputs to support data quality improvement. Tibebu et al. [19] evaluate fairness gaps at the intersection of race and gender, and toolkits such as AI Fairness 360 [20] and Fairlearn [21] provide metrics and mitigation approaches for such disparities. Despite these contributions, most existing methods and frameworks remain narrow in scope, focusing on a single dimension such as quality, privacy, or fairness, rather than enabling a comprehensive evaluation. Many overlook demographic-specific fairness and fail to provide mechanisms for auditing regulatory compliance.

The data subcard implements the System Cards framework's [15] data assessment dimension for tabular medical datasets. It utilizes raw data inspection and model-based diagnostics to assess data privacy, fairness, quality, and protection. Unlike existing approaches, the entire evaluation runs in a local environment to preserve privacy, producing a unified, interpretable subcard supported by visual summaries.

III. Method

The data subcard evaluates datasets using four criteria drawn from the System Cards framework's data assessment and assurance dimensions: data privacy (C211), data fairness (C212), data quality (C213), and data protection (C411). For each criterion, we define measurable metrics and apply them to the dataset. The evaluation integrates structured data-level assessments with optional model-based diagnostics, depending on dataset characteristics. In the first stage, the raw dataset is evaluated to generate an initial subcard that visualizes its overall state. Users can then apply preprocessing and train a diagnostic model to capture additional properties such as label consistency and fairness across groups. The final subcard summarizes both data-level and model-based results through quantitative scores and visual representations. Designed to run locally, the system preserves data privacy while providing actionable insights with minimal setup. Once the evaluation is complete, users submit their subcard to the main System Cards framework for rating and certification. An overview of the method is shown in Fig. 1.

A. Data Privacy (C211)

Protecting patient privacy is essential when analyzing medical datasets. We assessed privacy risk along three complementary dimensions: detection of Protected Health Information (PHI), attribute-level re-identification risk, and group-level re-identification risk.

1) PHI Detection: Direct identifiers are detected using Microsoft Presidio (https://microsoft.github.io/presidio/), which combines pre-trained named-entity-recognition models with rule-based pattern matching to flag all 18 HIPAA Safe Harbor categories. For every dataset, Presidio returns the PHI type, the columns affected, and the proportion of records containing each identifier, thereby enabling systematic de-identification before further analysis.

2) Attribute-Level Re-Identification: Let $Q \subseteq \{1, \ldots, p\}$ be the indices of quasi-identifier attributes. For each $q \in Q$ we tabulate the frequency of every value v over the n records; values that appear only once (or very rarely) are marked high-risk, as an adversary observing $x_{iq} = v$ can single out an individual. The attribute-level risk for q is therefore proportional to the fraction of unique or near-unique values observed for that attribute.

3) Group-Level Re-Identification: To capture risk from combinations of quasi-identifiers, we draw a random subset $QQ' \subseteq Q$ and form equivalence classes G whose members share identical values on all attributes in Q''. The deterministic risk for a class is defined as $r_G = 1/|G|$, reflecting the probability of unique identification when an attacker knows the exact attribute combination. Because an attacker may exploit any subset of quasi-identifiers, we estimate worst-case risk using a Monte Carlo procedure [22]. IIn each of $M = 1,000$ iterations, we create a subset Q'_m by including each attribute in Q with independent probability $p_t = 0.5$. This value was selected to represent a neutral assumption, where an adversary is equally likely to include or exclude any given quasi-identifier, resulting in a balanced distribution of subset sizes. It is user-configurable to adapt to different threat models or dataset characteristics. After sampling, equivalence classes G_m are formed and the associated risks $r_{G,m}$ are recorded. The empirical distribution of $r_{G,m}$ is then summarized by its mean, standard deviation, and maximum, providing a robust estimate of re-identification risk.

