Communications Medicine

https://doi.org/10.1038/s43856-026-01433-9

Article in Press

Integration of fairness-awareness into clinical language processing models

Received: 9 July 2025

Accepted: 3 February 2026

Cite this article as: Abulibdeh, R., Lin, Y., Ahmadi, S. *et al.* Integration of fairness-awareness into clinical language processing models. *Commun Med* (2026). https://doi.org/10.1038/s43856-026-01433-9

Rawan Abulibdeh, Yihang Lin, Sepehr Ahmadi, Ervin Sejdić, Leo Anthony Celi, Qiuyi Zhao & Karen Tu

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Integration of fairness-awareness into clinical language processing models

Rawan Abulibdeh[1], **Yihang Lin**[1], **Sepehr Ahmadi**[1, 2], **Ervin Sejdić**[1,3,*], **Leo Anthony Celi**[4, 5, 6], **Qiuyi Zhao**[1], **and Karen Tu**[7,8,3]

[1]Department of Electrical and Computer Engineering, University of Toronto, 10 King's College Rd, Toronto, M5S 3G8, Ontario, Canada
[2]Neurosciences and Mental Health Research Program, The Hospital for Sick Children, 170 Elizabeth St, Toronto, M5G 1E8, Ontario Canada
[3]North York General Hospital, 4001 Leslie St, North York, M2K 1E1, Ontario, Canada
[4]Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA, United States
[5]Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, United States
[6]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, United States
[7]Department of Family and Community Medicine, University of Toronto, 500 University Ave, Toronto, M5G 1V7, Ontario, Canada
[8]Toronto Western Hospital Family Health Team, University Health Network, 3rd Floor, 440 Bathurst St, Toronto, M5T 2S6, Ontario, Canada
[*]Corresponding author: esejdic@ieee.org

ABSTRACT

Background: Equitable deployment of clinical artificial intelligence systems requires consistent performance across diverse patient populations. However, race information in electronic health records is often missing/inconsistently documented, limiting the ability to construct representative cohorts or assess algorithmic bias. This study evaluates model performance and fairness in predicting race from clinical text.
Methods: We compared four transformer-based deep learning models with a hierarchical convolutional neural network designed to capture the multilevel structure of clinical narratives. A two-phase active learning framework guided annotation of a primary care database. A fairness-aware loss function was applied to mitigate disparities across racial groups. Each model was trained with and without fairness-aware optimization. Performance and equity were evaluated using 10-fold cross-validation and subgroup audits across race, sex, age, and their intersections.
Results: Here we show that the hierarchical convolutional neural network achieves higher accuracy and performance equity than transformer models (macro F1 = 98.4%). Fairness constraints enhance parity across most transformer architectures, but degrade hierarchical model performance and cause one clinical model to collapse toward majority predictions, demonstrating that fairness interventions are highly model dependent. Persistent disparities across race, sex, and age indicate that inequities reflect architectural limitations and systemic biases.
Conclusions: This study demonstrates that fairness can be integrated into clinical language models, though effects vary by model type. Architectures aligned with clinical text structure inherently promote fairness, yet mixed fairness constraint outcomes highlight the need for tailored interventions. Persistent demographic disparities show that algorithmic bias often reflects upstream documentation inequities. This framework offers a scalable path toward equitable NLP for clinical artificial intelligence.

Plain language summary

Medical records often lack information about patients' race, making it hard to identify potential race-associated health inequalities. We developed computer programs to find race information in doctors' notes. We tested different types of artificial intelligence models and added special rules to make them work fairly for all racial groups. We found that a model designed to read notes the way doctors write them worked best. Adding additional fairness rules helped some models but hurt others, showing there is no one-size-fits-all solution. Many differences we saw came from how doctors write their notes differently for different patient groups. This research shows we can build fairer medical artificial intelligence, but fixing computer programs alone is not enough. We also need to improve how health information is recorded.

Introduction

Institutionalized and internalized forms of racial discrimination play a significant role in shaping both health outcomes and healthcare utilization[1]. A substantial body of research has shown that racial background is closely linked to disparities across a range of health indicators[2–8]. Race is a multidimensional construct that often serves as a proxy for exposure to systemic and structural health risks shaped by social, behavioral, and environmental conditions. These socialized expressions of race can result in biologically relevant outcomes, contributing to persistent disparities in morbidity and mortality[9]. As social constructs, race and ethnicity remain critical lenses through which to examine inequities in healthcare access, quality of care, and broader social determinants of health, including housing, education, and nutrition. Evidence consistently shows that ethnic minority populations, particularly non-white groups, experience worse health outcomes than their white counterparts[4]. However, the temporal development and distribution of racial disparities across specific health conditions, especially among underrepresented groups such as the Hispanic population in America, remain poorly understood. Enhancing the consistency and utility of patients' racial and ethnic backgrounds is essential for advancing equity-focused research, enabling the creation of demographically informed cohorts, supporting analyses of disease onset and risk stratification by ethnicity, and guiding the development of targeted interventions and equitable resource allocation. A critical first step toward this goal is evaluating the feasibility of developing fair and robust algorithmic approaches that can transform unstructured clinical text into structured representations of race and ethnicity, enabling consistent identification and classification within electronic health records (EHRs).

EHRs have emerged as a key source of real-world data for capturing social determinants of health factors, such as race, enabling the study of how social risk factors influence disease onset, treatment, and health service utilization[10]. However, many social risk factors—including race—are inconsistently documented, particularly in structured EHR fields, which are often incomplete, outdated, or poorly aligned with evolving social constructs[11,12]. Several studies have quantified the extent of missing or incomplete race and ethnicity information within structured EHR fields, revealing that these gaps remain a significant barrier to equitable data representation. Analyses of clinical text have revealed an additional 948 patients identified as Black (+26%) and 665 as Hispanic (+20%) beyond what was captured in structured EHR fields[13]. Across EHRs, missing or uninformative race entries range from 25% to over 57%[14], with misclassification disproportionately affecting multiracial patients and other minoritized groups[15]. These omissions are not merely technical deficiencies; they distort population denominators, obscure health inequities, and introduce systematic bias into downstream analyses. In contrast, unstructured clinical text provides a rich but underutilized source of social and behavioral information. Yet, extracting meaningful insights from free-text data poses its own challenges due to its unstructured nature, domain-specific language, and high variability in grammar and spelling[16–18]. Recent advances in machine learning, particularly in natural language processing (NLP), offer promising tools for identifying such factors from clinical narratives[12,19,20]. Applying these methods to identify race in EHRs can enable the creation of higher-quality datasets, facilitate the study of racial disparities, and ultimately support more equitable healthcare delivery.

Most existing approaches for extracting social determinants of health from clinical text rely on transformer-based or conventional deep learning models that treat text as a flat sequence of tokens[21,22]. However, clinical documentation is inherently hierarchical—composed of sentences within notes, and notes across multiple encounters—each contributing different layers of contextual meaning[23]. Ignoring this structure can limit a model's ability to capture important semantic and temporal dependencies. Hierarchical architectures explicitly model these nested relationships, learning both word-level and sentence-level representations[24], which is particularly valuable in clinical NLP tasks where meaning is distributed across multiple notes and encounters. Despite their potential, hierarchical models remain underutilized in health-related NLP applications.

While transformer-based and other deep learning models have dominated recent clinical NLP research, it is important to note that rule-based and hybrid approaches remain viable alternatives, particularly for large-scale deployment. Rule-based systems using predefined patterns and lexicons offer high interpretability and computational efficiency[13,21]. However, they struggle with linguistic variability and contextual ambiguity[25]. Hybrid architectures combining rule-based filtering with machine learning models have emerged as a pragmatic solution: a recent study processed 2.1 billion clinical notes in two weeks using dual A40 GPUs by integrating rule-based filtering with BERT[26]. The choice between approaches involves trade-offs among accuracy, computational cost, and scalability. This study focuses on developing fairness-aware models for research applications where accuracy and equity detection are prioritized over institutional-scale deployment efficiency.

As machine learning becomes increasingly integrated into healthcare systems, concerns about algorithmic bias and fairness have grown more pressing[27–29]. Deep learning models trained on real-world clinical data risk perpetuating or amplifying existing disparities if ethical considerations—such as accountability, transparency, and privacy—are not systematically addressed. Predictive models must not only be accurate but also equitable, ensuring fair treatment across patient populations. Fairness is especially critical in the healthcare domain, given the historical marginalization of certain groups (e.g. racial and ethnic groups) in clinical care and research[30]. This concern is further amplified when models are tasked with classifying patients based on their social determinants of health, where the outcome variables themselves often represent protected attributes (e.g. sexual orientation, marital status, employment status, income, housing status). In such contexts, ensuring equitable model behavior is essential, as biased predictions can exacerbate existing inequities in downstream healthcare applications, such as in healthcare access, treatment, and outcomes for already vulnerable populations.

Despite the urgency of these concerns, only a small number of studies have explicitly evaluated algorithmic bias in the extraction of social constructs from EHRs; to our knowledge, only two have evaluated bias in this context[31,32]. Yu et al. identified performance gaps exceeding 16% across racial groups when extracting social determinants of health, highlighting persistent disparities by race and gender[31]. Similarly, Guevara et al. assessed algorithmic bias by introducing race and ethnicity descriptors into clinical narratives, demonstrating that fine-tuned models exhibited lower bias than ChatGPT[32]. However, in both cases, bias evaluation was treated as a secondary consideration within broader studies of social determinants extraction, rather than as a central design objective. Moreover, neither incorporated fairness constraints during model training or systematically examined bias across intersecting demographic axes (race, sex, and age). There is a pressing need for more comprehensive assessments of how bias emerges during data collection, model development, and deployment, especially for underrepresented groups, to address bias in predictive modeling[33]. Furthermore, it is equally important to explore and evaluate strategies that can improve model fairness and promote more equitable predictive outcomes.

This study investigates the automatic document-level classification of patient race from unstructured clinical text within EHRs as a critical step toward capturing social constructs that inform care phenotypes, enabling the assessment of healthcare equity and the identification of populations receiving lower-quality care. Unlike named-entity recognition or sentence-level concept extraction, which identify local mentions of race within notes, our objective is to determine each patient's overall racial category based on their complete longitudinal documentation. This formulation aligns with the practical use case of augmenting or validating structured demographic fields and enables population-level analyses of equity and care quality. To achieve this and address the current gaps in the literature, we first implemented an active learning framework to efficiently guide the annotation process by prioritizing notes likely to contain race-related information. We then compared state-of-the-art transformer-based models with a hierarchical convolutional neural network designed to represent the multi-level structure of clinical documentation, thereby capturing contextual dependencies across sentences and encounters. Next, we evaluated the impact of fairness-aware optimization techniques, including a loss function inspired by equalized odds, to determine whether such constraints improved predictive balance across racial groups. Finally, we analyzed potential sources of bias by examining how dataset composition and sensitive attributes such as sex and age influenced model performance. This work seeks to develop more robust, interpretable, and equitable methods for racial, and more widely social determinants of health, information extraction from clinical text, ultimately supporting the creation of more representative datasets for health equity research.

Our results show that the hierarchical convolutional neural network achieves higher overall performance and greater inter-group equity than transformer-based models, with strong accuracy and near-zero false positive rates for most racial groups. Fairness constraints improve parity for several architectures, though not universally. One transformer model exhibits reduced stability under fairness-aware training, while the hierarchical model maintains balanced performance but with some modest subgroup disparities. Persistent inequities across race, sex, and age suggest that structural factors such as pretraining bias and documentation patterns, rather than class imbalance alone, drive residual disparities. These findings demonstrate that fairness can be integrated directly into model architecture and training, offering a scalable framework for equitable and trustworthy clinical NLP, though achieving truly fair clinical aritifical intelligence requires addressing documentation practices and structural biases at their source.

Methods

Dataset overview

The dataset used to evaluate the models was the University of Toronto Practice-Based Research Network (UTOPIAN) Data Safe Haven, which is a repository of de-identified EHR data on over 400 family physicians, 96 clinics, and ∼400,000 patients in Ontario[34]. The three EHR vendors from which the UTOPIAN database extracts data from are among the most commonly used EHR vendors in family physician practices in Ontario[35,36].

To define the baseline cohort, all physicians and their patients with insufficient or low-quality data were excluded. For each data cycle, physicians were removed if they had fewer than 20% of billing, laboratory, or medication records available, or fewer than 200 rostered patients. Patients were included only if their physician met these data-quality thresholds, and if they had a valid sex and age, an EHR start date at least one year before the data extraction cut-off (unless younger than one year of age), populated entries in any of the cumulative patient profile tables, and were either rostered to a physician or had at least two family physician visits within the preceding three years. The *social history* and *risk factor* sections represent semi-structured fields within the cumulative patient profile and contain routinely collected information on patients' social determinants of health, typically updated during clinical encounters. These sections comprised 561,210 patient entries. Each entry in the EHR was timestamped, but some patients had multiple, textually identical records over time. To reduce redundancy, only the most recent entry was retained when repeated entries began with identical text.

The modeling cohort included adults aged ≥18 years as of December 31st, 2021, as social and contextual constructs were infrequently documented for children and youth. Eligible entries were grouped by patient, and the semi-structured fields from the cumulative patient profile were merged. To ensure a representative sample of the UTOPIAN database, we randomly selected 1.5% of patients from each clinic, yielding a final cohort of 4,375 patients. We verified that the random sample reflected the overall database distribution by comparing age, sex, and EHR start date, and confirmed that all physicians within each clinic were represented, with a similar number of patients per physician as in the full dataset.

A reference standard was developed by an annotator (R.A.) who manually labeled social phrases in the cohort according to predefined annotation guidelines, creating a labeled dataset for supervised machine learning models. The labeling framework followed a two-tier structure for race/ethnicity classification. The primary tier consisted of binary labels: present and absent, where present indicates the mention of race or ethnicity within the clinical note, and absent signifies no such mentions. Entries labeled as present were further categorized into nine distinct subclasses, aligned with the Canadian Institute for Health Information guidance on the use of race-based and Indigenous identity data collection and health reporting in Canada: White, Black, East Asian, Southeast Asian, South Asian, Middle Eastern, mixed heritage, Latin American, and Indigenous[37]. Additionally, missingness was treated as informative and was maintained as a separate category (absent) in the second-tier classification scheme. This hierarchical labeling approach enables both broad detection of race-related information and fine-grained categorization of racial and ethnic origins when such details are available in clinical documentation. To ensure annotation reliability, approximately 5% of the sample (219 sentences) was independently annotated by a second annotator (S.A.), yielding an inter-rater reliability kappa value of 0.91. Furthermore, S.A. was responsible for annotating the labels produced from the active learning phase.

Ethics approval and consent to participate

This study was conducted in accordance with the ethical standards outlined in the Declaration of Helsinki and all applicable institutional guidelines. The use of de-identified electronic health record data was reviewed and approved by the University of Toronto Health Sciences Research Ethics Board (REB #40129) and the North York General Hospital Research Ethics Board (REB #20-0044). The requirement for informed consent was waived by both boards, as the study used only de-identified data and posed minimal risk to participants. All data access, storage, and analysis were performed within the University of Toronto Practice-Based Research Network Data Safe Haven under approved governance protocols.

