A framework for assessing algorithmic discrimination risks in training data: a case of pediatric type 1 diabetes

Ioannis Bilionis, MSc, Ricardo C Berrios, BA, Antonio de Arriba Muñoz, MD, Luis Fernandez-Luque, PhD, Carlos Castillo, PhD

JAMIA Open, 2026, **9**(4), ooag125
https://doi.org/10.1093/jamiaopen/ooag125
Research and Applications

A framework for assessing algorithmic discrimination risks in training data: a case of pediatric type 1 diabetes

Ioannis Bilionis, MSc[1,2,*], Ricardo C. Berrios, BA[1], Antonio de Arriba Muñoz, MD[3], Luis Fernandez-Luque, PhD[1], Carlos Castillo, PhD[2,4]

[1]Adhera Health, Santa Cruz, CA, 95060, United States
[2]Department of Engineering, Universitat Pompeu Fabra, Barcelona, 08018, Spain
[3]Department of Pediatric Endocrinology, University Hospital "Miguel Servet", Zaragoza, 50009, Spain
[4]ICREA, Catalonia, 08010, Spain
*Corresponding author: Ioannis Bilionis, MSc, Department of Engineering, Universitat Pompeu Fabra, Carrer Roc Boronat, 138, Barcelona, Catalonia 08018, Spain (ibilionis@adherahealth.com)

Abstract

Objectives: **To develop a generalizable framework for identifying algorithmic discrimination risks arising from subgroup imbalances in machine learning training data, with relevance to medical informatics applications where heterogeneous real-world data can bias model behavior.**

Materials and Methods: **We introduce a discrimination risk assessment framework for training datasets, a 4-step methodology integrating: (1) controlled representation sampling, (2) ensemble-based model training, (3) multilevel subgroup disparity quantification, and (4) mitigation-oriented interpretation. The framework systematically perturbs training subgroup composition while holding evaluation sets fixed to isolate representation effects. Validation was performed across 7 publicly available pediatric type 1 diabetes datasets using sociodemographic variables and continuous glucose monitoring data to assess robustness under heterogeneous data sources.**

Results: **Representation balance alone does not guarantee stable or equitable model outputs. Ensemble analyses revealed subgroup-dependent volatility, with some groups consistently contributing to model generalization, while others inducing instability despite increased representation. These findings demonstrate structural sensitivity to data composition that is not captured by standard performance-only evaluations.**

Discussion: **The framework exposes disparities and instability patterns that remain hidden under single-model or balanced-data evaluations. By quantifying representation-driven behaviors and subgroup-specific training value, it offers a diagnostic tool for understanding data-induced risks prior to model deployment.**

Conclusion: **This work provides a methodology for assessing discrimination risks in training datasets used in medical informatics. By integrating data composition effects, prediction stability, and subgroup-level disparity analysis, the framework supports more reliable and transparent development of machine learning systems across diverse clinical and nonclinical contexts.**

Clinical trial registration: **This research did not involve any new clinical trial. All analyses were performed on data from previously registered clinical trials, as listed below:**

1) **PEDAP dataset**: The source of the data is the PEDAP Trial Study Group. (2024). PEDAP Public Dataset (Release 4). Retrieved from https://public.jaeb.org/dataset/599. The analyses content and conclusions presented herein are solely the responsibility of the authors and have not been reviewed or approved by PEDAP Trial Study Group. *ClinicalTrials.gov* Identifier: NCT04796779; registered March 15, 2021.
2) **IOBP2 RCT dataset**: The source of the data is the Insulin Only Bionic Pancreas Pivotal Trial. Retrieved from https://public.jaeb.org/dataset/579. The analyses content and conclusions presented herein are solely the responsibility of the authors and have not been reviewed or approved by the Bionic Pancreas Research Group or Beta Bionics. *ClinicalTrials.gov* Identifier: NCT04200313; registered December 16, 2019.
3) **DCLP5 dataset**: The source of the data is the University of Virginia. DCLP5 Dataset. Retrieved from http://public.jaeb.org/dataset/535. The analyses content and conclusions presented herein are solely the responsibility of the authors and have not been

Received: January 2, 2026. **Revised:** March 30, 2026. **Accepted:** June 18, 2026

reviewed or approved by the University of Virginia. Associated *ClinicalTrials.gov* Identifier: NCT03844789; registered February 18th, 2019

4) **DCLP3 dataset**: The source of the data is the University of Virginia. The International Diabetes Closed Loop (iDCL) trial. DCLP3 Public Dataset (Release 3). Retrieved from http:///public.jaeb.org/dataset/573. The analyses content and conclusions presented herein are solely the responsibility of the authors and have not been reviewed or approved by the University of Virginia. *ClinicalTrials.gov* Identifier: NCT03591354; registered July 19, 2018.
5) **CITY dataset**: The source of the data is the Jaeb Center of Health Research. CGM Intervention in Teens and Young Adults with T1D (CITY Public Dataset). Retrieved from http://public.jaeb.org/dataset/565. The analyses content and conclusions presented herein are solely the responsibility of the authors and have not been reviewed or approved by the Jaeb Center of Health Research. *ClinicalTrials.gov* Identifier: NCT03263494; registered August 28, 2017.
6) **SENCE public dataset**: The source of the data is the Jaeb Center of Health Research. SENCE Public Dataset. Retrieved from http://public.jaeb.org/dataset/554. The analyses content and conclusions presented herein are solely the responsibility of the authors and have not been reviewed or approved by the Jaeb Center of Health Research. *ClinicalTrials.gov* Identifier: NCT02912728; registered September 23, 2016.

Key words: machine learning; health equity; bias; pediatrics; diabetes mellitus type 1.

Background and significance

Machine learning (ML) is routinely deployed for decision support in high-stakes domains such as health care, finance, and criminal justice.[1] While these systems promise gains in efficiency and precision, they also risk reinforcing structural inequalities through biases in the underlying data or through dynamics that emerge once deployed, making rigorous subgroup-level evaluation essential.[2] In health care, disparities in model performance across racial, ethnic, and socioeconomic groups are well documented, often reflecting historical inequities encoded in clinical records.[3] Such persistent performance gaps constitute statistical group discrimination.[4] Despite growing awareness, there remains a lack of standardized frameworks to identify, measure, and mitigate sociodemographic biases that can be traced to ML training data, which is a foundational step toward the development of equitable artificial intelligence (AI) systems.

These challenges are particularly salient in pediatrics, where technical, ethical, and data-quality limitations remain insufficiently addressed, and where most AI/ML work is concentrated in high-income countries, while efforts in low- and middle-income countries remain largely exploratory.[5] Pediatric clinicians highlight the importance of training on health disparities and implicit bias, yet structured education on AI's risks is limited.[6–9] Despite advances such as federated learning and explainable AI, persistent challenges—including limited data diversity, poor generalizability, and unresolved ethical concerns—underscore the need for more rigorous approaches and frameworks for pediatric AI.[10–13]

Bias plays a critical role in the management of type 1 diabetes (T1D), a chronic condition requiring lifelong management and showing substantial variation in incidence, outcomes, and access to care across demographic groups.[14–16] Black and Hispanic youth experience higher rates of diabetic ketoacidosis and face structural barriers to advanced diabetes technologies.[17,18] Although continuous glucose monitoring (CGM) and insulin pumps have improved diabetes management,[19] adoption remains uneven,[20] with non-Hispanic White and privately insured populations benefiting the most. These disparities contribute to differences in glycemic control, complication rates, and hospitalization.[21,22] Socioeconomic factors—including insurance, education, and family support—further shape disease burden and treatment outcomes.[23–26]

Existing bias-assessment approaches in health-care ML remain fragmented. Foundational studies and broad ethical reviews[27–29] retrospectively highlight how proxy variables and EHR disparities propagate downstream clinical biases. In response, governance frameworks such as ACCESS-AI,[30] BE-FAIR,[31] GUIDE,[32] HEALTH-ML,[33] and emerging clinical implementation guidelines[34–36] have introduced standardized documentation, physician perspectives, and qualitative equity audits. Concurrently, numerous technical methodologies and recent systematic evaluations[37–44] predominantly focus on quantifying post hoc performance disparities (eg, error rate gaps) and applying downstream mitigations like algorithmic reweighting or dataset balancing. While these existing works successfully establish retrospective auditing, qualitative reporting, and downstream corrections, they largely evaluate static parity metrics and treat the initial training-data composition as a fixed artifact rather than a dynamic variable.