B. Data Fairness (C212)

Fairness is examined using three dataset-level metrics. Let s denote a sensitive attribute such as gender, age, or race, and let g represent one of its possible groups (for example, $s =$ female). Let Y be the outcome variable taking classes c_k. Representational rate is defined as $P(s = g)$, measuring the share of records in each group. Outcome rate evaluates class balance as $P(Y = c_k \mid s = g)$ for every group. Intersectional outcome rate extends this idea to multiple sensitive attributes: for attributes $s_1, \ldots, s_K$ with values $a_1, \ldots, a_K$, the rate is given by $P(Y = c_k \mid s_1 = a_1, s_2 = a_2, \ldots, s_K = a_K)$. Together, these probabilities reveal both single-attribute and compound disparities in outcome distributions.

C. Data Quality (C213*)

Ensuring the integrity and reliability of medical datasets is essential for producing valid analyses and robust models. The data quality in the System Cards framework is limited to the labels of the dataset; in this method, we extend it to encompass the entire data quality measure. Given a dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$ with n records and p features, where $x_i = (x_{i1}, \ldots, x_{ip})$ and y_i is the ground-truth label, we applied the following metrics. *Completeness* is the proportion of non-missing values per feature x_{ij}. *Data-type validation* confirms that each x_{ij} matches the expected clinical format. *Duplicate detection* compares entire feature vectors x_i to identify repeated records. *Feature correlation* uses Pearson coefficients for continuous features and CCramér'sV for categorical features to flag multicollinearity. *Outlier detection* applies both univariate (inter-quartile range) and multivariate (Mahalanobis distance) tests. *Class imbalance* is also assessed from the empirical distribution of y_i across classes.

D. Data Protection (C411)

Data protection is a critical component of responsible dataset use in research and clinical practice and is closely tied to regulatory compliance. In our method, this dimension is assessed through a structured rule-based module that reviews uploaded dataset documentation against predefined validation checks. The module examines six criteria: institutional review

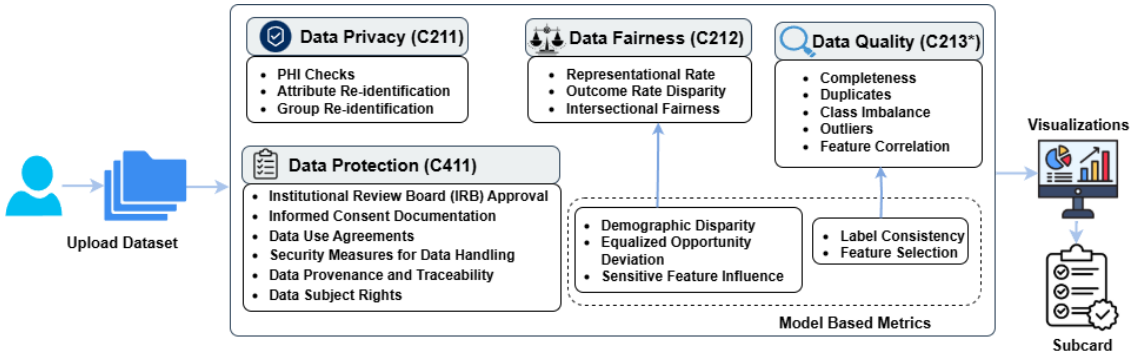

Fig. 1. Evaluation workflow and subcard evaluation. Uploaded datasets are assessed across four criteria (Data Quality, Fairness, Privacy, and Protection). The process outputs both a visual card and a structured JSON report, providing actionable insights for dataset accountability and transparency.

board approval, informed consent, data use agreements, security measures, data provenance, and data subject rights. For each criterion, the system verifies the presence and validity of the required information (for example, the correct IRB number format or the inclusion of a data use clause) and assigns a binary score of 1 if the requirement is satisfied or zero if not. The overall data protection score is computed as the proportion of criteria met, and the results are presented in a compliance card that reports individual scores together with recommendations for addressing any gaps.