Active learning pipeline

Extracting race and ethnicity information from clinical text presented significant challenges, driven by its predominantly missing-not-at-random distribution in the annotated baseline sample and the complexity of identifying race-related cues across diverse documentation styles. To address these limitations, we implemented a machine learning framework centered on active learning, inspired by the work of Lybarger, Ostendorf, and Yetisgen[38], to strategically guide sampling toward entries more likely to contain race-associated linguistic features. This approach aimed to improve annotation efficiency and mitigate class imbalance, while maintaining interpretability and reproducibility across the corpus.

As illustrated in Figure 1, the framework consisted of two main phases: an *initial training phase* and an *active learning phase*. In the initial training phase, labeled clinical text was preprocessed to standardize structure and formatting, after which numerical embeddings were generated to capture semantic relationships within the narratives. These embeddings and their corresponding labels were used to train a baseline classifier based on the BERT architecture, forming the foundation for subsequent iterative refinement.

The second phase employed an iterative active learning mechanism comprising four key components: (1) **text processing**, where unlabeled notes were preprocessed and converted into embedding vectors; (2) **classification and uncertainty assessment**, in which the trained classifier produced prediction scores, and a query function identified samples with the highest uncertainty; (3) **human annotation integration**, where these uncertain samples were prioritized for expert labeling to create high-quality reference data; and (4) **iterative refinement**, a feedback loop where the classifier was retrained with the newly annotated examples and remaining unlabeled samples were reassessed.

Hierarchical classifier

The classification approach used a two-level hierarchical model based on the BERT-base variant architecture, fine-tuned for race and ethnicity classification, and aligned with the two-tier labeling structure. At the first level, a binary classifier determined whether race-related information was present in the clinical text. If such information was detected, the second-level classifier assigned the text to one of the nine predefined racial categories. This hierarchical architecture was designed to address the extreme class imbalance in the EHR database, where notes lacking race-related content outnumber those with such mentions by a ratio of 71 to 1. By splitting the task into two stages, each model was optimized for a distinct objective: the first-level classifier was trained to detect race-related content under highly imbalanced conditions, while the second-level classifier performed fine-grained categorization without being affected by the majority of race-absent samples. This structure also offered several

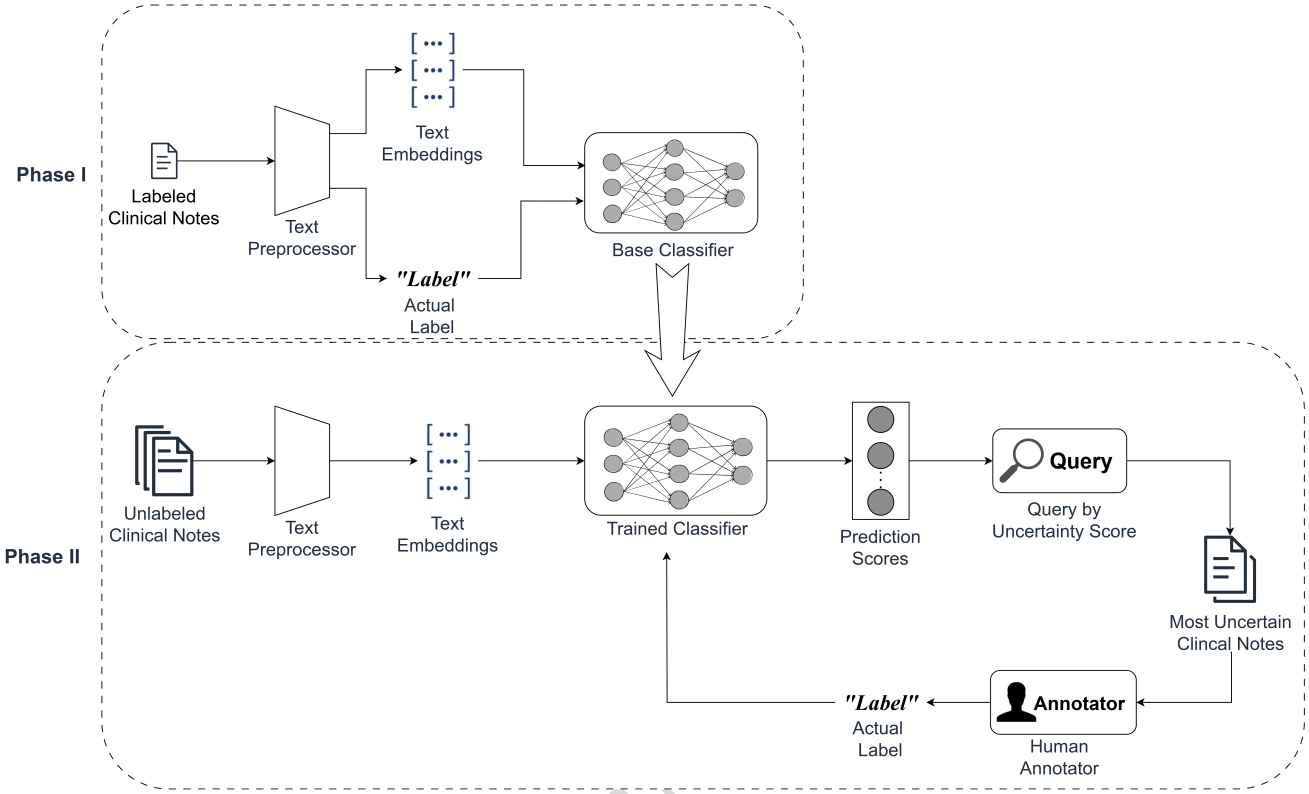


Figure 1. Schematic representation of the two-phase active learning pipeline for race/ethnicity classification from clinical text. Phase I is the initial training phase, where labeled clinical text is processed to train a base classifier. Phase II shows the active learning phase, where unlabeled notes are processed and classified using the model trained in Phase I, with uncertain predictions identified through a query mechanism. Human annotators provide labels for the most uncertain cases, which are then used to retrain the classifier in an iterative feedback loop.

advantages, including improved training efficiency, flexibility to apply targeted sampling and class weighting strategies at each level, and enhanced interpretability by decoupling detection from classification.

The BERT-base model used to guide sample selection was trained on the baseline annotated set without fairness constraints. This design choice was deliberate, intended to examine whether conventional uncertainty-based active learning (commonly applied in clinical NLP without fairness adjustments[38–40]) could inadvertently introduce demographic bias into the annotated dataset. Although fairness auditing of this preliminary model was not a primary objective of the study, we assessed the demographic composition of the actively sampled data to identify any sampling disparities.

The hierarchical CNN model architecture

Our model design was inspired in part by prior work combining CNN-based architectures with token-level selection strategies for EHR classification tasks[41]. The proposed hierarchical CNN model combined BERT-based contextual embeddings with convolutional neural networks at both the word and sentence levels to support hierarchical document classification by capturing both local n-gram features within sentences while identifying salient sentence-level information across the document. Unlike static embeddings (e.g., Word2Vec or GloVe), BERT provides dynamic, context-aware representations, enabling the model to better interpret domain-specific terminology, abbreviations, and polysemous language without requiring additional task-specific pretraining. We selected CNNs and attention mechanisms as a lightweight yet expressive alternative to stacking multiple transformer layers, reducing model complexity while enhancing interpretability. This structure enables direct insight into the model's decision-making process without relying on post-hoc explainability techniques. Figure 2 shows a schematic of the proposed hierarchical CNN model.

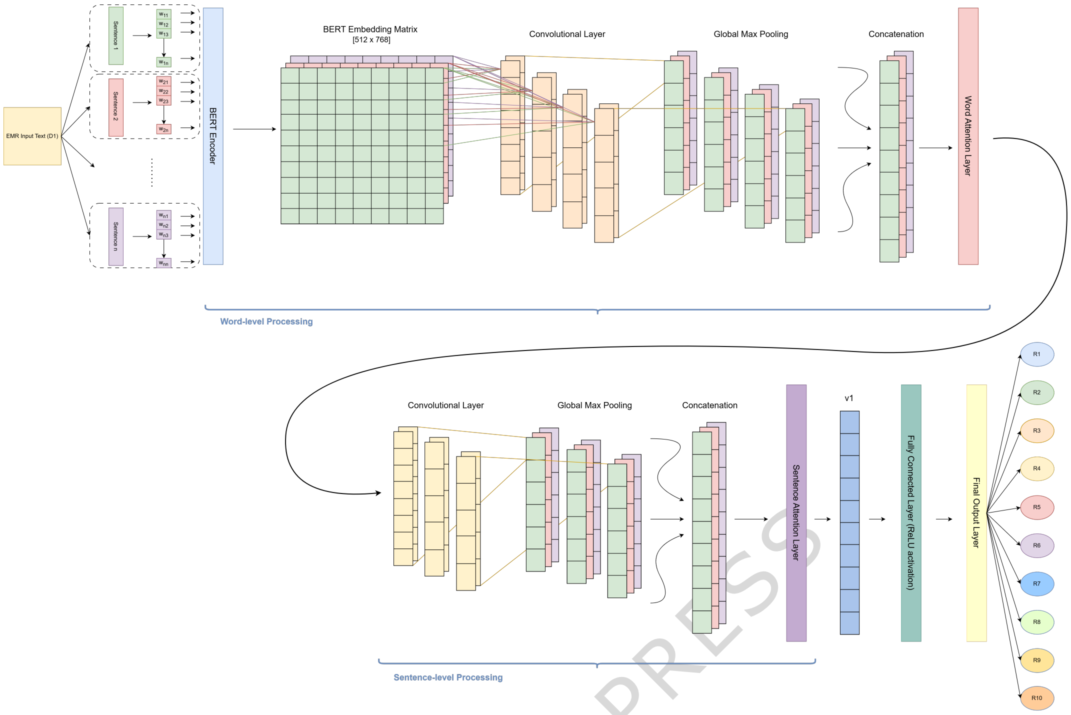


Figure 2. Schematic of the proposed hierarchical CNN model. Each EHR document (e.g. D1) is tokenized into sentences using a custom sentence tokenizer. Each sentence is further tokenized using the BERT tokenizer, generating subword tokens (w_{11} to w_{nn}). These tokens are passed through the BERT encoder, producing an embedding matrix for each sentence of size [512 x 768], where 512 represents the maximum token length and 768 is the hidden size of the BERT-base model. Each example sentence is color-coded throughout the figure for clarity. The BERT-encoded embeddings of each sentence are then processed through four sets of 1D convolutional filters with different kernel sizes. This is followed by global max pooling and concatenation into an n-dimensional feature vector for each sentence. The feature vectors are then passed through a word-level attention layer to produce sentence-level representations. A similar process is applied for sentence-level processing, where the stacked sentence features pass through three sets of 1D convolutional filters in sequence with different kernel sizes, followed by global max pooling and concatenation. This produces an n-dimensional feature vector for each sentence, which passes through the sentence-level attention layer to produce the final document vector (v_1) used for classification by the output layer after being processed through the final fully connected layer.

At the word level, each sentence was passed through a pre-trained BERT model to obtain contextual token embeddings. These embeddings were reshaped to fit a convolutional input format and passed through several 1D convolutional layers with different kernel sizes to capture local n-gram patterns. Each convolutional output underwent max pooling and was concatenated into a fixed-size vector. An attention mechanism was then applied to assign higher weights to more informative features within each sentence. Specifically, attention scores were computed using a tanh-transformed projection followed by a learned context vector, and the final weighted sum of these features forms the sentence representation.

These sentence representations were stacked to form a tensor representing the entire document. This tensor was passed to the sentence-level CNN module, which applied additional 1D convolution operations across the sentence sequence to detect higher-order inter-sentence dependencies. The outputs of these convolutions were pooled and further refined through a second attention mechanism that identified key sentences contributing most to the classification objective. Similar to the word-level attention, sentence-level attention weights were derived using a learned transformation and a context vector. The weighted sum of sentence representations produced a single document vector, which was passed through a fully connected layer and a softmax output layer to yield the final class probabilities.

Data preprocessing

For the BERT-based models (BERT, RoBERTa, DeBERTa, and BioClinicalBERT), we adopted the standard preprocessing pipeline used for transformer-based text classification. Each document was treated as a single flat sequence and tokenized using their corresponding pre-trained tokenizer. Truncation and padding were applied to a fixed maximum input length of 512 tokens, in accordance with model-specific constraints. Attention masks were generated to distinguish between true tokens and padding. Special tokens (e.g., `[CLS]` and `[SEP]`) were automatically added where required, following each model's architecture. All preprocessing steps—including casing, token handling, and normalization—were performed in accordance with the recommended settings of each pre-trained model to ensure compatibility and reproducibility.

Although the hierarchical CNN and transformer models differed in how they structured input (sentence-level vs. flat sequence), both pipelines employed the same BERT-based tokenizer to ensure consistency in token representation. To preserve the hierarchical document structure for the hierarchical CNN, we first applied sentence segmentation before encoding the individual sentences using the BERT tokenizer. Since EHRs often contain irregular punctuation and formatting, we developed a customized sentence tokenizer that segmented text using common end-of-sentence punctuation ('.', '!', '?') and dataset-specific delimiters (';', ','). Finally, the target variable—race/ethnicity—was encoded as a categorical numeric label for classification.

While preprocessing pipelines differed in structure, these differences reflect core architectural design rather than methodological choices. The hierarchical CNN's sentence-based segmentation is integral to its framework and cannot be replicated in standard BERT variants without architectural modification. Importantly, the corpus that was drawn from the social history and risk factor sections largely fell within the 512-token limit of transformer models, ensuring comparable information access across architectures. Although alternative preprocessing strategies (e.g., sliding windows, chunking) exist, our aim was to evaluate each model under its typical deployment configuration to isolate architectural, rather than preprocessing, effects.

Training and hyperparameter tuning

All models were trained and evaluated using 10-fold stratified cross-validation to assess generalizability and ensure class proportions were preserved across training and validation sets. Additionally, 10% of the annotated dataset was held out using a stratified shuffle split as an independent test set for final bias analysis across demographic subgroups (sex, age, and race). The remaining 90% was used for the 10-fold stratified cross-validation. To prevent data leakage, only one entry per patient was retained in the dataset, ensuring that training and validation samples remained independent. We saved the models with the highest validation performance across all folds and used it to evaluate the model performance on the held out sample. Training was conducted using the AdamW optimizer, with weight decay applied to the hierarchical CNN model to reduce overfitting and epsilon applied to the BERT-based models to ensure numerical stability during training. For the hierarchical CNN model, BERT parameters were fine-tuned separately using a learning rate set to 10% of the main model rate to avoid catastrophic forgetting while adapting the pre-trained embeddings to the dataset.

To determine optimal hyperparameters, we performed Bayesian optimization using Optuna, employing the Hyperband Pruner for early stopping of underperforming trials[42]. The final model configurations used for downstream evaluation and inference are provided in Table 1. To enhance computational efficiency and training stability, we implemented gradient checkpointing and automatic mixed precision. For EHR text with long sequences, we applied a sentence cap slightly above the average document length in the training set to maintain training feasibility while minimizing information loss.