To address these gaps, our framework shifts the paradigm from downstream correction to upstream, data-centric diagnosis. Unlike retrospective audits or qualitative checklists, our methodology explicitly isolates how training-data composition fundamentally drives model behavior prior to clinical deployment. By integrating controlled representation sampling as a pretraining diagnostic, we uniquely expose how shifting subgroup prevalence impacts dynamic prediction stability and systematic arbitrariness across minority groups.

Demographic imbalances introduced through data collection practices or clinical inclusion criteria remain a major barrier to equitable ML systems.[45–47] Even when datasets appear balanced, disparities may persist due to differences in model behavior, feature interactions, or decision boundaries.[48–51] New legal frameworks, such as Article 10 of the European AI Act (EU Regulation 2024/1689), mandate appropriate data governance and assessment practices. Understanding how training-data composition influences performance and stability is therefore essential for identifying discrimination risks.

To address this gap, we introduce a generalizable framework for assessing discrimination risks, using T1D as an application case. The framework examines how subgroup representation,

prevalence, and training stability shape model behavior through 4 structured stages: Controlled Representation Sampling, Multiple Model Training, Subgroup Disparity Analysis, and Risk Interpretation & Mitigation.

By generating multiple variations of the available data, the framework systematically evaluates representation-driven risks and performance dynamics across demographic groups. By systematically varying subgroup representation and analyzing performance outcomes, and model arbitrariness,[52] that is, the proportion of test samples receiving conflicting predictions across repeated bootstrap model instances, we identify populations with reduced "training value,"[53,54] where models struggle to learn stable or accurate patterns. This approach highlights the limitations of conventional dataset balancing and uncover how inconsistencies in model performance can persist despite apparent representational equity. Beyond technical evaluation, our framework introduces a novel dual focus: it equips data providers with guidelines to validate datasets prior to dataset release, and it offers AI developers practical routines for subgroup-sensitive validation prior to model deployment.

Methods

We propose a 4-step *discrimination risk assessment framework* for systematic evaluation of model performance and stability as illustrated in Figure 1, which presents a flow diagram showing four connected stages for assessing discrimination risks in machine learning training datasets: Controlled Representation Sampling, Multiple Model Training, Subgroup Disparity Analysis, and Risk Interpretation and Mitigation. The key methodological steps are summarized in Table 1, while the details of the proposed validation framework are described in the following subsections. A central motivation of the framework is to inform both model evaluation and upstream data design, offering practical guidance for recruitment or sampling strategies before large-scale datasets are even constructed. To facilitate adoption, a complete checklist and a reporting template are provided in Supplementary Material S4 for use by third parties.

Step 1: Controlled representation sampling

We propose a structured, representation-aware sampling pipeline designed to isolate the effects of subgroup representation on model performance while minimizing confounding factors. Prior to controlled sampling, we recommend integrating a preliminary outlier detection step (eg, utilizing *Z*-score, interquartile range or isolation forests). In complex medical datasets, filtering extreme physiological anomalies ensures that rare outliers do not disproportionately skew the representation dynamics of sparse subpopulations. To achieve this, we generate $K = 50$ fixed test sets for each dataset, meeting 3 key design criteria: (1) balanced subgroup proportions (eg, equal numbers of subjects per subgroup) to facilitate valid between-group comparisons; (2) prevalence-aligned class distributions, ensuring clinical outcome frequencies were preserved across subgroups; and (3) equivalence in mean CGM time series lengths across subgroups to control for differences in monitoring exposure. These test sets are held constant throughout all experiments to maintain evaluation consistency and avoid artifacts due to test distribution shifts or unequal data granularity.

For training-data preparation, subgroup representation is systematically varied across 5 levels (0%, 25%, 50%, 75%, and 100% of the training cohort) while preserving the original outcome prevalence within each split (see Figure 2). For each representation level and subgroup comparison (such as race, insurance status, or income level), we create 50 distinct stratified training sets through random sampling, ensuring that outcome class balance remained intact to prevent artificial bias. This sampling approach disentangles the influence of subgroup inclusion from class distribution effects, enabling a robust, interpretable analysis of how variations in training-data composition affect downstream model behavior. The controlled stratification by subgroup distribution, outcome prevalence, and CGM data

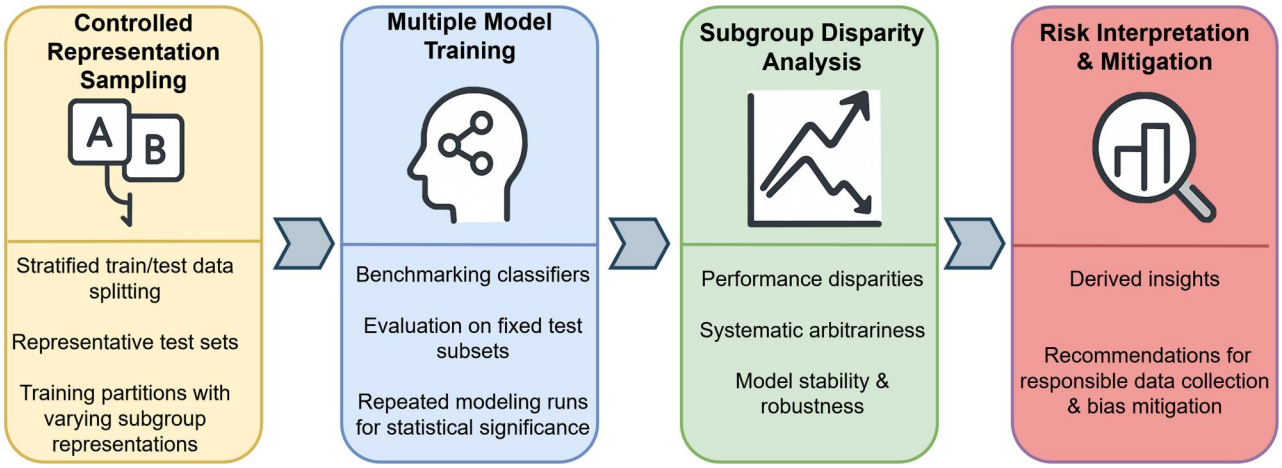

Figure 1. Overview of the proposed validation framework for assessing discrimination risks in machine learning training datasets comprised of 4 key stages: Controlled Representation Sampling, Multiple Model Training, Subgroup Disparity Analysis and Risk Interpretation & Mitigation.

Table 1. Discrimination risk assessment framework—checklist.