E. Model-Based Assessment

After the data-level evaluation, users who seek additional diagnostic insights can train a predictive model to evaluate the dataset from a modeling perspective. The method includes an integrated preprocessing module that addresses common data quality issues, such as missing values, incorrect data types, and categorical encoding, to ensure the dataset is ready for training. It supports several interpretable diagnostic models, including XGBoost, LightGBM, and logistic regression. Users can evaluate and compare model performance, select the most suitable model for their dataset, and configure key hyperparameters, such as the learning rate, number of estimators, and maximum tree depth. Once trained, the selected model serves as a diagnostic instrument to identify mislabeled records, unstable feature behavior, and potential biases associated with sensitive attributes.

Label consistency is evaluated by comparing the model's predictions with the actual labels, allowing the identification of mislabels or ambiguous entries. Feature relevance is measured using model-derived importance scores to identify which variables consistently contribute to accurate predictions and to flag those with unstable influence across different training subsets. Fairness is assessed using group-based performance metrics, including demographic disparity, equalized opportunity deviation, and the influence of sensitive attributes on model decisions.

a) Demographic Disparity [23]: Demographic disparity is measured by comparing the model's predicted positive rates across different groups defined by a sensitive attribute s. For each group g and class c, we compute:

$$\mathrm{PPR}_{g,c} = P(\hat{Y} = c \mid s = g).$$

The Demographic Disparity (DD) for sensitive attribute s is defined as:

$$\mathrm{DD}_s = \max_{g,g',c} \left| \mathrm{PPR}_{g,c} - \mathrm{PPR}_{g',c} \right|.$$

b) Equalized Opportunity Deviation [24]: We evaluate the model's consistency in identifying true positive cases across groups. For each group g and class c, we compute the true positive rate:

$$\mathrm{TPR}_{c,g} = P(\hat{Y} = c \mid Y = c,\, s = g).$$

The Equalized Opportunity Deviation (EOD) for sensitive attribute s is:

$$\mathrm{EOD}_s = \max_{g,g',c} \left| \mathrm{TPR}_{c,g} - \mathrm{TPR}_{c,g'} \right|.$$

c) Sensitive Feature Influence: To understand how sensitive attributes influence model predictions, we utilize SHAP values [25]. For each sensitive attribute s, we compute:

$$\mathrm{SI}_s = \frac{1}{n} \sum_{i=1}^{n} \left| \phi_{i,s} \right|.$$

Here $\phi_{i,s}$ is the SHAP value of attribute s for sample i. Higher SI_s values indicate stronger dependence on the sensitive attribute.

F. Data Subcard

The Data Subcard is the primary outcome of the evaluation, consolidating results from quality, fairness, privacy, and protection assessments into a single, interpretable report. Each criterion is represented by a numerical score, a color-coded rating, explanatory remarks, and targeted improvement suggestions. Scores are normalized to the interval $[0, 1]$, where higher values indicate stronger alignment with responsible data practices. Fig. 2 (L) illustrates how the System Cards framework employs this color scheme across all accountability dimensions, while Fig. 2 (R) presents the Data Subcard template, which applies the same scoring logic to dataset evaluation.

Data Subcard	
Basic Metadata	
Dataset Name:	[Name of the dataset]
Version:	[Version of the dataset]
Creation Date:	[Date created]
Source:	[Origin of the dataset]
Dataset Evaluation	
Data Privacy (C211) :	Score [X.XX] – Color: Remark
Data Fairness (C212):	Score [X.XX] – Color: Remark
Data Quality (C213):	Score [X.XX] – Color: Remark
Data Protection (C411):	Score [X.XX] – Color: Remark
Improvement Suggestions	
Data Privacy (C211):	[Suggestion]
Data Fairness (C212):	[Suggestion]
Data Quality (C213):	[Suggestion]
Data Protection (C411):	[Suggestion]

Fig. 2. System cards and their subcard implementation. (L) System Cards framework visualization [15] with four concentric circles; criteria shown as arcs colored from red (worst) to green (best). (R) Data Subcard template for the four data assessment criteria.