Fairness constraints

To assess the impact of fairness-aware interventions on improving equitable predictive performance across race/ethnicity groups, we developed a multi-staged approach that integrated pre-processing, in-processing, and post-processing fairness components[43]. Pre-processing included the use of stratified sampling to ensure representative distributions of race/ethnicity labels, thereby minimizing sampling bias. In-processing fairness was addressed through two mechanisms. First, we used a class-weighted loss function to mitigate the impact of class imbalance by assigning higher weights to underrepresented race categories, thereby ensuring that their contributions were amplified during optimization. We scaled the class weights using min-max normalization[44]. This normalization prevented excessively large weight differences that could destabilize training while ensuring that minority classes retained a higher influence relative to majority classes. Second, to directly optimize for fairness during training, we introduced a fairness-aware loss function that incorporated a penalization term inspired by equalized odds regularization[45] into the model's objective. This fairness term explicitly penalized the model for discrepancies in true and false positive rates across predicted race/ethnicity classes, encouraging the model to minimize disparities in classification performance across groups. The total loss used during training is defined as:

$$\mathscr{L}_{\mathrm{fair}} = \mathscr{L}_{\mathrm{CE}} + \lambda \cdot \mathscr{L}_{\mathrm{PP}} \tag{1}$$

where $\mathscr{L}_{\mathrm{CE}}$ denotes the standard (class-weighted) cross-entropy loss, and λ is a tunable hyperparameter that controls the trade-off between classification performance and fairness. The pairwise parity loss term $\mathscr{L}_{\mathrm{PP}}$ is defined as:

$$\mathscr{L}_{\mathrm{PP}} = \frac{1}{\binom{C}{2}} \sum_{i=1}^{C} \sum_{j=i+1}^{C} \left[(\mathrm{TPR}_i - \mathrm{TPR}_j)^2 + (\mathrm{FPR}_i - \mathrm{FPR}_j)^2 \right] \tag{2}$$

Table 1. Optimized hyperparameters for the trained models

Fairness-unaware models					
Hyperparameter	BERT-base	RoBERTa-base	DeBERTa-base	BioClinicalBERT	Hierarchical CNN
Learning rate	1.68×10^{-5}	1.75×10^{-5}	1.81×10^{-5}	4.41×10^{-5}	1.23×10^{-4}
Weight decay	-	-	-	-	0.01
Epsilon	4.95×10^{-7}	6.25×10^{-8}	9.88×10^{-7}	2.69×10^{-8}	-
Number of training epochs	9	3	9	6	9
Warm-up steps	136	338	58	46	303
Word-level filters	-	-	-	-	45
Word-level filter sizes	-	-	-	-	[1, 6, 6, 6]
Sentence-level filters	-	-	-	-	50
Sentence-level filter sizes	-	-	-	-	[1, 6, 6]
Word-level attention dimension	-	-	-	-	170
Sentence-level attention dimension	-	-	-	-	135
Dropout rate	-	-	-	-	0.24
Fully connected layer size	-	-	-	-	145
Lambda fairness	-	-	-	-	-

Fairness-aware models					
Hyperparameter	BERT-base	RoBERTa-base	DeBERTa-base	BioClinicalBERT	Hierarchical CNN
Learning rate	1.56×10^{-5}	3.70×10^{-5}	1.15×10^{-5}	2.08×10^{-5}	8.34×10^{-5}
Weight decay	-	-	-	-	0.01
Epsilon	1.87×10^{-8}	1.57×10^{-8}	2.67×10^{-8}	3.24×10^{-7}	-
Number of training epochs	7	3	8	8	6
Warm-up steps	99	295	236	400	167
Word-level filters	-	-	-	-	30
Word-level filter sizes	-	-	-	-	[1, 6, 6, 1]
Sentence-level filters	-	-	-	-	55
Sentence-level filter sizes	-	-	-	-	[1, 6, 6]
Word-level attention dimension	-	-	-	-	50
Sentence-level attention dimension	-	-	-	-	75
Dropout rate	-	-	-	-	0.24
Fully connected layer size	-	-	-	-	75
Lambda fairness	2.92×10^{-3}	1.96×10^{-1}	2.60×10^{-2}	2.49×10^{-2}	1.39

where C is the total number of race/ethnicity classes, and TPR_i, FPR_i represent the true and false positive rates for class i, respectively. The fairness penalty is computed by taking all possible pairs of race/ethnicity classes and measuring the squared differences in their true and false positive rates, explicitly penalizing pairwise disparities across groups to encourage equitable predictive performance. This formulation enables joint optimization of accuracy and inter-class performance parity, a fairness criterion adapted for multi-class classification tasks where the output itself is a sensitive attribute. While prior work has explored fairness-aware regularization strategies and constraint-based optimization for fairness metrics such as equalized odds and demographic parity[46–49], our approach introduces a simple, interpretable formulation tailored for multi-class classification tasks where the target label corresponds to a sensitive demographic attribute (i.e., race).

Finally, we implemented a post-processing threshold adjustment step to mitigate any residual biases that persisted after training. Class-specific thresholds were calibrated to improve precision for underrepresented classes, with higher decision thresholds applied to these classes and lower thresholds used for overrepresented ones (ranging from 0.50 to 0.90). To evaluate the impact of fairness interventions on both improving equitable predictive performance and overall classification performance, we trained two versions of each model: a fairness-unaware model and a fairness-aware variant with constraints applied. The fairness-unaware models were trained without class balancing techniques to establish a baseline representing common deployment scenarios where imbalance is not explicitly addressed. This allowed us to assess whether fairness-aware techniques improved the model's predictive balance across racial groups or introduced trade-offs in overall performance.

Bias analysis

Assessing potential bias introduced from active learning

To assess potential biases introduced by the active learning sampling process, we conducted a series of distributional analyses comparing the active learning samples to both the baseline annotated sample and the full UTOPIAN database. Specifically, we examined subgroup representation across sex and age attributes, as well as patient coverage, to evaluate whether the sampling strategy disproportionately favored or excluded specific populations. For each sensitive attribute, we computed sampling parity ratios with 95% confidence intervals to measure relative subgroup inclusion, along with Jensen-Shannon divergence and Shannon diversity index to quantify distributional shifts and within-group diversity, respectively. Chi-square tests were also performed to assess statistical significance of the observed differences in group distributions. We analyzed patient-level coverage by calculating the proportion of patients per provider in each dataset and testing for distributional shifts using chi-square tests.

Assessing performance bias among sex and age groups

We evaluated model performance across demographic subgroups defined by age—young adults (18–34 years), adults (35–49 years), seniors (50–64 years), and elders (65+ years)—and sex (female and male), including their intersections. For each subgroup, we computed standard classification metrics (F1-score, precision, and recall) to assess performance variability. To disentangle the source of bias, we first computed the distribution of sex and age groups within each predicted race category and compared them to their baseline proportions in the annotated dataset. This allowed us to assess whether the model mirrors or amplifies existing distributional biases. We then evaluated true and false positive rates across sensitive attributes (i.e., sex and age) for each predicted race, to assess whether the model introduced algorithmic disparities in performance (i.e., violated equalized odds)[43]. This analysis was conducted using the best-performing model (selected based on cross-validation) and evaluated on the held-out 10% of the annotated dataset.

Evaluation criteria

Given the skewed label distribution in our dataset, we evaluated model performance using precision, recall, F1 score, average precision, and the area under the precision-recall curve. Since the choice between optimizing for precision or recall typically depends on the clinical application, which in this case is unknown, we selected the F1 score as the primary evaluation metric, as it balances both precision and recall[50]. To assess overall model effectiveness, we report both micro- and macro-averaged scores across all race/ethnicity labels. Macro-averaging treats all classes equally by computing the unweighted mean of per-class metrics, allowing minority classes to contribute equally to the evaluation[51]. In contrast, micro-averaging aggregates contributions from all samples, which can bias results toward the majority class in imbalanced datasets. Given the class imbalance in our data, the macro-F1 score was deemed the most appropriate metric, as it provides a fairer assessment of model performance across all classes.

To assess whether the fairness-aware techniques improved the model's predictive balance across racial groups, we evaluated the disparities in false positive rates and false negative rates to help identify if any group was disproportionately affected by higher error rates. We also reported the representation ratio, which evaluates whether the model's predicted distributions across racial groups reflect the actual distribution in the dataset, allowing us to assess whether the model perpetuates or amplifies the biases present in the data.

To evaluate whether the differences in model performance were statistically significant, we employed a combination of parametric and non-parametric tests, depending on the distribution of the data. Shapiro-Wilk tests[52] were conducted to assess data normality. If normality was satisfied ($p < 0.05$, for all metrics), we used repeated measures ANOVA[53], followed by Tukey's honestly significant difference test[54] to control for multiple comparisons in pairwise model evaluations. If normality was violated, we applied the Friedman test[55], a non-parametric alternative, with the Nemenyi post-hoc test[56] to identify significant differences while adjusting for multiple comparisons. To directly compare the fairness-aware and fairness-unaware variants of each model, we conducted Wilcoxon signed-rank tests[57] on paired samples for each model–metric combination. These tests evaluated whether the inclusion of fairness constraints led to statistically significant changes in performance.

Since bias metrics often involve group-wise comparisons across categorical attributes (i.e., sex, age), we used the Mann–Whitney U test[58], a rank-based method that does not assume normality, for comparisons between two demographic groups (i.e., sex), and the Kruskal–Wallis test[59], a generalization of the Mann–Whitney U test for comparisons across more than two groups (i.e., age groups). To statistically assess whether the demographic composition of predicted race classifications deviated significantly from the actual dataset distribution, we employed the chi-squared goodness-of-fit test[60]. This test compares the observed subgroup proportions for each race to the expected proportions in the actual data.

Code Availability

All models were implemented using the PyTorch framework (version 2.3.1+cu121)[61], with transformer-based architectures developed using the HuggingFace Transformers library (version 4.37.1)[62]. Model development and analysis were conducted in Python 3.10.12 using NumPy 1.26.4, pandas 2.1.1, scikit-learn 1.4.dev0, Matplotlib 3.8.1, Seaborn 0.13.0, and NLTK 3.8.1. Training was performed on an NVIDIA Quadro RTX 6000 GPU using CUDA 12.2 (driver version 535.247.01). Hyperparameters and training configurations for all models are provided in the Methods section and summarized in Table 1.

The code for the active learning pipeline used for data annotation is publicly available at `https://github.com/seperahm/EMR_Race_Classification`. The remaining modeling code, developed for model training and fairness-aware loss implementation, are stored within the secure University of Toronto Data Safe Haven environment alongside the study data and cannot currently be exported for public release following archival of the environment under institutional privacy and security regulations. All transformer-based models used are standard, publicly available pre-trained architectures, and the hierarchical CNN—the primary methodological contribution of this work—is fully specified in the Methods section, including architectural details, optimized hyperparameters, and training procedures, enabling independent reimplementation.

Researchers seeking further methodological clarification or architecture-level guidance may contact the corresponding author for additional details or code review under appropriate data-sharing agreements.

Results

Data annotation using active learning

Figure 3 presents the distribution of race/ethnicity labels across the baseline annotated sample, the active learning samples, and the full dataset (a combination of the two sampling strategies). During the initial labeling phase, random sampling resulted in 61 race-present labels out of 4,375 annotated entries (1.39%). In contrast, active learning produced 910 race-present labels from 1,386 annotated entries (65.66%), reflecting a significant increase in sample selection. To create the full dataset, the baseline annotated sample was adjusted by downsampling the absent class from 4,314 to 3,019 instances ($\sim$30%), reducing its prevalence while preserving representational diversity. Furthermore, we excluded the newly identified absent samples from active learning in the final training set since the dataset already contained a substantial number of absent labels. This resulted in a combined dataset of 3,990 samples, composed of 3,019 absent labels (75.66%) and 971 present labels (24.34%). The dot plot visualizing the distribution of race/ethnicity labels across the three data subsets shows that certain classes, such as East Asian, and even more prominently, white, were sampled more frequently, while others, like Indigenous, remained underrepresented even after active learning. A detailed comparison of observed and expected sample counts by race/ethnicity groups across the active learning samples is presented in Supplementary Table S1. Although the active learning process improved the proportion of race-present to race-absent samples, substantial class imbalance persisted (illustrating the need for incorporating fairness-aware modeling strategies).

Dataset demographics

The final training dataset consisted of 3,990 patients, with a sex distribution of 57.82% female and 42.18% male. Race was documented for 24.34% of patients, with the majority classified as white (9.27%), followed by East Asian (3.31%), Black (2.61%), Middle Eastern (2.31%), South Asian (2.28%), Southeast Asian (1.80%), mixed heritage (1.53%), Latin American (1.05%), and Indigenous (0.18%). For age group distribution, 22.18% of patients were categorized as young adults (18–34 years), 27.34% as adults (35–49 years), 23.63% as seniors (50-64 years), and 26.84% as elders (65+ years).

Model performance

Overall performance across models

Table 2 presents a comparative evaluation of the five models—BERT, RoBERTa, DeBERTa, BioClinicalBERT, and the hierarchical CNN—using micro- and macro-averaged performance metrics across multiple evaluation measures. These metrics were calculated to represent overall model performance, so the absent class was considered in the calculation. For the transformer-based models, we employed their standard pre-trained architectures as described in the literature[63–66] without modifications to their internal structure. The hierarchical CNN significantly outperformed all transformer-based models across most evaluation metrics ($p < 0.05$), achieving the highest macro-averaged F1 score (98.44%), micro-averaged F1 score (99.64%), and particularly excelling in macro-averaged precision (98.81%), average precision (93.85%), and area under the precision-recall curve (95.76%). In contrast, DeBERTa achieved the highest macro- and micro-averaged recall (99.43% and 99.70%, respectively), but this did not consistently translate into higher F1 or area under the precision-recall curve scores. Among the transformer models, BioClinicalBERT, RoBERTa, and BERT performed competitively in micro-averaged scores, with BioClinicalBERT showing strong micro- and macro-averaged average precision and area under the precision-recall curves compared to the other transformers, but lower macro-averaged recall and F1 scores.

Behavior on longer notes. We quantified note lengths in the held-out test set using each model's corresponding tokenizer. Token counts varied across models due to differences in tokenizer vocabularies and subword segmentation algorithms (Supplementary Table S1). For all models, only one note (0.25% of test data) exceeded 512 tokens and would be subject to truncation. The 99th percentile ranged from 292-309 tokens depending on the tokenizer, confirming that the vast majority of notes fell well within the 512-token window and that truncation affected minimal data.

To directly probe model behavior on the long-note tail, we conducted a targeted case analysis by selecting the longest test notes (top 15) for each model's tokenizer. Among the 15 longest notes in the held-out test set, 10 contained no explicit mention of race, while 5 included race-related terms. On this subset, fairness-unaware models achieved accuracies of 93.3% (BERT), 80.0% (DeBERTa), 86.7% (RoBERTa), and 93.3% (BioClinicalBERT). In comparison, the corresponding fairness-aware variants correctly classified 14 of the 15 notes (93.3% accuracy) across all models. This pattern suggests that, for longer inputs, fairness-aware training may improve robustness without incurring the performance degradation often associated with fairness–accuracy trade-offs.