Step	Checklist items
1. Representation sampling	□ Define target subgroups □ Fix test sets with balanced subgroup sizes □ Align outcome prevalence across subgroups □ Match data quality (eg, CGM duration) □ Vary training representation (0%-100%) □ Repeat sampling *K* times
2. Model training	□ Select diverse algorithms □ Standardize preprocessing □ Train multiple models per configuration □ Store predictions and metrics
3. Disparity analysis	□ Compute subgroup performance metrics □ Assess self-consistency □ Quantify systematic arbitrariness □ Compare across representation levels
4. Risk interpretation	□ Identify high training-value subgroups □ Flag destabilizing subgroups □ Derive mitigation actions □ Document tradeoffs

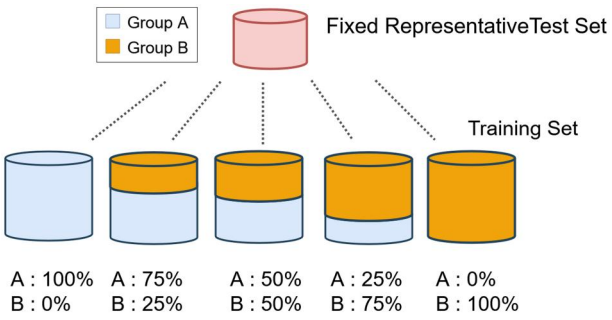

Figure 2. Schematic overview of the controlled stratified sampling strategy used to construct evaluation-ready train/test datasets. The figure illustrates five training dataset configurations based on different A/B subgroup compositions: 100%-0%, 75%-25%, 50%-50%, 25%-75%, and 0%-100%. Across all five configurations, the test dataset composition remains fixed, allowing evaluation of model performance under controlled variations in training subgroup representation.

quality provided a reproducible, clinically grounded foundation for subsequent modeling phases.

Step 2: Multiple model training

To evaluate model performance and model variability under controlled data compositions, the modeling phase involves training ensembles of classifiers for each unique training configuration. This enables robust estimation of both average performance and prediction stability. For each training condition (ie, a specific level of subgroup representation), the idea is to generate multiple independent models using randomized initializations and sampling seeds. This strategy[51] captures both stochastic noise (from training variability) and structural effects (from data composition), allowing the framework to move beyond single-point estimates.

We implement this modeling step across multiple supervised learning algorithms representing a range of complexity and inductive bias (eg, linear, tree-based, probabilistic). For each algorithm and training condition, we compute model-level metrics such as Area Under the Curve (AUC), precision-recall, and class-wise performance, and then aggregate these metrics across the ensemble to derive confidence intervals and behavioral distributions.

Step 3: Subgroup disparity analysis

To evaluate how subgroup representation influences model behavior, we implemented a multilevel disparity analysis framework that incorporates both predictive performance and model robustness. This process begins by aggregating evaluation outputs from ensembles of trained models under varied subgroup representation levels. For each configuration, we derive performance distributions stratified by subgroup and compute inter-group differences in key classification metrics. These differences are treated not as single-point gaps but as statistically grounded disparities derived from repeated model evaluations, allowing us to assess both magnitude and stability of group-level performance across conditions.

In parallel, we analyze prediction instability within subgroups to assess model consistency, focusing on metrics that reflect model arbitrariness and internal agreement across bootstrap replicates.[52] Instability is operationalized through distributional comparisons of prediction outputs and classification decisions, enabling fine-grained identification of groups for which the model exhibits unstable or sensitive behavior. These analyses are conducted across all sociodemographic dimensions varied in the training data, with fixed test distributions ensuring confounding factors are minimized. The resulting outputs—capturing both disparities and model stability—form the basis for downstream risk interpretation and inform whether certain subgroups contribute reliably to model generalization or exhibit signs of structural marginalization.

Step 4: Risk interpretation and mitigation

The final phase of the framework integrates the outputs from performance analysis, model consistency evaluation, and training value estimation to generate a composite view of discrimination risk. Rather than treating these metrics as isolated findings, this step synthesizes them to identify subgroup-specific patterns that influence model generalizability and reliability. This includes mapping subgroups along axes of predictive contribution, model stability, and representation level to detect cases where model robustness is disproportionately affected by specific population strata. The integration process is guided by an evidence-driven schema that flags subgroups for targeted

review or intervention based on their quantitative profile across prior analyses.

This synthesized risk profile is then used to support principled decisions around data curation and model refinement. By linking each subgroup's inclusion to its empirical impact on model performance and stability, stakeholders can prioritize high-value populations during dataset design, particularly in scenarios where limited resources require strategic sampling. The framework also supports identification of low-value or destabilizing subgroups—those whose inclusion contributes minimal or negative impact—prompting further investigation into potential data-quality issues or confounding sources of bias. This interpretive phase transforms analytic insights into practical guidance for improving equity and trustworthiness in clinical ML systems.

Results

T1D datasets

We applied the discrimination risk assessment framework to 6 publicly available US-based T1D datasets with a total of 19k weekly data points and $n = 690$ participants (range: 50-181 per study) (see "Data Availability" for details). Each database (DB_i, where $i = 1, 2, \ldots, 6$) contains structured sociodemographic information (including age, sex, race/ethnicity) and detailed CGM data. The prediction task was a binary classification of weekly hyperglycemia severity, where participants were labeled as either low risk (≤ 3 severe hyperglycemic events per week) or high risk (>3 events). Severe events were defined as glucose levels exceeding 250 mg dL^{-1} for more than 180 min, in accordance with established CGM clinical guidelines.[55–58] The framework was applied to each dataset individually and to a harmonized joint T1D database combining all 6 sources. Demographic description is shown in Table 2, with CGM summaries and preprocessing details in Supplementary Material S1.

Step 1: Impact of controlled representation sampling on evaluation cohorts

We first evaluated how controlled representation sampling influences training and evaluation cohorts. To ensure consistent and unbiased assessment, we generated $K = 50$ fixed test sets per dataset that satisfied 3 criteria: (1) balanced subgroup proportions (eg, $N_a = N_b$), (2) prevalence-aligned class distributions ($S_a \approx S_b$) to preserve real-world clinical outcome frequencies, and (3) comparable CGM exposure across subgroups. These test sets were used across all experiments to avoid shifts in evaluation conditions.

Training sets systematically varied subgroup representation across 5 levels (0%, 25%, 50%, 75%, and 100%) for each protected attribute (eg, sex, insurance, income). For each level, we created 50 distinct stratified training sets using random sampling ensuring outcome prevalence was preserved to prevent artificial class imbalance. This representation-aware sampling disentangled subgroup inclusion effects from class imbalance and provided the foundation for examining representation-driven model behavior.

Step 2: Performance and stability across multiple model ensembles

To quantify both accuracy and variability, we trained ensembles of 50 independently initialized models for each training configuration and dataset, using logistic regression (LR), random forest (RF), naive Bayes, and CatBoost. The objective was to predict the weekly acute hyperglycemia risk, defined as distinguishing between low risk (3 or fewer severe hyperglycemic events over the next week) and high risk (more than 3 such events during the next week). All algorithms showed similar patterns, but LR most consistently achieved the strongest performance across 4 of the 6 datasets and the joint DB, while also providing interpretability aligned with clinical use (see Tables S3 and S4 in Supplementary Material S2).