IV. Experimental Evaluation

A. Experimental Setup

The evaluation method was implemented in Python using `pandas` for data handling, `NumPy` for numerical operations, `scikit-learn` for model training and evaluation, and `Matplotlib` and `Seaborn` for visualization. The pipeline begins with direct data-level assessments of privacy, fairness, quality, and protection, and compiles the results into a structured subcard that integrates quantitative metrics with concise visual summaries. For model-based diagnostics, we adopted `XGBoost` after preliminary comparisons showed it consistently outperformed alternative algorithms in both predictive accuracy and interpretability. The final output of the evaluation is a structured card that visualizes the dataset status with color-coded scores and explanatory remarks. Scoring follows the Likert-style method defined in the System Cards framework: $t_i \geq 0.80$ appears in green to indicate strong alignment, $0.60 \leq t_i < 0.80$ appears in yellow to indicate adequacy with room for improvement, and $t_i < 0.60$ appears in red to highlight deficiencies that compromise transparency, reproducibility, or overall quality. These thresholds were calibrated through pilot evaluations.

B. Datasets

We evaluated the subcard using two medical datasets from the UCI Machine Learning Repository. First, the Diabetes readmission dataset (https://archive.ics.uci.edu/ml/datasets/Diabetes+) includes over 100K inpatient records with demographic, laboratory, medication, and discharge details. Second, the Heart disease dataset (https://archive.ics.uci.edu/ml/datasets/Heart+Disease) comprises four sources in one repository, with experiments typically focusing on the Cleveland subset of 303 records and 13 features. The UCI datasets used here are suitable for prototyping, but are simpler than real hospital datasets, which are inaccessible, poorly curated, and fragmented across multiple data dictionaries.

C. Results

The privacy of the datasets was assessed by scanning for protected health information; no direct identifiers were detected in the Diabetes or Heart Disease files. Attribute-level inspection then showed virtually no single-field risk: gender and race values are never unique, and age is unique in only 0.3% of Heart Disease records. When common demographic fields are combined, the chance of isolating an individual increases. The mean re-identification probability is 0.08 in Diabetes and 0.25 in Heart Disease, and each dataset contains at least one record that can be singled out, as indicated by a maximum risk of 1.0. Table I summarizes the risks of re-identification, showing that most records are protected; however, rare combinations of attributes pose an exposure risk.

TABLE I
PRIVACY RISK METRICS SUMMARIZING ATTRIBUTE-LEVEL UNIQUENESS AND GROUP-LEVEL RE-IDENTIFICATION PROBABILITIES.

Metric	Attribute	Diabetes	Heart Disease
Attribute-Level Risk	Gender	0.000	0.000
	Race	0.000	0.000
	Age	0.000	0.003
Group-Level Risk	Mean	0.082	0.245
	Std Dev	0.216	0.304
	Max	1.000	1.000

Fairness assessment of the two medical datasets reveals distinct demographic patterns. The outcome distributions in the Heart Disease dataset reveal apparent disparities by gender and age. Female records concentrate in the no disease category, while male records are more evenly spread across all severity levels, and patients older than sixty appear in every severity class. The intersection matrix in Fig. 3 shows that young women form the healthiest group, whereas older men experience the broadest range of outcomes. Complementary model-based metrics provide a broader context for these findings. Table II shows that age drives the largest demographic gap in both datasets, with race and gender gaps minimal in Diabetes but a pronounced gender effect in Heart. Shapley attributions explain this contrast by showing that the Diabetes model derives most of its predictive power from age. In contrast,

the Heart model relies more heavily on gender, mirroring the distributional imbalances highlighted in the data-level analysis.

TABLE II
FAIRNESS METRICS AND SHAP-BASED INFLUENCE SCORES FOR SENSITIVE ATTRIBUTES.