Sampling method	Race present	Race absent	Total annotated
Baseline annotated sample	61 (1.39%)	4314 (98.61%)	4375
Active learning samples	910 (65.66%)	476 (34.34%)	1386
Full dataset (final training dataset)	971 (24.34%)	3019 (75.66%)	3990

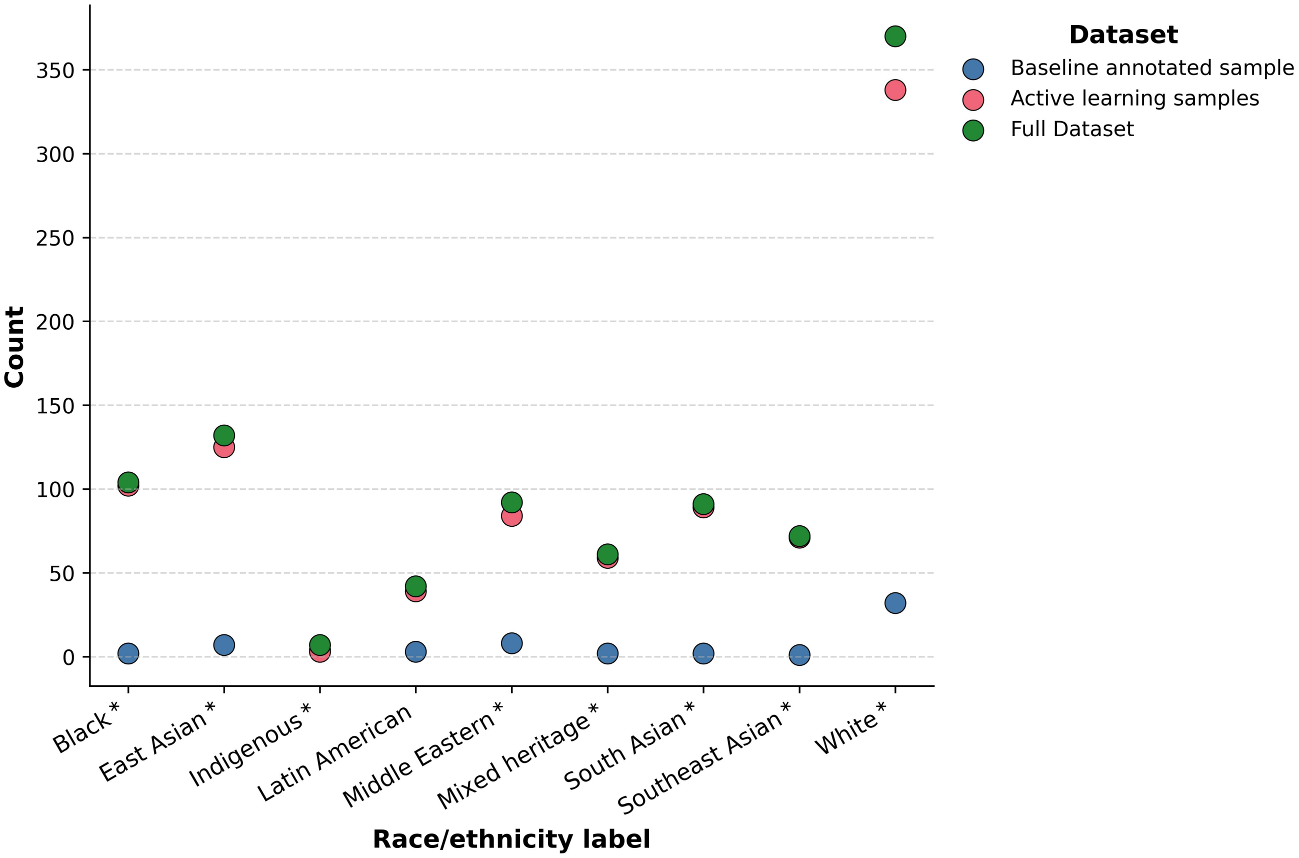


Figure 3. Overview of race/ethnicity annotation across the dataset. The table summarizes the proportion of records with present vs. absent race/ethnicity labels across the baseline sample, active learning subset, and final training data after downsampling. The dot plot shows the count distribution of individual race/ethnicity classes across the three sampling phases. Asterisks (*) in the plot indicate statistically significant deviations from the expected values under proportional sampling. Significance was determined using standardized residuals from a two-sided chi-square goodness-of-fit test, with $|z| > 2$ corresponding approximately to $p < 0.05$. Exact p-values for each group comparison are provided in Table S1.

Table 2. Model performance across 10-fold cross-validation on the training set (90% of data), reported as micro- and macro-averages for each evaluation metric.

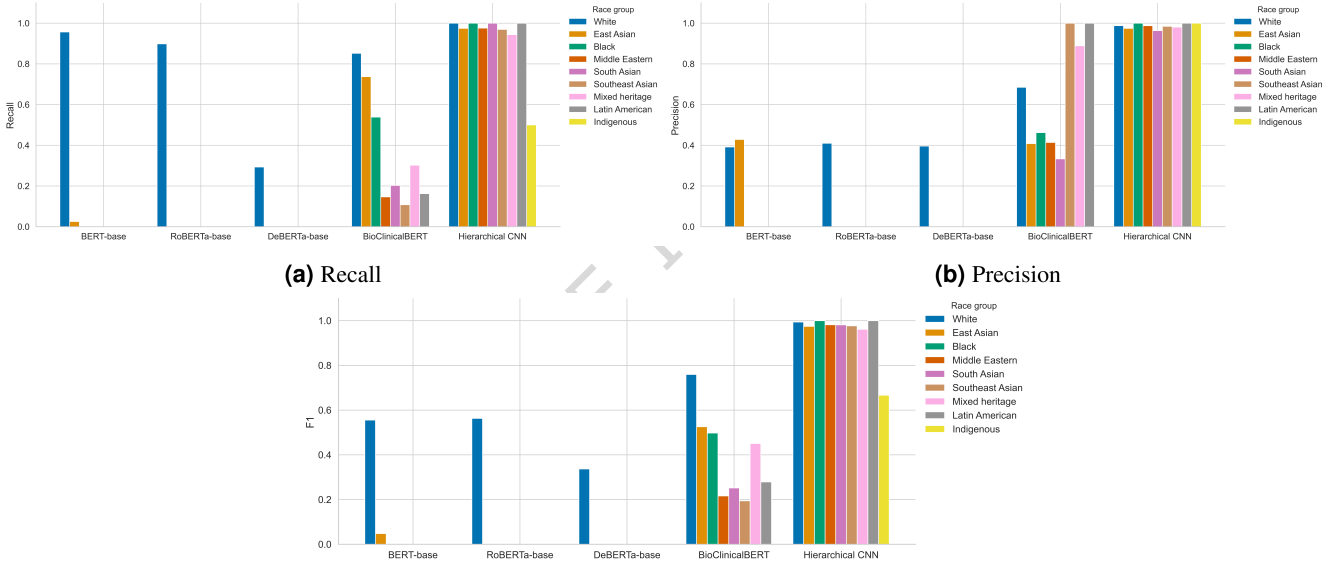
Model	Recall		Precision		F1 Score		Avg Prec		AUC PR	
	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro
BERT-base	83.98 (0.183)	97.68 (0.023)	63.13 (0.111)	84.38 (0.006)	66.84 (0.131)	90.54 (0.012)	27.39 (0.024)	92.05 (0.007)	27.33 (0.033)	92.19 (0.006)
RoBERTa-base	94.78 (0.064)	98.50 (0.013)	67.97 (0.014)	83.68 (0.012)	76.58 (0.023)	90.49 (0.012)	25.53 (0.020)	90.76 (0.010)	25.64 (0.031)	90.74 (0.011)
DeBERTa-base	**99.43** (0.011)	**99.70** (0.006)	73.58 (0.037)	78.27 (0.037)	83.54 (0.043)	87.64 (0.021)	20.23 (0.036)	84.87 (0.056)	21.29 (0.027)	85.38 (0.049)
BioClinicalBERT	63.55 (0.156)	92.99 (0.030)	62.56 (0.153)	88.78 (0.032)	58.99 (0.139)	90.79 (0.024)	53.47 (0.138)	94.84 (0.015)	53.97 (0.149)	95.21 (0.016)
Hierarchical CNN	98.48 (0.036)**	99.69 (0.006)**	**98.81** (0.019)††‡**	**99.60** (0.007)†‡‡§§	**98.44** (0.031)††‡**	**99.64** (0.007)†‡§§*	**93.85** (0.060)†‡‡§§	**99.88** (0.003)†‡‡§§	**95.76** (0.052)†‡‡§§	**99.88** (0.003)†‡‡§§

Area under the precision-recall curve is denoted as AUC PR. Values are reported as percentages, with standard deviations in parentheses. Bold-faced numbers represent the highest performance among the models. Superscripts indicate significance ($p < 0.05$) between the best-performing model (hierarchical CNN) and the other models as follows: † compared to BERT, ‡ compared to RoBERTa, § compared to DeBERTa, * compared to BioClinicalBERT. Double symbols (e.g., ††) indicate strong significance at $p < 0.001$. Of note, statistical differences were observed between BioClinicalBERT and DeBERTa in most metrics and some cases between BioClinicalBERT and RoBERTa. *Statistical analysis:* Normality of performance distributions was assessed using the Shapiro–Wilk test. Where normality was satisfied, two-sided repeated-measures ANOVA followed by Tukey's honestly significant difference test was used for pairwise comparisons, adjusting for multiple comparisons. Where normality was violated, the two-sided Friedman test with Nemenyi post-hoc correction was applied. Exact p-values were below the indicated thresholds ($p < 0.05$, $p < 0.001$) for all significant comparisons; non-significant comparisons exceeded $p > 0.05$.

Notably, the single note exceeding the 512-token limit imposed by standard transformer architectures (true label: Black) was misclassified by all fairness-unaware models, whereas all fairness-aware variants correctly predicted the true class. Although based on a single example, this reversal may suggest that fairness-aware optimization could reduce sensitivity to truncation effects, potentially by discouraging over-reliance on early or majority-associated tokens. Given the limited number of very long notes in this primary care dataset, these observations should be interpreted cautiously. Nonetheless, they indicate that truncation does not necessarily preclude accurate classification and that fairness-aware training may confer benefits beyond group-level parity, including improved generalization to distributional edge cases. Confirmation of this effect will require evaluation in datasets with a higher prevalence of long or multi-document clinical records.

Per-class performance by race

Figure 4 displays grouped bar charts comparing recall, precision, and F1 scores across all race categories for the five evaluated models. The hierarchical CNN consistently outperformed other models in all groups, especially the underrepresented ones, achieving the highest scores across all three metrics. In contrast, transformer-based models (particularly, BERT, RoBERTa, and DeBERTa) demonstrated substantial variability and generally underperformed on non-white classes, with several instances of zero precision, recall, and F1 scores. BioClinicalBERT showed comparatively higher precision for certain minority groups (i.e., Southeast Asian, mixed heritage, and Latin American) but suffered from notably low recall, indicating that the model tended to miss relevant cases.



(a) Recall

(b) Precision

(c) F1 score

Figure 4. Per-class performance comparison across models. Grouped bar charts show per-class recall (a), precision (b), and F1 score (c) for five models: BERT-base, RoBERTa-base, DeBERTa-base, BioClinicalBERT, and the hierarchical CNN. Bars represent mean performance across cross-validation folds per race group. The race categories are ordered by their frequency in the training dataset, from most to least represented.

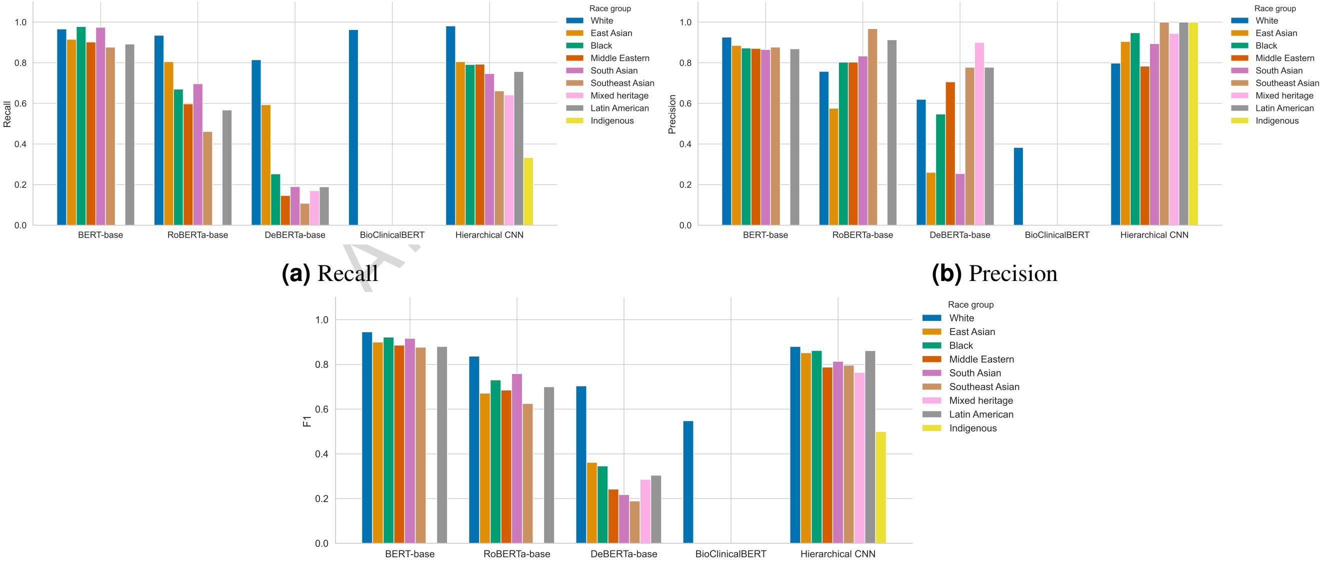
Fairness analysis

Impact of fairness constraints on model performance

Table 3 presents a comparative analysis of performance variations across the fairness-aware models (i.e., with the addition of fairness constraints) and their fairness-unaware counterparts (i.e., without fairness constraints). These metrics were calculated to represent overall model performance, so the absent class was considered in the calculation. The inclusion of fairness constraints led to varying impacts across models and race groups. Overall, fairness-aware BERT consistently demonstrated strong performance, outperforming all other fairness-aware models in macro- and micro-averaged precision (90.20% and 96.71%, respectively) and F1 scores (91.48% and 97.53%, respectively), while achieving the second-best performance in recall, average precision, and area under the precision-recall curve. The hierarchical CNN showed competitive results, outperforming all other models in macro- and micro-averaged average precision and area under the precision-recall curve, and ranking second in macro- and micro-averaged F1 (87.79% and 96.98%, respectively) and precision scores (89.38% and 95.91%, respectively). Interestingly, BioClinicalBERT exhibited the highest macro- and micro-averaged recall performance among fairness-aware models (97.96% and 99.23%, respectively)— a marked contrast to its fairness-unaware counterpart, which had the lowest macro- and micro-averaged recall values (63.55% and 92.99%, respectively)— though this came at the cost of a notable drop in micro-averaged precision (from 88.78% to 84.30%) and generally a low macro-precision overall (68.21% and 62.56% for fairness-aware and fairness-unaware versions, respectively). While the hierarchical CNN maintained strong micro-level performance post-fairness constraints, its macro-level metrics dropped substantially, from the 93–98% range in the fairness-unaware version to 82–90% in the fairness-aware version, indicating that fairness regularization may have compromised its ability to generalize across underrepresented classes.