For instance, LR achieved an AUC of 0.805 ± 0.046 in DB1 and 0.819 ± 0.017 in the T1D joint DB, outperforming more complex models like CatBoost in these settings. Naive Bayes exhibited the highest variance in performance and lower average AUC, suggesting a tendency toward underfitting in complex, low-sample environments. Meanwhile, RF and CatBoost delivered robust but slightly less consistent results across the board, with CatBoost performing particularly well in high-signal datasets, meaning datasets where predictor variables show clear separation between outcome classes, low noise and missingness, balanced subgroup representation, and minimal confounding, such as DB3.

Model stability was characterized using (1) the standard deviation of AUC across bootstrapped evaluations and (2) systematic arbitrariness,[52] which quantifies prediction inconsistency across model instances trained on slightly different samples. Lower values indicate stable behavior under small perturbations. DB6 exhibited low AUC but also very low arbitrariness (3%), suggesting consistent yet underperforming models (see Table 3). In contrast, DB4 showed high arbitrariness (51.2%) despite moderate AUC, signaling hidden volatility likely driven by heterogeneity or noise. The joint DB demonstrated high mean performance and relatively low arbitrariness (13%), illustrating how harmonized multisource datasets can improve model stability. These findings reinforce the need to assess both accuracy and consistency when evaluating model reliability for clinical applications.

Step 3: Subgroup disparities in performance, stability, and consistency

We analyzed subgroup robustness by examining AUC disparities, instability patterns, and self-consistency gaps across representation levels. Systematic arbitrariness quantified within-group prediction instability, while Kolmogorov-Smirnov (KS) distances between subgroup self-consistency distributions captured differences in reliability between demographic strata (see details in Supplementary Material S3).

Significant disparities emerged across multiple sociodemographic dimensions (Table 4). Sex-based performance gaps

Table 2. Baseline sociodemographic characteristics of participants across datasets.

Feature	DB1 (*n* =103)	DB2 (*n* =181)	DB3 (*n*=99)	DB4 (*n*=50)	DB5 (*n*=113)	DB6 (*n*=144)	Joint DB (*n*=690)
Age (median ± SD, years)	4±1.2	12±3.6	11±2.0	16±1.2	16±1.6	5±1.7	10±5.1
Years since diagnosis (median ± SD, years)	1±.0	5±.6	5±.8	6±.7	7±4.3	2±1.9	4.8±3.8
Sex (%)							
Female	51.5	42.5	50.5	46.0	52.2	50	48.4
Male	48.5	57.5	49.5	54.0	47.8	50	51.6
Unknown	–	–	–	–	–	–	–
Race (%)							
White	84.5	74.6	85.9	86.0	77.0	71.5	78.3
Black/African American	5.8	12.7	–	–	8.8	15.3	8.8
Asian	1.9	2.8	2.0	4.0	4.4	0.7	2.5
American Indian/ Alaska Native	–	1.7	1.0		0.9		0.7
More than one race	7.8	6.6	10.1	10.0	4.4	9.7	7.8
Unknown	–	1.7	1.0	–	4.4	2.8	1.9
Ethnicity (%)							
Hispanic or Latino	14.6	14.4	8.1	8.0	22.1	11.1	13.6
Non-Hispanic or Latino	85.4	85.6	91.9	92.0	77.0	87.5	85.9
Unknown	–	–	–	–	0.9	1.4	0.4
BMI classification (%)							
Normal weight	27.2	3.9	5.1	18.0	23.9	14.6	14.1
Overweight/obese	16.5	2.8	4.0	20.0	21.2	13.9	11.6
Unknown	56.3	93.4	90.9	62.0	54.9	71.5	74.3
Insurance type (%)							
Public	22.3	19.9	10.1	10.0	43.4	35.4	25.2
Private	67.0	76.8	86.9	90.0	54.0	51.4	68.7
Both	8.7	1.7	3.0	3.0	2.7	10.4	4.8
No insurance	1.0	0.6	–	–	–	1.4	0.6
Unknown	1.0	1.1	–	–	–	1.4	0.7
Education level (%)							
Less than high school	–	1.1	–	8.0	14.2	0.7	3.3
High school graduate	8.7	26.5	5.1	10.0	44.2	47.2	26.8
Technical/Vocational	2.9	0.6	–	–	–	–	0.6
College/Bachelor's degree	35.0	29.3	40.4	26.0	20.4	28.5	29.9
Associate degree	10.7	9.9	3.0	2.0	9.7	7.6	8.0
Advanced degree	42.7	30.9	51.5	54.0	9.7	11.1	29.7
Unknown	–	1.7	–	–	1.8	4.9	1.7
Annual income (%)							
<$25k	–	2.2	–	4.0	5.3	11.8	4.2
$25k-$35k	3.9	2.8	2.0	–	15.0	5.6	5.2
$35k-$50k	9.7	5.0	3.0	–	15.9	17.4	9.4
$50k-$75k	12.6	8.3	5.1	8.0	13.3	20.8	11.9
$75k-$100k	17.5	14.9	17.2	12.0	11.5	16.7	15.2
$100k-$200k	32.0	31.5	34.3	38.0	22.1	18.8	28.3
$≥200k	18.4	24.3	32.3	22.0	4.4	0.7	16.2
Unknown	5.8	11.0	6.1	16.0	12.4	0.3	9.6

Continuous variables are summarized as median ± SD; categorical variables are shown as percentages.

appeared in 3 datasets, with consistently higher AUC for male participants. Higher parental education and higher household income were associated with improved performance in 2 individual datasets and the joint DB. One dataset showed substantially better AUC for privately insured individuals.

Importantly, these disparities remained across different subgroup representation levels (see Figure 3A). For example, in one dataset, the AUC for the male subgroup consistently outperformed the female subgroup, even when using a training set comprised entirely of female participants (100% F/0% M).

In contrast, when the model was trained either on balanced data (50% F/50% M) or exclusively on male data (0% F/100% M), female AUC performance degraded notably (see Figure 3B). This asymmetry reveals that data from certain subgroups contributes more effectively to the model's learning process—suggesting unequal *training value* between subgroups.

Next, to evaluate the impact of subgroup representation on model stability, we measure arbitrariness from 2 complementary perspectives: (1) the overall arbitrariness across different representation levels and (2) subgroup-specific disparities in systematic arbitrariness, quantified using the KS distance between the empirical cumulative distribution functions (CDFs) of prediction self-consistency across subgroups. Overall arbitrariness typically varied by less than 8% across representation. The largest variation was observed in DB4, where overall arbitrariness increased by up to 8% across representation scenarios (Figure 4A). However, the subgroup-level analysis revealed more pronounced disparities—particularly for sex, where the normalized KS distance reached 12% in DB3 (Figure 4B). These results highlight why aggregate metrics can obscure subgroup-specific vulnerabilities.

Table 3. Predictive performance and model consistency across datasets.

Dataset	Mean AUC ± SD	Systematic arbitrariness (%)
DB1	0.805 ± 0.046	31.8
DB2	0.722 ± 0.051	24.4
DB3	0.794 ± 0.050	32.4
DB4	0.734 ± 0.064	51.2
DB5	0.645 ± 0.051	19.8
DB6	0.567 ± 0.054	3.0
Joint DB	0.819 ± 0.017	13.0

Mean AUC and systematic arbitrariness for logistic regression models trained on each dataset ($N = 50$ repetitions per dataset).