Metric	Attribute	Diabetes	Heart Disease
Demographic Disparity	Race	0.112	–
	Gender	0.011	0.521
	Age	0.388	0.609
Equalized Opportunity	Race	0.360	–
	Gender	0.037	0.307
	Age	0.434	0.182
SHAP Sensitive Influence	Race	0.143	–
	Gender	0.221	0.671
	Age	0.636	0.329

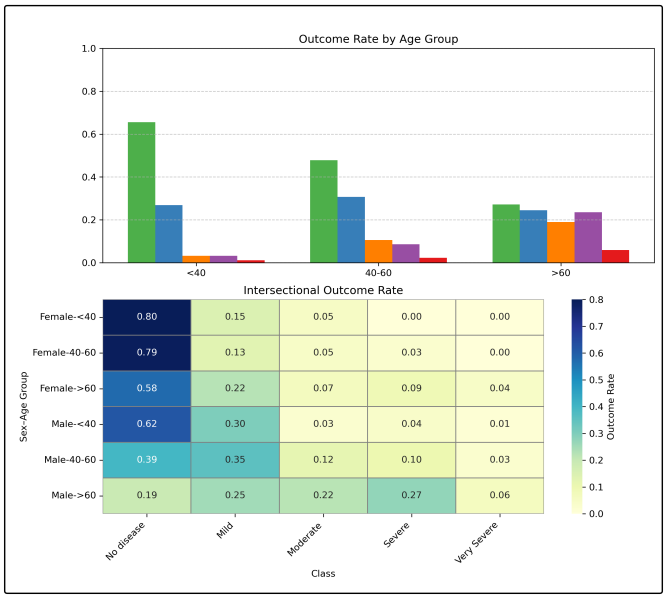

Fig. 3. Outcome distribution in the Heart Disease dataset by age (top) and sex–age groups (bottom), showing disparities across demographic subgroups.

The quality of the datasets is evaluated using the metrics defined in Section III (C). As Fig.4 shows, the Heart Disease data are almost complete, with missing values concentrated in fluoroscopy vessel count, thalassemia type, and ST-segment slope. The label distribution is moderately skewed toward the ”Disease” class, while the ”very Severe” class represents fewer than 5% of the records. Cholesterol has the highest outlier fraction, followed by resting blood pressure and vessel count. Correlation analysis reveals only modest associations, with the largest being a positive correlation between age and vessel count. Fig.5 presents the class imbalance and anomaly profile for Diabetes encounters. Readmissions marked ”O” exceed 50% of cases, whereas readmissions at 30 days account for approximately one in ten. Outliers cluster in encounter-frequency variables, with outpatient visits showing the highest anomaly rate, and emergency and inpatient visits contributing smaller shares. Laboratory tests, diagnoses, and medication counts show fewer extremes.

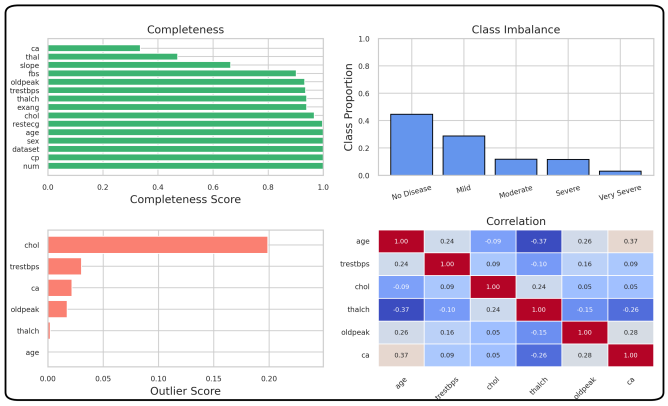

Fig. 4. Data quality metrics for the Heart Disease dataset, including completeness, class imbalance, outlier scores, and feature correlation.