Macro-level results revealed that BERT, RoBERTa, and, to an extent, DeBERTa and BioClinicalBERT benefited from fairness constraints, particularly regarding average precision and area under the precision-recall curve for the first three models. For example, BERT improved its macro-averaged average precision from 27.39% to 80.53%, and area under the precision-recall curve from 27.33% to 81.99%. RoBERTa (in all macro-averaged metrics except recall) and BioClinicalBERT (in macro-averaged precision, recall, and F1 score) exhibited similar upward trends, while DeBERTa suffered substantial drops in macro-averaged recall, precision, and F1 score under fairness constraints but had an increase in average precision and area under the precision-recall curve scores. At the micro-level, fairness-aware BERT again showed consistent improvements across all metrics, significantly outperforming its fairness-unaware counterpart. Most transformer models, including DeBERTa and RoBERTa, also exhibited micro-level gains, particularly in precision and F1 score, with the exception of BioClinicalBERT, which experienced drops in micro-averaged precision, average precision, and area under the precision-recall curve.

Figure 5 illustrates the per-class precision, recall, and F1 scores for all five models with fairness constraints. Fairness constraints substantially improved per-class performance for BERT, RoBERTa, and DeBERTa and reduced inter-group performance disparities, especially for BERT. However, performance inconsistency across race/ethnicity groups persisted or even worsened in some models; for example, RoBERTa, and even more so, DeBERTa, exhibited high variance across race groups, especially in recall and F1 scores, where performance on non-white groups often lagged. BioClinicalBERT had a drastic drop in per-class performance, with scores collapsing to zero for all non-white groups. The hierarchical CNN also saw a modest increase in disparities between groups when fairness constraints were applied, with its F1 score range widening by nearly 5% compared to its unconstrained form.



(a) Recall

(b) Precision

(c) F1 score

Figure 5. Per-class performance across fairness-aware models. Grouped bar charts compare per-class recall (a), precision (b), and F1 score (c) for the fairness-aware versions of the five models. Bars represent mean performance across cross-validation folds per race group. The race categories are ordered by their frequency in the training dataset, from most to least represented.

Performance inconsistencies were observed across race groups that could not be fully explained by sample size differences alone. For example, fairness-aware RoBERTa and DeBERTa achieved substantially higher recall for East Asian patients compared to Black patients—80.51% vs. 67.03% for RoBERTa and 59.32% vs. 25.27% for DeBERTa—even though the difference in sample size between the two groups was only ∼23% (132 vs. 104). This recall gap was larger than the difference observed between white (n = 370) and East Asian patients in both models. Fairness-unaware BERT (see Figure 4) also exhibited

Table 3. Performance comparison of fairness-aware and fairness-unaware models across 10-fold cross-validation on the training set (90% of data), reported as micro- and macro-averages for each evaluation metric.

Fairness-aware models

Model	Recall		Precision		F1 Score		Avg Prec		AUC PR	
	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro
BERT-base	93.75 (0.020)	98.36 (0.005)	**90.20 (0.027)***	**96.71 (0.005)***	**91.48 (0.018)***	**97.53 (0.005)***	80.53 (0.028)*	98.21 (0.004)*	81.99 (0.040)*	98.46 (0.003)*
RoBERTa-base	83.83 (0.087)*	96.92 (0.014)	83.45 (0.093)*	92.91 (0.046)*	80.51 (0.115)	94.83 (0.030)*	67.02 (0.155)*	96.19 (0.023)*	68.34 (0.150)*	96.52 (0.022)*
DeBERTa-base	64.66 (0.194)*	92.89 (0.037)*	57.46 (0.175)	86.79 (0.039)*	57.17 (0.171)*	89.68 (0.031)	44.80 (0.172)*	94.00 (0.020)*	44.75 (0.191)*	94.22 (0.020)*
BioClincalBERT	**97.96 (0.015)***	**99.23 (0.003)***	68.21 (0.006)	84.30 (0.003)*	76.82 (0.008)*	91.16 (0.003)	24.97 (0.014)*	91.87 (0.005)*	25.09 (0.032)*	91.83 (0.007)*
Hierarchical CNN	90.19 (0.135)	98.20 (0.023)	89.38 (0.139)*	95.91 (0.057)*	87.79 (0.145)*	96.98 (0.037)*	**82.47 (0.194)***	**98.68 (0.023)***	**83.74 (0.213)***	**98.68 (0.023)***

Fairness-unaware models

Model	Recall		Precision		F1 Score		Avg Prec		AUC PR	
	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro
BERT-base	83.98 (0.183)	97.68 (0.023)	63.13 (0.111)	84.38 (0.006)	66.84 (0.131)	90.54 (0.012)	27.39 (0.024)	92.05 (0.007)	27.33 (0.033)	92.19 (0.006)
RoBERTa-base	94.78 (0.064)	98.50 (0.013)	67.97 (0.014)	83.68 (0.012)	76.58 (0.023)	90.49 (0.012)	25.53 (0.020)	90.76 (0.010)	25.64 (0.031)	90.74 (0.011)
DeBERTa-base	99.43 (0.011)	99.70 (0.006)	73.58 (0.037)	78.27 (0.037)	83.54 (0.043)	87.64 (0.021)	20.23 (0.036)	84.87 (0.056)	21.29 (0.027)	85.38 (0.049)
BioClinicalBERT	63.55 (0.156)	92.99 (0.030)	62.56 (0.153)	88.78 (0.032)	58.99 (0.139)	90.79 (0.024)	53.47 (0.138)	94.84 (0.015)	53.97 (0.149)	95.21 (0.016)
Hierarchical CNN	98.48 (0.036)	99.69 (0.006)	98.81 (0.019)	99.60 (0.007)	98.44 (0.031)	99.64 (0.007)	93.85 (0.060)	99.88 (0.003)	95.76 (0.052)	99.88 (0.003)

Fairness-aware values reflect models trained with fairness constraints. Area under the precision-recall curve is denoted as AUC PR. Values are reported as percentages, with standard deviations in parentheses. Bold-faced numbers indicate the highest performance across all fairness-aware models. Superscripts indicate statistically significant differences ($p < 0.05$) between each fairness-aware model and its corresponding baseline (fairness-unaware) model, based on the two-sided Wilcoxon signed-rank test. Superscripts are applied only to fairness-aware models, reflecting statistically significant differences from their baseline counterparts. Exact p-values were below the indicated threshold ($p < 0.05$) for all significant comparisons; non-significant comparisons exceeded $p < 0.05$.

stark differences between East Asian and Black patients, achieving zero precision for Black patients and 42.86% precision for East Asian patients. Disparities were also evident between other groups with comparable sample sizes. For example, fairness-aware DeBERTa had a 45.16% higher precision rate for Middle Eastern patients compared to South Asian patients despite a difference of only one training sample. Similarly, South Asian (n = 91) and Southeast Asian (n = 71) groups showed widely different results across many transformer models (fairness-aware and fairness-unaware). Underrepresented groups such as mixed heritage and Indigenous were consistently underserved by transformer-based models. Both BERT and RoBERTa scored zero across all metrics for these groups, and DeBERTa and BioClinicalBERT failed entirely on the Indigenous class. While the extremely low performance for Indigenous patients may reflect the small sample size, the persistent failure on the mixed heritage class—despite the presence of smaller or comparably sized groups that did not show this behavior—suggests additional contributing factors.

Impact of fairness constraints on bias

To examine group-level disparities, we calculated representation ratios for each racial and ethnic group, defined as the predicted frequency of a class divided by its true frequency. Table 4 presents these ratios across all models under fairness-aware and fairness-unaware training conditions. Among fairness-unaware models, transformer-based architectures such as BERT and RoBERTa showed stark disparities, with predictions overwhelmingly concentrated on white patients (representation ratios of 2.45 and 2.20, respectively) and nearly all other groups receiving ratios of 0.00. In contrast, their fairness-aware counterparts exhibited substantially improved representational balance across most groups. DeBERTa similarly benefited from fairness-aware training, but its outputs still skewed heavily toward majority classes, overrepresenting white (1.32) and East Asian (2.31) patients while underrepresenting others. Interestingly, BioClinicalBERT reversed this pattern: while the fairness-unaware version provided moderate representation across several groups (e.g., white = 1.25, Black = 1.17, South Asian = 0.60), fairness-aware training caused a collapse in diversity, with predictions skewed almost exclusively toward white patients. The hierarchical CNN was the most balanced model in its fairness-unaware form, maintaining ratios near 1.00 across groups; however, applying fairness constraints introduced slight overrepresentation of white patients (1.24) and underrepresentation of others.

Table 4. Representation ratios across race/ethnicity groups for fairness-aware and fairness-unaware models.

Fairness-aware models

Model	White	East Asian	Black	Middle Eastern	South Asian	Southeast Asian	Mixed heritage	Latin American	Indigenous
BERT-base	1.04 (0.058)	1.03 (0.126)	1.12 (0.104)	1.04 (0.191)	1.13 (0.118)	1.00 (0.184)	0.00 (0.000)	1.02 (0.242)	0.00 (0.000)
RoBERTa-base	1.23 (0.421)	1.38 (0.970)	0.84 (0.504)	0.76 (0.460)	0.83 (0.500)	0.46 (0.405)	0.00 (0.000)	0.65 (0.489)	0.00 (0.000)
DeBERTa-base	1.32 (0.658)	2.31 (1.582)	0.45 (0.492)	0.22 (0.354)	0.75 (0.586)	0.15 (0.309)	0.17 (0.385)	0.22 (0.478)	0.00 (0.000)
BioClinicalBERT	2.52 (0.109)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)
Hierarchical CNN	1.24 (0.517)	0.89 (0.399)	0.83 (0.445)	1.03 (0.440)	0.85 (0.420)	0.68 (0.451)	0.68 (0.426)	0.74 (0.409)	0.20 (0.000)

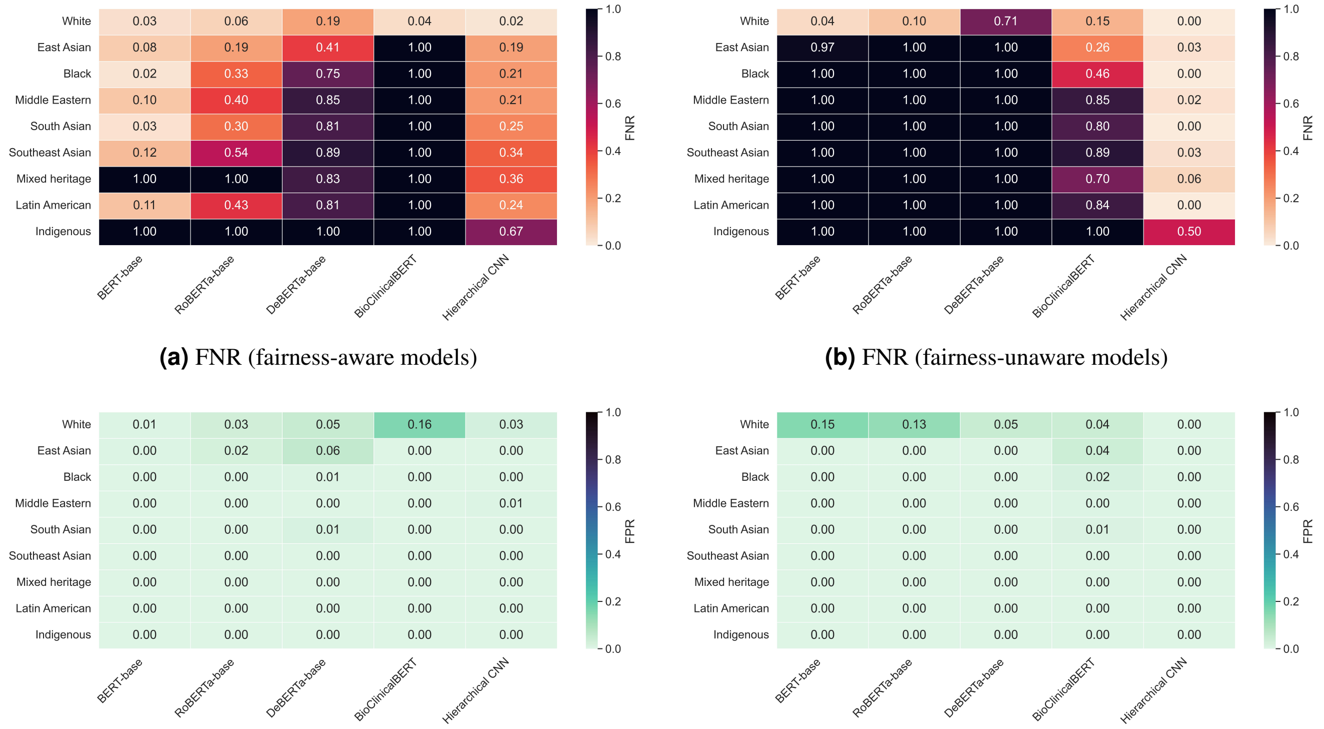
Fairness-unaware models

Model	White	East Asian	Black	Middle Eastern	South Asian	Southeast Asian	Mixed heritage	Latin American	Indigenous
BERT-base	2.45 (0.100)	0.06 (0.089)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.02 (0.053)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)
RoBERTa-base	2.20 (0.411)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)
DeBERTa-base	0.73 (1.186)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)
BioClinicalBERT	1.25 (0.465)	1.81 (0.874)	1.17 (0.615)	0.36 (0.337)	0.60 (0.507)	0.10 (0.316)	0.32 (0.266)	0.15 (0.269)	0.00 (0.000)
Hierarchical CNN	1.01 (0.031)	1.00 (0.056)	1.00 (0.000)	0.99 (0.040)	1.04 (0.058)	0.99 (0.105)	0.96 (0.207)	1.00 (0.000)	0.30 (0.000)

Fairness-aware versions reflect models trained with fairness constraints. Representation ratios were calculated as the predicted frequency of each class divided by its true frequency. Values are reported as mean (standard deviation) over 10-fold cross-validation. A value close to 1.0 indicates proportional representation. Values above 1.0 indicate over-representation, while values below 1.0 indicate under-representation.

Figure 6 illustrates group-level disparities in false negative and false positive rates across racial groups. False negative rate analysis showed that all models consistently struggled to correctly identify Indigenous patients, even when fairness constraints were applied. The fairness-unaware hierarchical CNN had the lowest false negative rates overall, ranging from 0.00 to 0.06 for most groups, but reached 0.50 for Indigenous patients. It also exhibited fewer disparities across groups than the fairness-aware version of BERT. Among the fairness-unaware transformer models, false negative rates were generally high for non-white groups, while DeBERTa performed poorly across all race categories. Fairness-unaware BioClinicalBERT showed a sharp jump in false negative rate from 0.46 for Black patients to 0.85 for Middle Eastern patients, despite a close range in sample sizes (n = 104 vs. 92). Furthermore, BioClinicalBERT performed better on identifying mixed heritage patients (0.70) than on other similarly represented groups, suggesting again that performance disparities were not solely driven by sample size.

With fairness constraints applied, BERT, RoBERTa, and DeBERTa saw substantial reductions in false negative rates, particularly BERT and RoBERTa. However, disparities persisted. Fairness-aware DeBERTa identified white and East Asian patients more accurately (false negative rates of 0.19 and 0.41), while false negative rates for all other groups ranged from



(a) FNR (fairness-aware models)

(b) FNR (fairness-unaware models)

(c) FPR (fairness-aware models)

(d) FPR (fairness-unaware models)

Figure 6. False negative and false positive rate disparities across racial groups for fairness-aware and fairness-unaware models. Heatmaps show the average false negative rate (FNR) and false positive rate (FPR) across the racial groups for the five models over 10-fold cross-validation. Panels (a) and (b) compare group-level false negative rates for fairness-aware and fairness-unaware models, respectively, reflecting disparities in recall. Panels (c) and (d) present false positive rates, capturing over-classification errors. Darker shades indicate higher error rates, with greater intensity representing more severe group-level misclassification.