Step 4: Actionable insights and mitigation strategies from disparity patterns

The final step of our discrimination risk assessment framework synthesizes the multilevel insights generated across previous stages to support informed, actionable decisions. The introduction of training value as an operational metric provides practical guidance for optimizing data composition. In synthesis, subgroups with low representation but high training value should be prioritized during data collection and model refinement, whereas subgroups whose inclusion does not enhance (or even degrades) generalization must also be well represented, but need further investigation—there might be issues of data quality, noise, or sampling bias.

Applying this framework to the databases, we study how relying solely on performance metrics can obscure critical dimensions of model behavior, as shown in Figure 5 for DB3 in the case of sex and for the joint DB in the case of age. In one instance, despite stable classification accuracy across subgroup proportions, the underlying prediction consistency varied significantly—pointing to hidden instability not captured by AUC. In another, overall model reliability improved, yet disparities in performance widened between subgroups, revealing a disconnect between model robustness and outcome equity. In light of the T1D validation examples, we provide targeted recommendations for mitigating discrimination risk based on observed representation-performance patterns. Specifically, in DB3 (sex), the configuration yielding the lowest disparity metrics involved oversampling the male subgroup at a 3:1 ratio relative to the female subgroup. This suggests that higher male representation stabilizes model behavior and reduces intergroup performance divergence in this dataset. Conversely, increased representation of females was associated with amplified systematic arbitrariness (KS distance), indicating that current feature sets may inadequately

Table 4. Subgroup-level AUC disparities across representation levels and demographic attributes.

Dataset	Attribute	0/100	25/75	50/50	75/25	100/0	Acute risk disparity
DB1	Sex	NS	NS	NS	NS	NS	NS
	Education	**L<H	**L<H	**L<H	**L<H	**L<H	**
	Income	***L<H	***L<H	***L<H	***L<H	***L<H	***
DB2	Sex	NS	NS	NS	NS	NS	***
DB3	Sex	***F>M	***F>M	***F>M	***F>M	***F>M	***
	Education	***L<H	***L<H	***L<H	***L<H	***L<H	*
DB4	Sex	*F<M	*F<M	*F<M	*F<M	**F<M	**
	Education	NS	NS	NS	NS	NS	NS
DB5	Sex	NS	NS	NS	NS	NS	NS
	Insurance	***PR>PU	***PR>PU	***PR>PU	***PR>PU	***PR>PU	***
DB6	Sex	NS	NS	**F<M	**F<M	**F<M	NS
Joint DB	Sex	NS	NS	NS	NS	NS	NS
	Income	***L<H	***L<H	***L<H	***L<H	***L<H	***
	Age	***C>A	***C>A	***C>A	***C>A	***C>A	***
	Time since Dx	***L>H	***L>H	***L>H	***L>H	***L>H	ns

Notation indicates statistically significant differences (*$P<.1$; **$P<.05$; ***$P<.01$), with subgroup direction (eg, F < M indicates higher accuracy for males).
Abbreviations: A, adolescents; C, children; H, high; L, low; NS, nonsignificant; PR, private insurance; PU, public insurance.

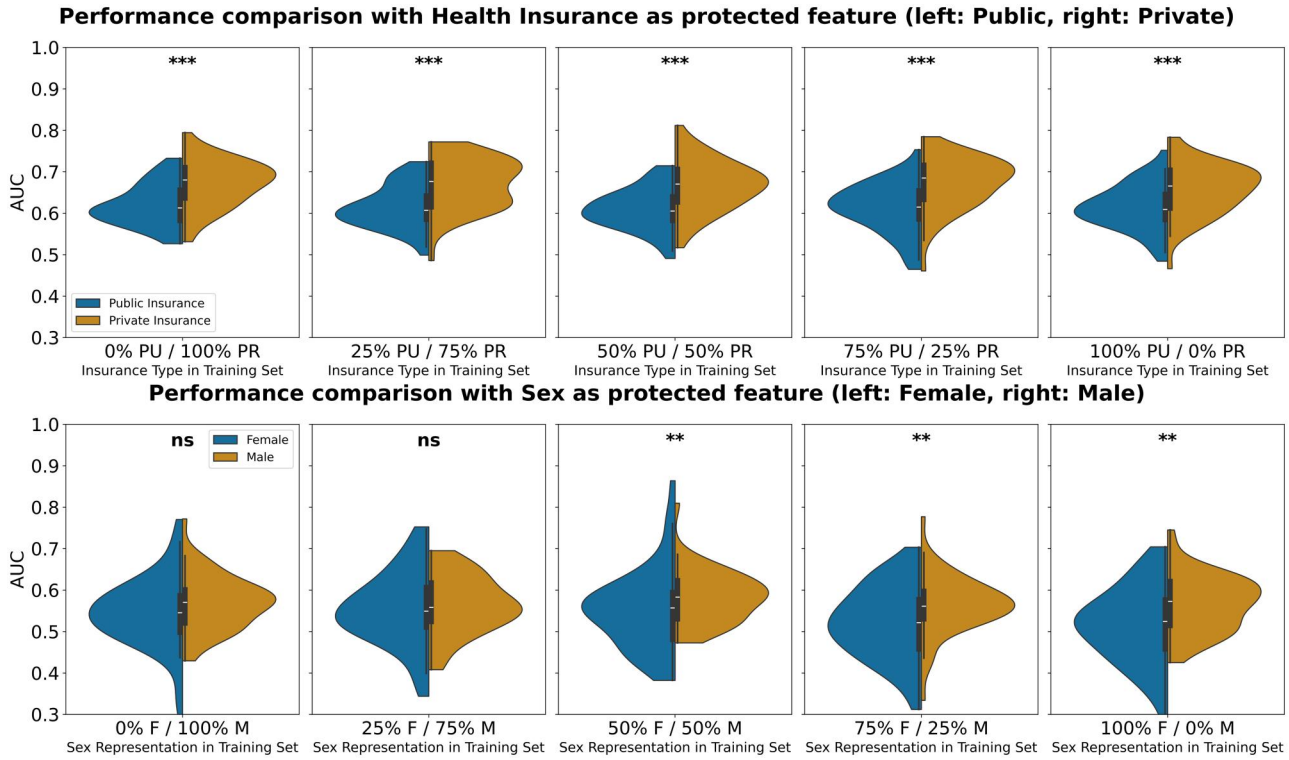

Figure 3. Examples of persistent performance disparities across subgroup representation levels. (A, top) DB4: models trained primarily on privately insured individuals consistently outperform those trained on publicly insured ones, with AUC gaps remaining stable across training proportions. (B, bottom) DB6: male subgroup AUC does not exceed female subgroup AUC when males make up the entire training set (0% F/100% M). Notably, female AUC performance drops substantially, with a pronounced disparity emerging in which models trained exclusively on female data (100% F/0% M) achieve markedly higher performance for males than for females.