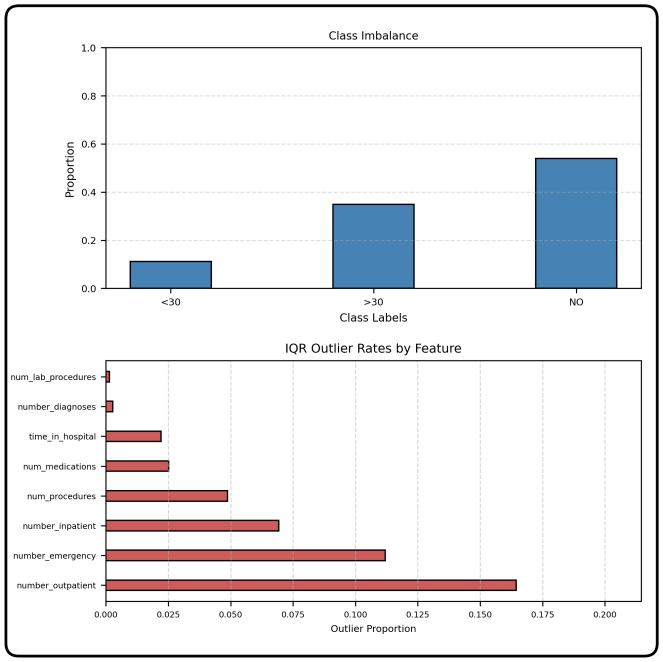

Fig. 5. Class imbalance and outlier rates in the Diabetes dataset, with skewed distributions and high anomaly concentrations in visit-related features.

D. Discussion

The evaluation results indicate that the datasets examined here provide a strong foundation for analysis, with most variables well-defined and complete. However, missing values in attributes such as vessel count and thalassemia, imbalanced outcome distributions, and outliers in clinical measures such as cholesterol and blood pressure indicate areas that require targeted preprocessing before model development. Fairness assessments reveal measurable differences across demographic groups, particularly in the Heart Disease dataset, where model output shows a higher sensitivity to gender and age. Privacy evaluations indicate the absence of direct identifiers and low uniqueness at the attribute level. Yet, the combination of quasi-identifiers still results in moderate re-identification risks in

some instances. These results suggest that, while the datasets are broadly suitable for analytical use, addressing the identified gaps will improve both reliability and trustworthiness.

V. Conclusion

We presented the Data Subcard, a locally executable method that applies the System Cards framework to audit tabular datasets across privacy, fairness, quality, and protection. The approach integrates structured data-level checks with model-based diagnostics, yielding interpretable artifacts that support pre-model risk assessment. Empirical evidence from public clinical datasets reveals that subcard surfaces, re-identification exposure, demographic disparities, and data defects can erode downstream validity, thereby improving dataset accountability before model training and evaluation.

Future work will complete the protection dimension with operational compliance checks and document verification, refine scoring through calibration and uncertainty quantification, and expand evaluation to longitudinal and heterogeneous settings to strengthen external validity. We also plan to conduct sensitivity analyses under alternative threat models and subgroup definitions, as well as multi-institution benchmarking to assess generalizability. Additionally, we aim to integrate the framework more tightly with institutional governance workflows through a reproducible local package and programmatic interfaces. These steps will advance the subcard from a practical auditing tool to a validated component of data stewardship and assurance.

Acknowledgment

This research was supported in part by the National Institutes of Health (NIH) under award number UM1TR004539 and by the Hugh Roy and Lillie Cranz Cullen Endowment Fund at the University of Houston. All statements of fact, interpretations, opinions, and conclusions expressed in this paper are solely those of the authors and should not be construed as representing the official views or policies of the sponsors.

The authors acknowledge the use of AI-assisted tools for grammar correction and language polishing.

References

[1] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Y. Yee, K. Zhang, Y. Zhang, G. Flores, G. Duggan, J. Irvine, Q. V. Le, J. Dean, and G. S. Corrado, "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, p. 18, 2018.

[2] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: Past, present and future," *Stroke and Vascular Neurology*, vol. 2, no. 4, pp. 230–243, 2017.

[3] E. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, pp. 44–56, 2019.

[4] M. Benjamens, P. Dhunnoo, and B. Meskó, "The state of artificial intelligence–based FDA-approved medical devices and algorithms: an online database," *NPJ digital medicine*, vol. 3, no. 1, p. 118, 2020.