0.75 to 1.00. Fairness-aware BioClinicalBERT became highly biased toward white patients, showing a sharp imbalance in recall. The hierarchical CNN, when fairness constraints were applied, also exhibited increased false negative rates, especially for Indigenous (0.67), mixed heritage (0.36), and Southeast Asian (0.34) groups. False positive rates were generally much lower than false negative rates across all models, with most values near zero. White patients consistently had the highest false positive rates, especially in transformer-based models, suggesting a tendency to over-predict this category. The fairness-unaware hierarchical CNN did not produce any false positives, reflecting high specificity. Slightly elevated false positive rates were observed in fairness-unaware BioClinicalBERT and fairness-aware DeBERTa.

Bias diagnostics

Analysis of potential bias introduced from active learning

The baseline annotated sample was carefully constructed to reflect the broader UTOPIAN database. Using stratified random sampling, it preserved key distributions—such as age, sex, and EHR start date—and ensured representation from all physicians and clinics. Table 5 reports distributional metrics such as sampling parity, Jensen-Shannon divergence, Shannon diversity index, and chi-square p-values, comparing the distribution of sex and age groups across the UTOPIAN database, the baseline annotated sample, and the active learning subset. While the baseline sample closely mirrored the demographic composition of the full dataset, the active learning process introduced moderate shifts in subgroup representation. For instance, sampling parity for adults aged 35–49 reached 1.29, indicating overrepresentation in the active learning set, while young adults aged 18-34 were underrepresented with a parity score of 0.84. A similar pattern was observed for sex: males showed slight underrepresentation (sampling parity = 0.90 vs. 1.01 in the baseline), while females were proportionally overrepresented (1.08 vs. 0.99). These patterns are supported in Supplementary Figure S1, which visualizes the proportional distributions of sex and age groups across the UTOPIAN database, the baseline annotated sample, and the active learning subset.

The other distributional metrics corroborate these findings. The Jensen-Shannon divergence for age increased from 0.00017 (baseline vs. UTOPIAN database) to 0.00289 (active learning vs. UTOPIAN database), and the Shannon diversity index

Table 5. Distributional metrics across sex and age groups for each sampling method. Sampling parity with 95% confidence intervals is reported per group. Jensen-Shannon divergence (JSD), Shannon diversity index (SDI), and chi-square p-values are reported per attribute.

Attribute	Group	Baseline annotated sample				Active learning samples			
		Sampling parity (95% CI)	JSD	SDI	p-value	Sampling parity (95% CI)	JSD	SDI	p-value
age	young adult	1.05 (0.99–1.11)	0.00017	2.00	1.00	0.84 (0.75–0.93)	0.00289	1.97	1.00
	adult	1.02 (0.97–1.08)				1.29 (1.19–1.39)			
	senior	0.98 (0.93–1.03)				0.89 (0.80–0.98)			
	elder	0.96 (0.91–1.00)				0.97 (0.88–1.05)			
sex	female	0.99 (0.96–1.02)	0.00002	0.99	0.99	1.08 (1.04–1.13)	0.00106	0.97	0.93
	male	1.01 (0.98–1.05)				0.90 (0.84–0.96)			

Sampling parity compares the proportion of each subgroup across different datasets to assess representational balance. JSD quantifies distributional shifts (0 = identical, higher = more different). SDI reflects diversity of group representation (higher = more diverse). P-values are from two-sided chi-square tests assessing distributional similarity with exact p-values reported.

declined slightly from 2.00 to 1.97, suggesting a mild reduction in age diversity. Regarding sex, the Jensen-Shannon divergence remained low at 0.00106; however, it still marked an increase from 0.00002 in the baseline sample, indicating a measurable shift despite the overall low divergence. Despite these shifts, no statistically significant differences were detected across comparisons ($p > 0.93$), suggesting that demographic representation remained within acceptable bounds throughout the sampling process.

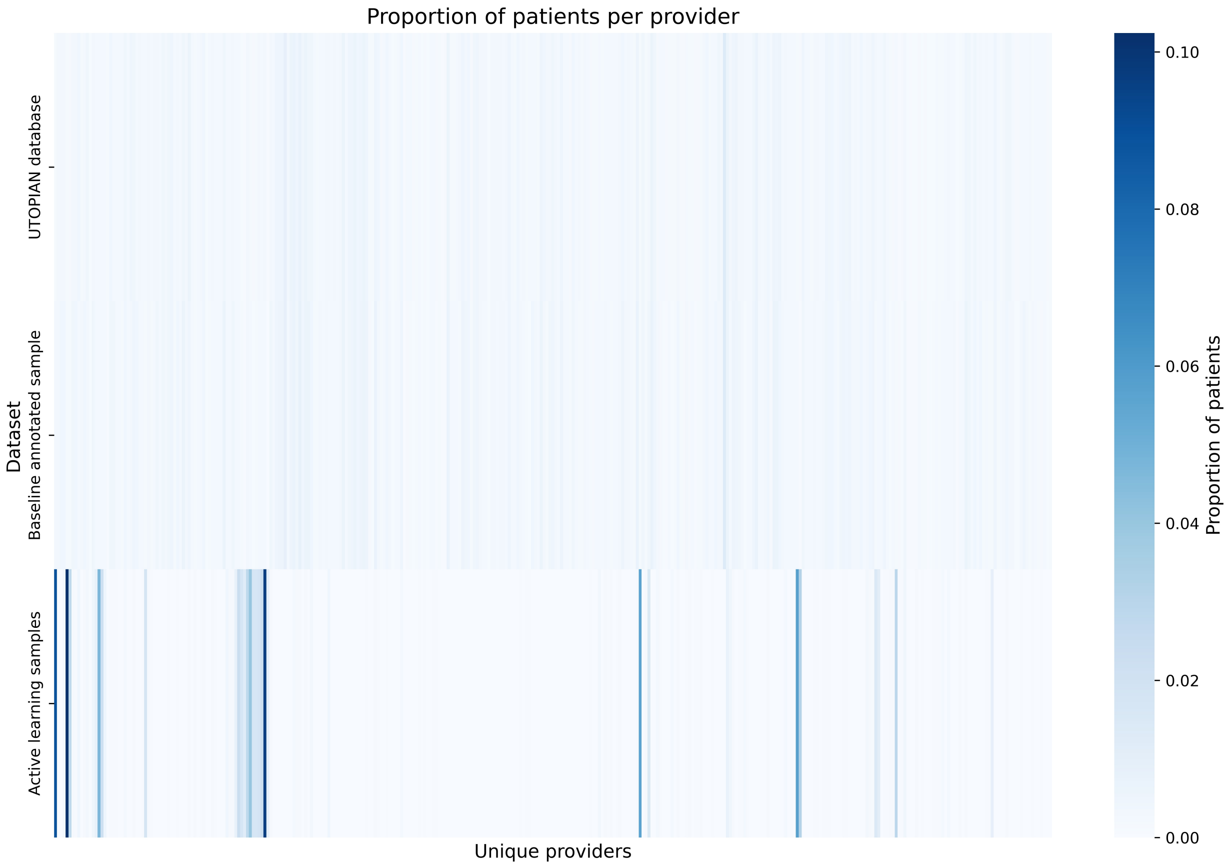
Figure 7 shows the distribution of patients per provider across the full UTOPIAN database, the baseline annotated sample, and the active learning subset. The baseline sample preserved the patient/provider composition of the original database and did not differ significantly from it ($p = 0.08$). In contrast, the active learning subset showed a substantial deviation in patient proportions across providers, where notably, 43 clinics and 209 providers present in the UTOPIAN database were not sampled in the active learning subset. Patient/provider distributions in the active learning data differed significantly from both the baseline annotated sample ($p < 10^{-87}$) and the full UTOPIAN database ($p < 10^{-95}$), indicating a degree of sampling concentration where the active learning strategy introduced a shift in patient coverage that was not present in the original annotation process.

Analysis of model bias across sex and age

All reported results in this section are based on evaluations performed on the held-out 10% of the annotated dataset, conducted using the best-performing model (selected based on cross-validation). Figure 8 compares the actual demographic distribution in the annotated dataset to the distributions of sex and age within the predicted race categories for each model. For clarity and to emphasize meaningful performance patterns, we report the results of the better-performing variant for each model, using the fairness-unaware versions of BioClinicalBERT and the hierarchical CNN. Among all models, RoBERTa exhibited the largest deviations from the true distribution, most notably overrepresenting subgroups such as male East Asians (+2.33 percentage points) and Black young adults (+2.08), which were patterns not as pronounced as in other models. In contrast, BERT and BioClinicalBERT more closely matched the original demographic distribution, though BioClinicalBERT still introduced notable shifts; for example, underrepresenting East Asian young adults (–1.77) and overrepresenting seniors within the same group (+2.19), as well as South Asian seniors (+2.58). Model-specific disparities also emerged: the hierarchical CNN underrepresented white young adults (–2.78) while overrepresenting them within the mixed heritage predictions (+1.85), and DeBERTa overrepresented Southeast Asian adults (+1.84).

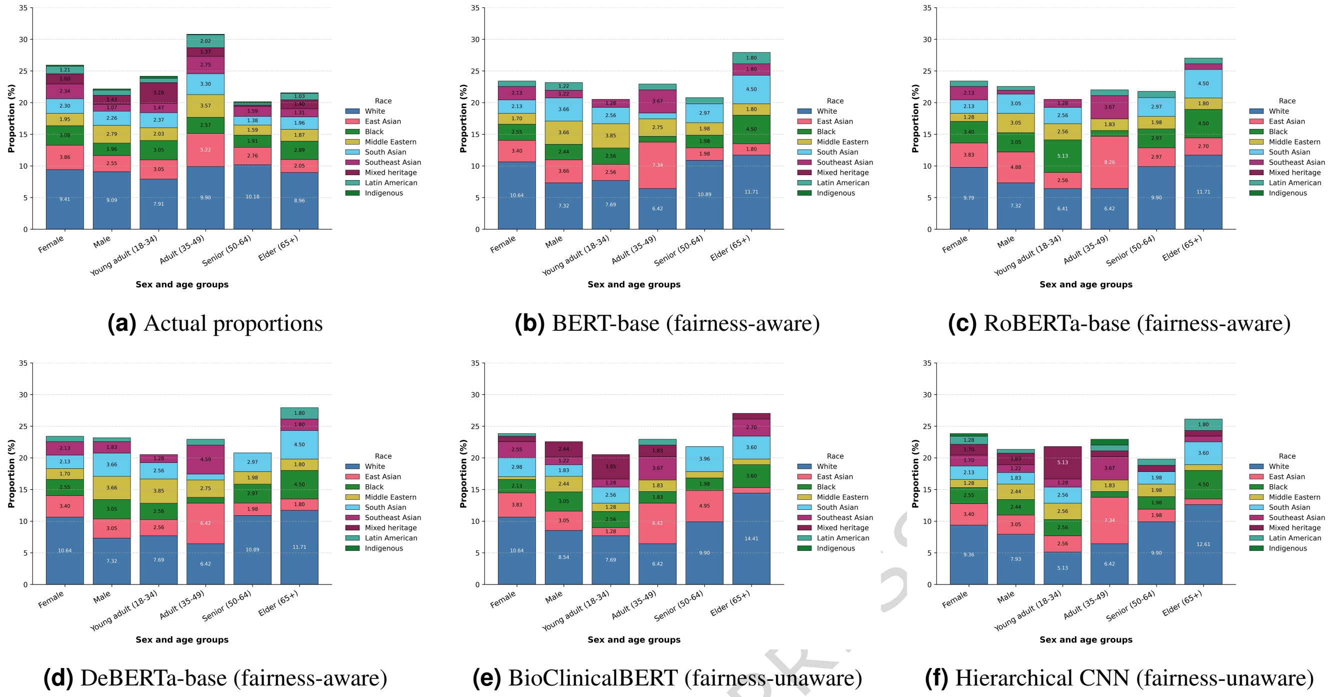
Seniors and adults were consistently absent or severely underrepresented as demographic subgroups among predicted Southeast Asian and South Asian individuals across most models (particularly Southeast Asian seniors and South Asian adults). Similarly, all models lacked any representation of Latin American young adults in their predictions. BioClinicalBERT only included Latin Americans within the female and adult subgroups, while the hierarchical CNN showed a comparable absence of Indigenous individuals. White patients' demographic groups were consistently underrepresented, whereas Black patients' demographic groups tended to be overrepresented across predicted outputs. The most consistent trend across models was the underrepresentation of adults and overrepresentation of elders, suggesting possible age-related linguistic bias in the prediction process. These age-related discrepancies were more pronounced than those associated with sex, where female subgroups exhibited relatively minimal variation across models. All models demonstrated statistically significant disparities ($p < 0.05$) in the distribution of sex and age within predicted race categories, with the exception of white patients, which showed no significant deviation. The most frequent and pronounced disparities occurred for East Asian, Southeast Asian, and, most notably, across all models, Latin American groups.

Figure 9 presents the classification performance of the best-performing version of each model across the sensitive demo-



Metric	Value
Total number of clinics in database	96
Total number of providers in database	408
Missing clinics in active learning samples	43
Missing providers in active learning samples	209
No. patients per provider distribution comparison (chi-square test, p-values)	
Active learning vs. UTOPIAN database	$< 10^{-95}$
Active learning vs. baseline annotated sample	$< 10^{-87}$
Baseline annotated sample vs. UTOPIAN database	0.08

Figure 7. Distribution of patients per provider across datasets. The heatmap shows the proportion of patients per provider in the UTOPIAN database, baseline annotated sample, and active learning samples. The table summarizes database representation (clinics and providers) in the active learning samples and the statistical analysis of the number of patients per provider across datasets using two-sided chi-squared tests, with corresponding exact *p*-values reported.