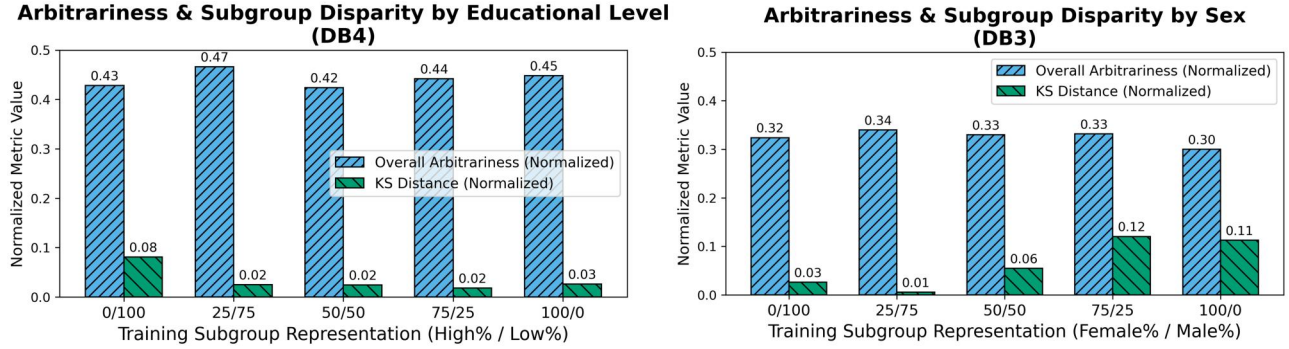

Figure 4. Examples of arbitrariness and subgroup-level inconsistency across representation levels. (A, left) Maximum deviation in overall model arbitrariness (DB4) for caregiver education. Normalized AUC-based arbitrariness shows a deviation of up to 8% across training representation levels (0% to 100% low vs high education). (B, right) Maximum inconsistency between subgroups, measured by normalized Kolmogorov-Smirnov (KS) distance. For sex in DB3, subgroup consistency diverges by 12%, with female predictions exhibiting greater internal consistency than male counterparts.

capture predictive signals for this subgroup. We recommend improving feature design or refining data collection for the female subgroup.

Similarly, in the joint DB (age), the lowest subgroup disparities emerged when the training configuration included only children data. Increasing adolescent representation widened stability and performance gaps, suggesting adolescent-specific dynamics were insufficiently captured. Mitigation requires targeted feature review or enhanced adolescent-specific data acquisition.

These examples show that equal representation does not guarantee fairness: some subgroups contribute disproportionately to model robustness, while others reveal structural data issues. The framework therefore supports proactive strategies such as targeted oversampling, stratified recruitment, subgroup-specific feature engineering, and strategic dataset design.

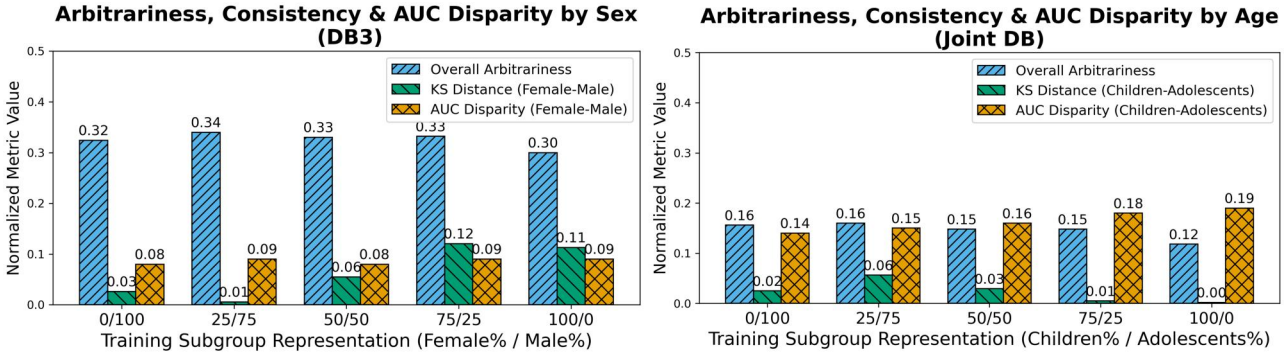

Figure 5. Examples of performance parity and model instability under varying subgroup representation. (A, left) DB3 (sex): AUC disparity remains stable across subgroup representation levels, despite variability in overall model arbitrariness and subgroup-level consistency (KS distance). (B, right) Joint DB (age): AUC disparity increases even as overall arbitrariness and consistency disparity between age groups decrease, highlighting a divergence between performance and model stability metrics.

Discussion

This study introduces a modular framework for discrimination risk assessment in clinical ML, designed to uncover not only disparities in predictive performance but also instabilities in model behavior across subgroups and data configurations. By incorporating repeated sampling, ensemble training, and distributional aggregation, the approach provides a more comprehensive diagnostic view than traditional single-model evaluations. Applied across multiple real-world datasets, the framework reveals hidden consistency issues and subgroup-specific vulnerabilities that standard metrics such as AUC often fail to surface.

A key contribution of this work is the explicit separation of data representation effects from model behavior, operationalized through metrics such as systematic arbitrariness, training value, and subgroup self-consistency. These measures enable characterization of reliability—not only whether the model performs well, but whether it behaves stably across demographic strata. Our findings demonstrate that demographic balancing alone does not guarantee equitable outcomes; subgroup-specific instability can persist even under symmetric representation, aligning with emerging evidence of age-related bias in health information technologies.[59] By enabling multidimensional subgroup evaluation, the framework supports earlier detection of risks that would otherwise remain obscured in aggregate validation metrics.

From an informatics perspective, the framework provides a structured and transparent method for integrating fairness-oriented diagnostics into standard ML validation pipelines, supporting reproducibility, auditability, and governance-ready evaluation of clinical AI systems.[60] It assists data curators in assessing whether specific subgroups contribute meaningfully to model robustness, informing recruitment or sampling strategies. For model developers, it offers criteria for prioritizing when disparities merit targeted interventions such as data augmentation, weighting, or subgroup-specific evaluation. More broadly, it supplies data providers with practical guidance[61,62] to avoid upstream biases during dataset creation, improving the suitability of real-world clinical datasets for downstream algorithm development.[63,64]

While the framework is general by design, empirical validation relied on T1D datasets and CGM-derived outcomes, which limits generalizability of outcome-specific insights. The binary prediction setting may not capture discrimination dynamics present in more complex tasks, and some protected attributes could not be evaluated due to imbalances in CGM monitoring duration. Crucially, the framework requires adequate subgroup sample sizes; smaller datasets limit analyses to coarse single-attribute splits, whereas larger datasets permit fine-grained, multiattribute stratification. When data sparsity is severe, training a reliable statistical model becomes fundamentally difficult, limiting the feasibility of controlled subsampling. Additionally, while the framework quantifies behavioral disparities, it does not establish causal mechanisms underlying them, which remains an important direction for future study.

Future work should explore applicability across broader informatics contexts, including deep learning, unsupervised methods, reinforcement learning, and generative AI, where discrimination mechanisms may differ. Furthermore, we should examine how this framework can be adapted for postdeployment monitoring, as discrimination risk may emerge or change over time due to data drift and evolving patient populations.[65] Validation in additional clinical and nonclinical domains—and with multimodal datasets integrating structured data, imaging, or clinical text—will be essential to fully assess generalizability. These directions would broaden the framework's generalized utility and support its adoption as a standard tool for fairness evaluation in diverse health-care AI systems.

Conclusion

Our framework introduces a novel and scalable approach for discrimination risk assessment in ML pipelines that embeds subgroup representation, model stability, and training value directly into the model validation process. In medical informatics, 3 persistent barriers limit equitable model development. First, social determinants of health, including socioeconomic status, insurance type, and caregiver education, remain vastly underreported in clinical datasets and underrepresented in the design of medical AI tools, including devices like continuous glucose monitors. Second, while model validation protocols are gaining sophistication, most remain narrowly focused on performance metrics, with limited incorporation of fairness-aware model

diagnostics. Third, when working with pediatric populations, existing guidelines rightly call for special considerations in the use of children's data, but these do not yet extend to evaluating discrimination risk, despite the fact that model bias may emerge not only against the child, but also against caregivers, due to intertwined sociodemographic and model dynamics. By embedding such considerations into a unified validation framework, we provide actionable guidance for developers, regulators, and clinicians working at the intersection of pediatric care and ML. This dual emphasis on rigorous, representation-aware validation and data-efficient mitigation planning positions the framework as a practical tool for advancing responsible AI within clinical and medical informatics contexts.