[5] M. Moor, B. Rieck, M. Horn, C. R. Jutzeler, and K. Borgwardt, "Early prediction of sepsis in the ICU using machine learning: A systematic review," *Frontiers in Medicine*, vol. 8, p. 607952, 2021.

[6] F. Dexter and R. H. Epstein, "Operating room scheduling and postanesthesia care unit staffing for surgeries with times estimated from historical data," *Anesthesia & Analgesia*, vol. 127, no. 2, pp. 585–593, 2018.

[7] C. Ordonez and J. García-García, "Referential integrity quality metrics," *Decision Support Systems*, vol. 44, no. 2, pp. 495–508, 2008.

[8] K. B. Bayley, T. Belnap, L. A. Savitz, A. L. Masica, N. Shah, N. S. Fleming, and T. S. Carey, "Challenges in using electronic health record data for CER: Experience of four learning organizations," *Medical Care*, vol. 51, no. 8 Suppl 3, pp. S62–S69, 2013.

[9] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[10] U.S. Department of Health and Human Services, "Standards for privacy of individually identifiable health information," 65 Fed. Reg. 82462, codified at 45 CFR Part 160 and 164, 2000.

[11] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation)," Official Journal of the European Union, L119, pp. 1–88, 2016.

[12] National Institute of Standards and Technology, "Artificial intelligence risk management framework (AI RMF 1.0)," U.S. Department of Commerce, Technical Report NIST AI 100-1, 2023.

[13] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[14] T. K. Bahiru, H. Tibebu, and I. A. Kakadiaris, "AI data development: A scorecard for the system card framework," in *Proc. the International Conference on Information Technology and Artificial Intelligence (ITAI 2025)*, Gurgaon, India, January 17–19, 2025.

[15] F. Gursoy and I. A. Kakadiaris, "System cards for AI-based decision-making for public policy," *arXiv preprint arXiv:2203.04754*, 2022.

[16] S. Afzal, C. Rajmohan, M. Kesarwani, S. Mehta, and H. Patel, "Data readiness report," in *in Proc. IEEE International Conference on Smart Data Services (SMDS)*, Chicago, IL, July 5–10, 2021, pp. 42–51.

[17] K. Hiniduma, S. Byna, J. L. Bez, and R. K. Madduri, "AI data readiness inspector (AIDRIN) for quantitative assessment of data readiness for AI," in *Proc. the International Conference on Scientific and Statistical Database Management*, New York, NY, July 1–3, 2024.

[18] N. Gupta, H. Patel, S. Afzal, N. Panwar, R. Mittal, S. Guttula, A. Jain, L. Nagalapatti, S. Mehta, S. Hans, P. Lohia, A. Aggarwal, and D. Saha, "Data quality toolkit: Automatic assessment of data quality and remediation for machine learning datasets," in *Proc. the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Virtual Event (Singapore), August 14–18, 2021.

[19] H. Tibebu and I. A. Kakadiaris, "Addressing the intersection of race and gender in AI bias," in *Proc. the 13th SAI Computing Conference (SAI)*, London, United Kingdom, June 19–20, 2025.

[20] R. K. E. Bellamy, K. Dey, M. Hind, S. Hoffman, S. Houde, K. Kannan, Q. V. Liao, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4:1–4:15, 2019.

[21] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker, "Fairlearn: A toolkit for assessing and improving fairness in AI," Microsoft Research, Tech. Rep. MSR-TR-2020-32, 2020.

[22] N. Metropolis and S. Ulam, "The monte carlo method," *Journal of the American Statistical Association*, vol. 44, no. 247, pp. 335–341, 1949.

[23] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proc. SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, Australia, August 10–13, 2015, pp. 259–268.

[24] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proc. International Conference on Neural Information Processing Systems*, Barcelona, Spain, Dec. 5–10, 2016, pp. 3315–3323.

[25] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. International Conference on Neural Information Processing Systems*, Long Beach, CA, Dec. 4–9, 2017, pp. 4765–4774.