(a) Actual proportions

(b) BERT-base (fairness-aware)

(c) RoBERTa-base (fairness-aware)

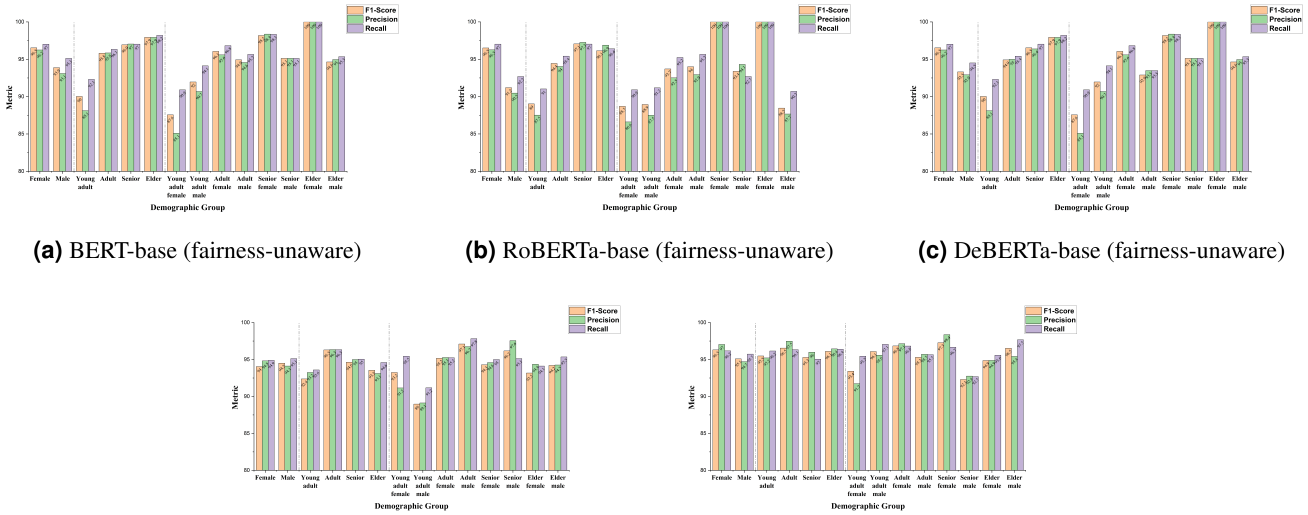
(d) DeBERTa-base (fairness-aware)

(e) BioClinicalBERT (fairness-unaware)

(f) Hierarchical CNN (fairness-unaware)

Figure 8. Comparison between the actual demographic distribution in the dataset and the distribution of predicted race outputs (in percentages). The top-left panel (a) shows the true proportions of sex and age subgroups in the overall annotated dataset. Panels (b) through (f) display the distribution of these subgroups within the model's predicted outputs.

graphic attributes, including their intersection. The most pronounced differences were observed at the intersection of sex and age, rather than when considering either attribute alone. BERT, RoBERTa, and DeBERTa exhibited lower performance for young adults compared to other age groups, a trend that was more evident among young adult females, with F1 scores of 87.60 (BERT), 88.70 (RoBERTa), and 87.60 (DeBERTa). Senior and elder subgroups generally achieved the highest performance across the same models, a trend more pronounced among females. In RoBERTa, for example, male seniors and elders had lower F1 scores (93.40 and 88.50, respectively) compared to a perfect 100.00 for their female counterparts. DeBERTa demonstrated the highest variability across sex-age intersections, while RoBERTa exhibited the highest drop in performance in these groups. The hierarchical CNN remained relatively stable across intersectional groups, with a slight drop for senior males (F1 = 92.30). It also showed the most consistent performance across age groups, and BioClinicalBERT showed the least discrepancies in performance across sex groups. The observed differences in model performances across sex and age groups were not statistically significant ($p > 0.05$).



(a) BERT-base (fairness-unaware)

(b) RoBERTa-base (fairness-unaware)

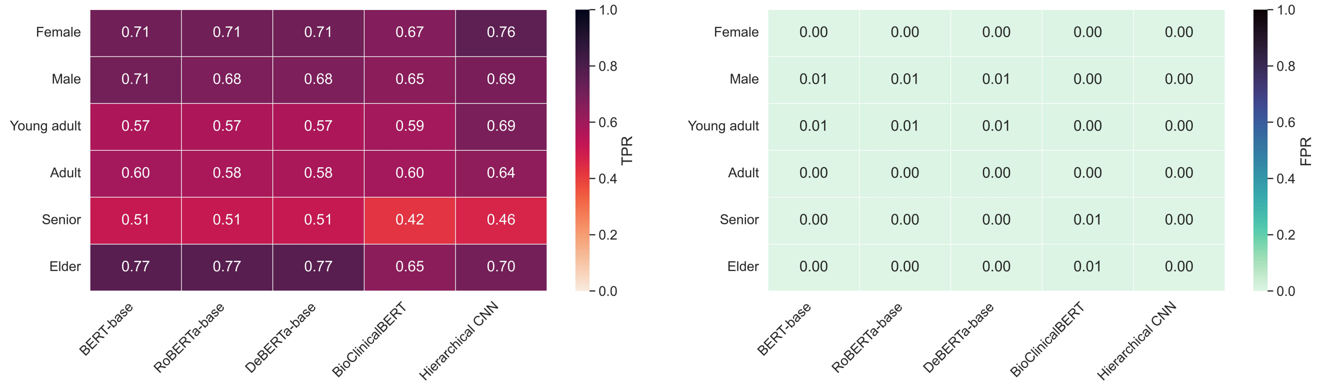
(c) DeBERTa-base (fairness-unaware)

(d) BioClinicalBERT (fairness-unaware)

(e) Hierarchical CNN (fairness-unaware)

Figure 9. Model performance across sensitive attributes. The panels present weighted-averaged F1-score, precision, and recall across age, sex, and the intersections for each model: **(a)** BERT-base, **(b)** RoBERTa-base, **(c)** DeBERTa-base, **(d)** BioClinicalBERT, and **(e)** Hierarchical CNN (all fairness-unaware). The absent class was included in this analysis.

Considering the race classes alone (i.e., excluding the absent class), Figure 10 displays the true and false positive rates across sex and age subgroups for each model, using macro averaging to place equal emphasis on all race categories. Across all models, true positive rates varied more notably by age group than by sex, with a consistent pattern of higher performance for elders and lower performance for seniors. The difference in true positive rates between these two groups ranged from 0.23 to 0.26, highlighting a marked decline in predictive accuracy for the senior subgroup. While transformer-based models exhibited minimal differences in performance between males and females, the hierarchical CNN showed a more pronounced sex-based discrepancy, achieving a higher true positive rate for females (0.76) than for males (0.69). False positive rates remained uniformly low across all models and demographic subgroups, indicating a low rate of over-prediction regardless of age or sex. None of the observed disparities across subgroups reached statistical significance ($p < 0.05$) for any model.



(a) True positive rates

(b) False positive rates

Figure 10. Group-level true and false positive rates across models and sensitive attributes. Heatmaps present the macro-averaged true positive rate (a) and false positive rate (b) for each sensitive attribute (sex and age groups) across the five models. Results reflect the best-performing variant of each model, with fairness-unaware versions shown for BioClinicalBERT and the hierarchical CNN. Colour intensity corresponds to the magnitude of the metric, with darker shades indicating higher values.

Supplementary Figures S2 and S3 provide a more granular view of true and false positive rates disaggregated by race groups. While average true positive rates by sex appear comparable in the group-level heatmap, the supplementary breakdown reveals that the slightly lower overall performance in males was primarily driven by sharp declines in specific race categories, while female performance was at times lower than male performance, even in a broader range of race groups for some models (i.e., in BERT), but the discrepancies were generally more modest. For RoBERTa and DeBERTa, the observed drop in true positive rates among males and adults-relative to BERT-was attributable to poorer performance on East Asian individuals for those groups. BioClinicalBERT and the hierarchical CNN showed more pronounced sex-based differences in true positive rates, as well as greater variability across senior and elder age groups. BioClinicalBERT, in particular, exhibited extensive performance discrepancies by age and consistently underperformed for Black and Middle Eastern individuals across demographic subgroups.

Discussion

Our findings challenge the prevailing assumption that larger, deeper transformer-based models are inherently superior in clinical NLP tasks. Although transformer architectures such as BERT, RoBERTa, DeBERTa, and BioClinicalBERT have rapidly gained traction due to their extensive pretraining and contextual embedding capabilities, our comparative analysis revealed that a thoughtfully designed hierarchical CNN, which explicitly mirrored the structured nature of clinical documentation, consistently achieved higher predictive performance and improved inter-class performance parity, with the added benefit of being more interpretable than the black-box large language model alternatives. The hierarchical CNN consistently outperformed transformer models across macro- and micro-averaged F1 scores, precision, and area under the precision-recall curve, suggesting that capturing intra- and inter-sentence dependencies is especially valuable when identifying sparse, demographically meaningful information such as race from clinical text. This performance gap is particularly noteworthy given the inherent data-scarce and imbalanced nature of real-world EHRs, conditions under which many transformers struggle (as noted by the fairness-unconstrained results of the transformer models). Importantly, the hierarchical CNN's superior macro-F1 score (98.44%) reflects better balance across racial subgroups, reinforcing the need for task-specific architectural choices rather than defaulting to scale. We can find a broader lesson for clinical AI development here: When dealing with the messy nature of clinical data, models explicitly tailored to reflect the clinical texts structure and context may offer greater utility and fairness than state-of-the-art black-box alternatives.

Introducing fairness constraints revealed not a universal performance penalty, but rather model-specific effects that offer crucial insights for clinical deployment. Fairness-aware versions of transformer models, particularly BERT and RoBERTa, demonstrated substantial improvements in both predictive accuracy and equity, elevating performance from negligible levels (mostly zero precision and recall) for non-white groups to significantly higher scores. Even DeBERTa, which experienced a decline in macro-level metrics under fairness constraints, maintained improved micro-averaged scores and better class balance, suggesting that certain performance trade-offs may be acceptable when weighed against downstream patient impact. However, BioClinicalBERT and the hierarchical CNN both experienced performance trade-offs under fairness constraints, with BioClinicalBERT showing a notable decline in average precision and area under the precision-recall curve, and although its macro-averaged precision, recall, and F1 scores improved, performance for all non-white groups collapsed entirely at the per-class level (showing the limitations of relying solely on aggregate metrics when evaluating fairness).

In clinical settings, these trade-offs between fairness and accuracy are not merely mathematical, they carry real-world consequences. Reduced performance for historically marginalized groups can delay diagnoses, result in unfair resource allocations, and ultimately lead to worse health outcomes for these communities[30], even when overall model performance appears strong. From this perspective, fairness constraints are not just algorithmic improvements but ethical imperatives, especially when, as shown in this study, they often improve both equity and performance. From a deployment perspective, our results emphasize the importance of selecting fairness-aware models based on task-specific and population-specific needs. For instance, stakeholders focused on optimizing precision and F1 scores may benefit most from fairness-aware BERT, while those seeking robust classification across imbalanced groups may prefer the fairness-aware hierarchical CNN, given its high average precision and area under the precision-recall curve. However, the performance degradation and increased false negatives observed in the fairness-aware hierarchical CNN shows the necessity of evaluating fairness-aware models holistically before real-world clinical deployment, as model-specific characteristics and pretraining biases may significantly influence how fairness-aware methods affect group-level predictions. We recommend that explicit clinical and ethical criteria guide the adoption of fairness-aware approaches. Models should be selected not only for their ability to maintain clinical accuracy but also for their capacity to reduce disparities. Tailoring interventions based on population demographics, clinical context, and equity considerations is essential for responsible, real-world deployment of clinical AI systems.

Importantly, many disparities in model performance were observed across racial groups. Even with fairness constraints improving overall group-level performance in the transformer models, disparities persisted, pointing to the limitations of these interventions when upstream data sources carry embedded biases. For example, Indigenous patients were consistently underserved, and transformer-based models frequently failed to predict any cases for mixed heritage patients. These failures reflect broader issues of underrepresentation: approximately 84% of the global population, primarily non-Caucasian groups, remain underrepresented in clinical omics data[30,33], which are heavily skewed toward individuals of European ancestry.

Similar imbalances have been observed in many types of clinical data[67–70], including clinical notes, which pose challenges for commonly used machine learning models when applied to data-disadvantaged populations.

Not all observed disparities in model performance could be explained by sample size alone, we also identified notable discrepancies among racial groups with similar representation levels. This suggests that other factors, such as pretraining biases in large language models[71–76], may be at play. For example, active learning further contributed to representational imbalances by producing non-proportional sampling across racial groups. While underlying data imbalance likely influenced these selections, our post-hoc analysis revealed notable deviations from expected sampling rates. These disparities may reflect how uncertainty estimates from the BERT model interact with underrepresented or linguistically diverse patient subgroups[67] (since fairness was not explicitly considered in the sampling design, these findings should be interpreted with caution). Many active learning studies in clinical NLP emphasize annotation efficiency rather than the representational fairness of the sampling model itself[38–40]. Because the query model's uncertainty is often correlated with data density or group frequency, these strategies can unintentionally amplify existing disparities (See Supplementary Table S2). This highlights an important yet understudied source of bias propagation: the fairness of the model guiding data collection. Our findings emphasize that fairness considerations should extend beyond the final predictive model to the entire data pipeline, including active learning, to prevent amplification of representational bias in the labeled dataset.

The fairness-unaware hierarchical CNN, by contrast, exhibited more stable performance across racial groups, with representation ratios closer to 1.00 and reduced disparities in false positives and negatives. That the fairness-unaware CNN could reliably identify race information, whereas the transformer counterparts could not, indicates that such cues are indeed present in the clinical narratives. The discrepancy therefore reflects how architectural and optimization characteristics of transformer models can interact with pretraining biases to suppress minority-group signals, rather than an absence of these signals in the data itself. This suggests that architectural features such as document-structure preservation may help mitigate some pretraining-related biases. However, even this model showed increased false negatives for Indigenous patients, demonstrating that architecture alone cannot fully resolve entrenched disparities.

Pretraining biases of the transformer models can be further amplified during fine-tuning if clinical notes reflect uneven documentation practices. Prior work has shown that physician notes can encode implicit bias, given that physicians too are not immune to the influence of societal biases, shaping how patient characteristics are recorded based on race[68,77]. This may reduce the semantic consistency of race-related expressions and lead to discriminatory outcomes by models trained on this data. Notably, AI systems have been shown to infer race from redacted clinical text and medical images with surprising accuracy, often outperforming human experts[67,70]. This shows how deeply racial signals are embedded in clinical data and raises ethical concerns for the use of "race-agnostic" models[30]. Across models, the most prominent source of error was under-classification, particularly for marginalized groups, where performance remained almost consistently higher for white patients compared to other racial groups. While this may partly reflect sample size imbalances, it also points to deeper causes, such as lexical variation and race-specific documentation norms, that reflect structural inequities in healthcare and are now encoded in clinical NLP models. This illustrates the multi-layered nature of racial bias in medical AI, rooted not only in training data but also in the systems that generate it. While fairness-aware training improved representation balance in many cases, it cannot fully address disparities stemming from linguistic, institutional, or systemic inequities. We therefore advocate for routine bias audits using disaggregated metrics, such as representation ratios and false negative rates, prior to clinical deployment. Understanding why certain groups are consistently misclassified, even when adequately represented, is essential to developing equitable algorithms and ensuring that model outputs align with ethical standards and patient-centered care.

These risks become even more concerning when considering the emergence of performance disparities across sex and age groups, despite the fact that neither attribute was explicitly provided to the models, suggesting that demographic signals may be implicitly encoded in clinical text in ways that intersect with race and influence model behavior. While sex-based disparities were relatively modest, age-related variation was more pronounced and systematic. This bias was even more evident within certain racial groups; South Asian, Southeast Asian, and Latin American patients faced greater representational disparities across sex and age categories than any other group, with many models either underperforming or failing to represent them altogether. For Indigenous patients, the only model capable of classification (hierarchical CNN) did so exclusively for female adults, failing entirely for other sex and age subgroups. Models also varied in predictive accuracy across demographic intersections. For instance, BERT, RoBERTa, and DeBERTa achieved the highest performance for older females, while their lowest performance was observed in younger females. This suggests that documentation for certain demographic groups may contain more explicit or structured race-related language, while other groups may be documented with fewer or noisier race cues, possibly due to incomplete EHRs or inconsistent documentation practices, further complicating model behavior.