Acknowledgements

The authors thank the Jaeb Center for Health Research for providing access to the publicly available type 1 diabetes datasets. The authors further thank colleagues and collaborators for valuable discussions.

Author contributions

Ioannis Bilionis (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing—original draft, Writing—review & editing), Ricardo C. Berrios (Conceptualization, Funding acquisition, Resources, Supervision, Writing—original draft, Writing—review & editing), Antonio de Arriba Muñoz (Investigation, Validation, Writing—original draft, Writing—review & editing), Luis Fernandez-Luque (Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Validation, Writing—original draft, Writing—review & editing), and Carlos Castillo (Formal analysis, Investigation, Methodology, Supervision, Validation, Writing—original draft, Writing—review & editing)

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

This work was funded by the Spanish Ministry of Science and Innovation (REF: DIN2021-011865). It was also partially supported by the Department of Research and Universities of the Government of Catalonia (SGR 00930), the EU-funded project SoBigData++ (grant agreement 871042), and MCIN/AEI/10.13039/501100011033 under the Maria de Maeztu Units of Excellence Programme (CEX2021-001195-M). None of the funders played any role in the study design, data collection, analysis, interpretation of results, or the writing of this manuscript.

Conflicts of interest

Ioannis Bilionis, Ricardo C. Berrios and Luis Fernandez Luque are employees of Adhera Health. All other authors declare that they have no competing interests or conflicts of interest related to this manuscript.

Data availability

This study used publicly available, deidentified datasets coordinated by the Jaeb Center for Health Research (https://public.jaeb.org/datasets/diabetes). All datasets were obtained from previously completed and ethically approved clinical trials, with informed consent collected by the original investigators between 2008 and 2022. No new data collection or interaction with participants took place. All datasets were accessed in accordance with JAEB's data use terms. The code supporting the findings of this study is available from the corresponding author upon request.

References

1. An Q, Rahman S, Zhou J, Kang JJ. A comprehensive review on machine learning in healthcare industry: classification, restrictions, opportunities and challenges. *Sensors*. 2023; 23:4178.
2. Liao Y, Naghizadeh P. Social bias meets data bias: the impacts of labeling and measurement errors on fairness criteria. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence (AAAI); Vol. 37. 2023:8764-8772.
3. Perets O, Stagno E, Yehuda EB, et al. Inherent bias in electronic health records: a scoping review of sources of bias. ACM Transactions on Intelligent Systems and Technology. 2025 Aug 5.
4. Lippert-Rasmussen K. *Born Free and Equal? A Philosophical Inquiry into the Nature of Discrimination*. Oxford University Press; 2013.
5. Muralidharan V, Schamroth J, Youssef A, Celi LA, Daneshjou R. Applied artificial intelligence for global child health: addressing biases and barriers. *PLOS Digit Health*. 2024; 3:e0000583.
6. Barber Doucet H, Ward VL, Johnson TJ, Lee LK. Implicit bias and caring for diverse populations: pediatric trainee attitudes and gaps in training. *Clin Pediatr (Phila)*. 2021; 60:408-417.
7. Bhargava H, Salomon C, Suresh S, et al. Promises, pitfalls, and clinical applications of artificial intelligence in pediatrics. *J Med Internet Res.* 2024;26:e49022.
8. Johnson TJ. Racial bias and its impact on children and adolescents. *Pediatr Clin North Am*. 2020;67:425-436.
9. Bakken S. *Bias, Artificial Intelligence, and Humans*. Oxford University Press; 2025.
10. Ganatra HA. Machine learning in pediatric healthcare: current trends, challenges, and future directions. *J Clin Med*. 2025;14:807.

11. Ott MA. Bias in, bias out: ethical considerations for the application of machine learning in pediatrics. *J Pediatr*. 2022; 247:124.
12. Foote HP, Cohen-Wolkowiez M, Lindsell CJ, Hornik CP. Applying artificial intelligence in pediatric clinical trials: potential impacts and obstacles. *J Pediatr Pharmacol Ther*. 2024;29:336-340.
13. Gray KD, Subramaniam HL, Huang ES. Effects of racial bias in pulse oximetry on children and how to address algorithmic bias in clinical medicine. *JAMA Pediatr*. 2023; 177:459-460.
14. Gupta EA, Huang X, Velasquez HJ, et al. Age and sex mark clinical differences in the presentation of pediatric type 1 diabetes mellitus. *J Pediatr Endocrinol Metab*. 2025;38:22-28.
15. Lawrence JM, Divers J, Isom S, et al.; SEARCH for Diabetes in Youth Study Group. Trends in prevalence of type 1 and type 2 diabetes in children and adolescents in the US, 2001-2017. *JAMA*. 2021;326:717-727.
16. Fang M, Wang D, Selvin E. Prevalence of type 1 diabetes among US children and adults by age, sex, race, and ethnicity. *JAMA*. 2024;331:1411-1413.
17. Gandhi K, Tosur M, Schaub R, Haymond M, Redondo M. Racial and ethnic differences among children with new-onset autoimmune type 1 diabetes. *Diabetic Med*. 2017; 34:1435-1439.
18. Lipman TH, Smith JA, Patil O, Willi SM, Hawkes CP. Racial disparities in treatment and outcomes of children with type 1 diabetes. *Pediatr Diabetes*. 2021;22:241-248.
19. Yu TS, Song S, Yea J, Jang KI. Diabetes management in transition: market insights and technological advancements in CGM and insulin delivery. *Adv Sens Res*. 2024;3:2400048.
20. Patel PM, Thomas D, Liu Z, Aldrich-Renner S, Clemons M, Patel BV. Systematic review of disparities in continuous glucose monitoring and insulin pump utilization in the United States: key themes and evidentiary gaps. *Diabetes Obes Metab*. 2024;26:4293-4301.
21. Zakaria NI, Tehranifar P, Laferrère B, Albrecht SS. Racial and ethnic disparities in glycemic control among insured US adults. *JAMA Netw Open*. 2023;6:e2336307-7.
22. DeSalvo DJ, Noor N, Xie C, et al. Patient demographics and clinical outcomes among type 1 diabetes patients using continuous glucose monitors: data from T1D exchange real-world observational study. *J Diabetes Sci Technol*. 2023; 17:322-328.
23. Powell PW, Chen R, Kumar A, Streisand R, Holmes CS. Sociodemographic effects on biological, disease care, and diabetes knowledge factors in youth with type 1 diabetes. *J Child Health Care*. 2013;17:174-185.
24. Willers C, Iderberg H, Axelsen M, et al. Sociodemographic determinants and health outcome variation in individuals with type 1 diabetes mellitus: a register-based study. *PLoS One*. 2018;13:e0199170.
25. Scott A, Chambers D, Goyder E, O'Cathain A. Socioeconomic inequalities in mortality, morbidity and diabetes management for adults with type 1 diabetes: a systematic review. *PLoS One*. 2017;12:e0177210.
26. Estiri H, Strasser ZH, Rashidian S, et al. An objective framework for evaluating unrecognized bias in medical AI models predicting COVID-19 outcomes. *J Am Med Inform Assoc*. 2022; 29:1334-1341.
27. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science (1979)*. 2019;366:447-453.
28. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med*. 2018;169:866-872.
29. Sikstrom L, Maslej MM, Hui K, Findlay Z, Buchman DZ, Hill SL. Conceptualising fairness: three pillars for medical algorithms and health equity. *BMJ Health & Care Inform*. 2022; 29:e100459.
30. Garba-Sani Z, Farinacci-Roberts C, Essien A, Yracheta JM. ACCESS AI: a new framework for advancing health equity in health care AI. *Health Affairs Forefront*. 2024.
31. Gupta R, Sasaki M, Taylor SL, et al. Developing and applying the BE-FAIR equity framework to a population health predictive model: a retrospective observational cohort study. *J Gen Intern Med*. 2025;40:2537-2547.
32. Ladin K, Cuddeback J, Duru OK, et al. Guidance for unbiased predictive information for healthcare decision-making and equity (GUIDE): considerations when race may be a prognostic factor. *NPJ Digit Med*. 2024;7:290.
33. Ekoniak J, Asadinia M. HEALTH-ML: a machine learning framework for equity-driven public health outcome prediction. In: *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE; 2025:00491-00498.
34. Holstein K, Wortman Vaughan J, Daumé IH, Dudik M, Wallach H. Improving fairness in machine learning systems: what do industry practitioners need? In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019:1-16.
35. Jagtiani P, Karabacak M, Margetis K. A concise framework for fairness: navigating disparate impact in healthcare AI. *J Med Artif Intell*. 2025;8:51.
36. Matos J, Gallifant J, Chowdhury A, et al. A clinician's guide to understanding bias in critical clinical prediction models. *Crit Care Clin*. 2024;40:827-857.
37. Dykstra S, MacDonald M, Beaudry R, et al. An institutional framework to support ethical fair and equitable artificial intelligence augmented care. *NPJ Digit Med*. 2025;8:84.
38. Li F, Wu P, Ong HH, Peterson JF, Wei WQ, Zhao J. Evaluating and mitigating bias in machine learning models for cardiovascular disease prediction. *J Biomed Inform*. 2023; 138:104294.
39. Davis SE, Dorn C, Park DJ, Matheny ME. Emerging algorithmic bias: fairness drift as the next dimension of model maintenance and sustainability. *J Am Med Inform Assoc*. 2025; 32:845-854.
40. Colacci M, Pou-Prom C, Siddiqi A, Mamdani M, Verma AA. Evaluating sociodemographic bias in a deployed machine-learned patient deterioration model. *JAMIA Open*. 2025; 8:ooaf158.
41. Raza S, Shaban-Nejad A, Dolatabadi E, Mamiya H. Exploring bias and prediction metrics to characterise the fairness of machine learning for equity-centered public health decision-making: a narrative review. *IEEE Access*. 2024;12: 180815-180829.

42. Al-Zanbouri Z, Sharma G, Raza S. Equity in healthcare: analyzing disparities in machine learning predictions of diabetic patient readmissions. In: *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*. IEEE; 2024:660-669.
43. Hu L, Li D, Liu H, et al. Enhancing fairness in AI-enabled medical systems with the attribute neutral framework. *Nat Commun*. 2024;15:8767.
44. Liu M, Ning Y, Ke Y, et al. FAIM: Fairness-aware interpretable modeling for trustworthy machine learning in healthcare. *Patterns*. 2024;5:101059.
45. Hastings P, Hughes S, Blaum D, Wallace P, Britt MA. Stratified learning for reducing training set size. In: *International Conference on Intelligent Tutoring Systems*. Springer; 2016:341–346.
46. Puyol-Antón E, Ruijsink B, Piechnik SK, et al. Fairness in cardiac MR image analysis: an investigation of bias due to data imbalance in deep learning based segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2021:413-423.
47. Chinta SV, Wang Z, Palikhe A, et al. AI-driven healthcare: fairness in AI healthcare: a survey. *PLOS Digit Health*. 2025;4.
48. Rajput D, Wang WJ, Chen CC. Evaluation of a decided sample size in machine learning applications. *BMC Bioinformatics*. 2023;24:48.
49. Shen JH, Raji ID, Chen IY. The data addition dilemma. arXiv, arXiv:240804154, 2024, preprint: not peer reviewed.
50. Pot M, Kieusseyan N, Prainsack B. Not all biases are bad: equitable and inequitable biases in machine learning and radiology. *Insights Imaging*. 2021;12:13.
51. Raumanns R, Schouten G, Pluim JP, Cheplygina V. Dataset distribution impacts model fairness: single vs. multi-task learning. In: *MICCAI Workshop on Fairness of AI in Medical Imaging*. Springer; 2024:14-23.
52. Cooper AF, Lee K, Choksi MZ, et al. Arbitrariness and social prediction: the confounding role of variance in fair classification. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence (AAAI); Vol. 38. 2024:22004-22012.
53. Das S, Singh A, Chatterjee S, Bhattacharya S, Bhattacharya S. Finding high-value training data subset through differentiable convex programming. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer; 2021:666-681.
54. Ghorbani A, Zou J. Data shapley: equitable valuation of data for machine learning. In: *International Conference on Machine Learning*. PMLR; 2019:2242-2251.
55. Battelino T, Danne T, Bergenstal RM, et al. Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range. *Diabetes Care*. 2019;42:1593-1603.
56. Maiorino MI, Signoriello S, Maio A, et al. Effects of continuous glucose monitoring on metrics of glycemic control in diabetes: a systematic review with meta-analysis of randomized controlled trials. *Diabetes Care*. 2020;43:1146-1156.
57. Mouri M, Badireddy M. *Hyperglycemia*. StatPearls Publishing; 2023.
58. Care D, et al. Standards of care in diabetes—2023. *Diabetes Care*. 2023;46:S1-S267.
59. Van Kolfschooten H. The AI cycle of health inequity and digital ageism: mitigating biases through the EU regulatory framework on medical devices. *J Law Biosci*. 2023; 10:lsad031.
60. Chan A, Rahimi-Ardabilli H, Rogers WA, Coiera E. The real-world impact of artificial intelligence ethics frameworks across a decade in healthcare: a scoping review. *J Am Med Inform Assoc*. 2025;32:1767-1777.
61. Wang HE, Landers M, Adams R, et al. A bias evaluation checklist for predictive models and its pilot application for 30-day hospital readmission models. *J Am Med Inform Assoc*. 2022; 29:1323-1333.
62. Muralidharan V, Burgart A, Daneshjou R, Rose S. Recommendations for the use of pediatric data in artificial intelligence and machine learning ACCEPT-AI. *NPJ Digit Med*. 2023;6:166.
63. Muralidharan V, Adewale BA, Huang CJ, et al. A scoping review of reporting gaps in FDA-approved AI medical devices. *NPJ Digit Med*. 2024;7:273.
64. Liang W, Rajani N, Yang X, et al. Systematic analysis of 32,111 AI model cards characterizes documentation practice in AI. *Nat Mach Intell*. 2024;6:744-753.
65. Subasri V, Krishnan A, Kore A, et al. Detecting and remediating harmful data shifts for the responsible deployment of clinical AI models. *JAMA Netw Open*. 2025;8:e2513685-5.

JAMIA Open, 2026, 9, 1–12
https://doi.org/10.1093/jamiaopen/ooag125
Research and Applications