Given the intricate nature of biases, woven into both training data and model architectures, ensuring fairness in clinical AI systems requires more than technical solutions; it necessitates structural reforms, improved documentation practices, sociodemographic representation, and a sustained commitment to ethical design. Broader interventions, including interdisciplinary and community collaboration, AI literacy in medical education, and policy reforms, are essential to address the structural roots of sociodemographic inequities in healthcare and to promote equitable access to care. Improving how data is collected, recorded, and ultimately used in downstream AI applications is critical for reducing documentation biases and enhancing the reliability of social determinants of health information. One actionable step is to improve data representativeness through fairness-aware sampling strategies. As demonstrated in this study, sampling biases may emerge from the techniques used, which may directly affect model equity. Future efforts should consider combining fairness interventions with sampling strategies, such as uncertainty-based active learning coupled with stratified sampling. Fairness-aware modeling techniques[78–80], which were shown in this study to reduce racial disparities in prediction outcomes, should become standard practice in model development pipelines. Achieving long-term accountability requires integrating algorithmic fairness audits[81,82] as a routine component of model evaluation and deployment. Group, individual, and counterfactual fairness metrics[43] (such as those used in this study) can help uncover and address the underlying drivers of algorithmic bias. This requires testing across intersecting subpopulations as well, where we observed the greatest performance disparities. Another promising avenue involves leveraging explainable AI frameworks to "unbox" black-box systems, making hidden biases more transparent and actionable[83,84]. When paired with human-in-the-loop workflows, these tools enable clinicians to review, contextualize, and challenge algorithmic decisions, enhancing system-level accountability[85]. Such systems, however, require targeted clinician training focused on AI ethics, bias detection, and model evaluation.

While the collection and use of race data are critical for identifying and addressing healthcare disparities, it is essential to acknowledge that such efforts also carry significant ethical risks. The very act of classifying patients by race—a socially constructed and historically fluid category—may inadvertently reinforce stereotypes, contribute to stigma, or be misappropriated in ways that further marginalize vulnerable populations. There is a growing concern among minority communities regarding the potential misuse of their data, stemming from historical instances of medical exploitation and systemic bias[86]. This mistrust can, in turn, influence clinical encounters, as healthcare providers may anticipate patient discomfort and view this reluctance as a barrier to effectively assessing social determinants of health[87]. In the current socio-political climate, discussions around race and systemic inequities have become contentious. This environment necessitates a careful, transparent approach to research that emphasizes the ethical imperatives of equity and justice without alienating stakeholders. As AI becomes increasingly embedded in healthcare delivery, it is crucial to recognize that these systems reflect the societal contexts in which they are developed[30], and it is our collective responsibility to ensure that AI becomes a force for equity rather than an amplifier of existing disparities. This study does not claim to "solve" these deeply entrenched issues but rather seeks to navigate their complexity by promoting responsible, equity-focused use of data. Ultimately, while this path is fraught with challenges, a conscientious and inclusive approach can harness the power of AI to advance equitable healthcare outcomes.

Institutional-scale deployment of transformer-based architectures poses formidable computational and infrastructural challenges that warrant deliberate consideration. Processing hundreds of millions of clinical notes with these models would demand extensive GPU time—potentially months of computation—and substantial operational costs. For large-scale applications such as populating structured demographic fields across electronic health record systems, more computationally efficient methods, including rule-based or hybrid pipelines, may offer more practical and sustainable alternatives[13,26]. By contrast, the fairness-aware models developed in this study serve a more specialized function: enabling research, bias auditing, and the creation of representative cohorts where equity cannot be compromised. We advocate for a tiered implementation strategy in which efficient screening approaches first identify candidate records, followed by fairness-aware deep learning methods for cases requiring nuanced interpretation or equity-sensitive evaluation. Within this framework, the hierarchical CNN represents an optimal middle ground; combining the interpretability and computational efficiency of conventional architectures with the accuracy and equity benefits of fairness-aware learning. Ultimately, different use cases demand different solutions: lightweight, scalable methods for routine tasks, and fairness-aware, interpretable models for high-stakes research and decision-making contexts where equitable outcomes are paramount.

While this study provides valuable insights into fairness-aware modeling for race classification in EHRs, several limitations should be acknowledged. First, although our fairness-aware models incorporated multiple constraints, we were unable to isolate the individual effects of each intervention. Future work should conduct controlled experiments to evaluate these constraints independently, accounting for model configurations, to better determine which fairness strategies are most effective and for which model types. Second, we were unable to fully explain the performance disparities observed between racial groups with comparable sample sizes, suggesting that other latent factors may be driving these differences. Further research is needed to disentangle these sources of bias using tools such as linguistic audits and subgroup-specific error analysis. Importantly, while transformer-based models like BERT and RoBERTa benefited substantially from fairness constraints, other models, such as the hierarchical CNN, showed degraded performance and increased bias under the same conditions. This points to the critical role of model-specific characteristics and pretraining biases in shaping how fairness interventions impact group-level outcomes. As such, fairness constraints should not be assumed to yield uniformly positive results and must be evaluated on a per-model basis. Future work should further disentangle the respective contributions of representation learning and aggregation strategy by examining intermediate architectures, such as sentence-level transformer models with alternative aggregation schemes, to better understand how architectural choices mediate fairness outcomes in clinical NLP.

Additionally, our automated sampling strategy did not incorporate fairness objectives, which may have contributed to representational imbalances in the final training data. Future iterations should consider fairness-aware sampling methods that prioritize underrepresented groups while maintaining selection diversity. Such approaches may help mitigate, rather than reinforce, disparities over time. Although we examined performance disparities across intersecting demographic groups, these results should be interpreted with caution given the limited sample sizes of several minority subgroups. Intersectional metrics are inherently sensitive to data sparsity and can exhibit high variance under such conditions. Our intent was therefore to highlight potential patterns of inequity rather than to draw inferential conclusions. Nonetheless, reporting these estimates remains important for transparency and for guiding future efforts toward more representative and balanced datasets. Lastly, while this study was conducted using the UTOPIAN database, a diverse dataset spanning 96 primary care clinics using widely adopted EHR systems in Ontario, the generalizability of our findings remains to be tested in other settings, particularly hospital-based EHRs and those from different provinces or healthcare systems. These extensions are essential to assess the broader applicability, robustness, and equity of AI-driven demographic inference tools in real-world clinical environments.

Conclusions

This study demonstrates that equitable clinical NLP requires more than technical sophistication; it requires design choices that mirror the structure, context, and inequities of the data itself. By uniting hierarchical architectures, fairness-aware learning, and active annotation, we show that model performance and equity are not mutually exclusive goals, but complementary dimensions of responsible AI. The hierarchical CNN's superior performance under fairness-unaware training highlights the value of domain-aligned architectures, while the mixed effects of fairness constraints across models show the need for tailored, task-specific interventions rather than one-size-fits-all fairness solutions. Persistent disparities across race, sex, and age, even in fairness-constrained models, reveal that algorithmic inequities often originate upstream, in the documentation and data generation processes themselves. Addressing these inequities demands a paradigm shift: fairness must be treated not as an afterthought, but as an architectural principle guiding every stage of AI development, from data collection to deployment. Technical advances alone cannot rectify the structural biases embedded in health systems, but they can illuminate where reform is most urgently needed. By coupling rigorous quantitative evaluation with ethical and clinical insight, this work offers a path toward NLP systems that do more than predict; they help rebuild trust, representation, and accountability in digital health.

Acknowledgements

This work was supported by the Canadian Institutes of Health Research [grant number 173094]. Dr. K Tu receives a Chair in Family and Community Medicine Research in Primary Care at UHN and a Research Scholar Award from the Department of Family and Community Medicine, Temerty Faculty of Medicine, University of Toronto. Dr. L Celi is funded by the National Institute of Health through DS-I Africa U54 TW012043-01 and Bridge2AI OT2OD032701, and the National Science Foundation through ITEST #2148451.

Author contributions statement

K.T. and E.S. conceived the study. R.A. designed and conducted the study, developed and implemented the models, collected and processed the data, performed model and bias analyses, and drafted the manuscript. K.T. and E.S. supervised the study, provided resources, assisted in manuscript editing and review, and contributed to project administration. K.T. additionally curated data, and secured funding. Y.L. contributed to the conceptualization and development of the hierarchical CNN model and the active learning model. Y.L. also provided input on the methodology and interpretation of results. S.A. developed the active learning model, performed its analysis, and generated results. S.A. also assisted in drafting portions of the manuscript. L.A.C. contributed to the interpretation of findings, assisted in drafting the discussion and future directions, and provided critical feedback on the manuscript. Q.Z. provided support for data analysis and interpretation of results. All authors—R.A., Y.L., S.A., K.T., L.A.C., Q.Z., and E.S.—reviewed and approved the final manuscript.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The data used in this study are individual-level, de-identified electronic health record data. Policies, procedures, and Research Ethics Board (REB) regulations governing the source data prohibit public release of individual-level data; only aggregate data are permitted for disclosure. The nature of the data used in this particular project is such that there is no way to aggregate the data for public release. The dataset was derived from the University of Toronto Practice-based Research Network's (UTOPIAN) Data Safe Haven, a large primary care EHR repository encompassing over 400 clinics and 400,000 patients in Ontario, Canada. The parent database has been archived and is not currently accessible.

Access to the dataset may be considered in the future upon request and approval by the University of Toronto Health Sciences REB. Requests for data access should be directed to the Human Research Ethics Unit at `ethics.review@utoronto.ca` or to the research ethics coordinator, Mariya Gancheva (`m.gancheva@utoronto.ca`). Requests will be reviewed within approximately four weeks and are subject to applicable institutional data use agreements. All data are stored securely on encrypted institutional servers within the University of Toronto Data Safe Haven environment.

All aggregate numerical source data underlying the main and supplementary figures are provided in *Supplementary data 1 (Excel)*, which is sufficient to reproduce the analyses and visualizations presented in this paper. Numerical data underlying Figure 7 (provider-level proportions) are not publicly shared due to potential re-identification risk under UTOPIAN Data Safe Haven REB policy.

References

1. Ford, M. E. & Kelly, P. A. Conceptualizing and categorizing race and ethnicity in health services research. *Heal. Serv. Res.* **40**, 1658–1675 (2005).
2. Prus, S. G. Comparing social determinants of self-rated health across the United States and Canada. *Soc. Sci. Medicine* **73**, 50–59 (2011).
3. Morris, S. M. *et al.* Predictive modeling for clinical features associated with neurofibromatosis type 1. *Neurol. Clin. Pract.* **11**, e497–e505 (2021).
4. Brown, T. H., O'Rand, A. M. & Adkins, D. E. Race–ethnicity and health trajectories: Tests of three hypotheses across multiple groups and health outcomes. *J. Heal. Soc. Behav.* **53**, 359–377 (2012).
5. Lubetkin, E. I., Jia, H., Franks, P. & Gold, M. R. Relationship among sociodemographic factors, clinical conditions, and health-related quality of life: Examining the EQ-5D in the US general population. *Qual. Life Res.* **14**, 2187–2196 (2005).
6. Lingren, T. *et al.* Developing an algorithm to detect early childhood obesity in two tertiary pediatric medical centers. *Appl. Clin. Informatics* **7**, 693–706 (2016).
7. Ahuja, Y. *et al.* Leveraging electronic health records data to predict multiple sclerosis disease activity. *Annals Clin. Transl. Neurol.* **8**, 800–810 (2021).
8. Franks, P., Gold, M. R. & Fiscella, K. Sociodemographics, self-rated health, and mortality in the US. *Soc. Sci. Medicine* **56**, 2505–2514 (2003).
9. Freeman, H. P. The meaning of race in science–considerations for cancer research: Concerns of special populations in the national cancer program. *Cancer: Interdiscip. Int. J. Am. Cancer Soc.* **82**, 219–225 (1998).
10. Davidson, J., Vashisht, R. & Butte, A. J. From genes to geography, from cells to community, from biomolecules to behaviors: The importance of social determinants of health. *Biomolecules* **12**, 1449 (2022).
11. Bucher, B. T. *et al.* Determination of marital status of patients from structured and unstructured electronic healthcare data. In *AMIA Annu. Symp. Proc.*, vol. 2019, 267–274 (2019).
12. Han, S. *et al.* Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. *J. Biomed. Informatics* **127**, 103984 (2022).
13. Sholle, E. T. *et al.* Underserved populations with missing race ethnicity data differ significantly from those with structured race/ethnicity documentation. *J. Am. Med. Informatics Assoc.* **26**, 722–729 (2019).
14. Polubriaginof, F. C. *et al.* Challenges with quality of race and ethnicity data in observational databases. *J. Am. Med. Informatics Assoc.* **26**, 730–736 (2019).
15. Proumen, R., Connolly, H., Debick, N. A. & Hopkins, R. Assessing the accuracy of electronic health record gender identity and REal data at an academic medical center. *BMC Heal. Serv. Res.* **23**, 884 (2023).
16. Qing, L., Linhong, W. & Xuehai, D. A novel neural network-based method for medical text classification. *Futur. Internet* **11**, 255 (2019).
17. Nguyen, H. & Patrick, J. Text mining in clinical domain: Dealing with noise. In *Proceedings of the 22nd Association for Computing Machinery Special Interest Group on Knowledge Discovery in Data International Conference on Knowledge Discovery and Data Mining*, 549–558 (2016).

18. Abulibdeh, R. *et al.* Assessing the capture of sociodemographic information in electronic medical records to inform clinical decision making. *PloS One* **20**, e0317599 (2025).

19. Senior, M. *et al.* Identifying predictors of suicide in severe mental illness: A feasibility study of a clinical prediction rule (oxford mental illness and suicide tool or OxMIS). *Front. Psychiatry* **11**, 268 (2020).

20. Lybarger, K. *et al.* Leveraging natural language processing to augment structured social determinants of health data in the electronic health record. *J. Am. Med. Informatics Assoc.* **30**, 1389–1397 (2023).

21. Patra, B. G. *et al.* Extracting social determinants of health from electronic health records using natural language processing: A systematic review. *J. Am. Med. Informatics Assoc.* **28**, 2716–2727 (2021).

22. Bompelli, A. *et al.* Social and behavioral determinants of health in the era of artificial intelligence with electronic health records: A scoping review. *Heal. Data Sci.* **2021** (2021).

23. Zhang, D., Thadajarassiri, J., Sen, C. & Rundensteiner, E. Time-aware transformer-based network for clinical notes series prediction. In *Machine learning for Healthcare Conference*, 566–588 (PMLR, 2020).

24. Yang, Z. *et al.* Hierarchical attention networks for document classification. In *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489 (2016).

25. Abulibdeh, R., Tu, K. & Sejdić, E. Natural language processing methods for assessing social determinants of health in the electronic health records: A narrative review. *Expert. Syst. with Appl.* 127928 (2025).

26. Shi, J. *et al.* Accelerating clinical NLP at scale with a hybrid framework with reduced GPU demands: A case study in dementia identification. *arXiv preprint arXiv:2504.12494* (2025).

27. Flaxman, A. D. & Vos, T. Machine learning in population health: Opportunities and threats. *Public Libr. Sci. Medicine* **15**, e1002702 (2018).

28. Weissler, E. H. *et al.* The role of machine learning in clinical research: Transforming the future of evidence generation. *BioMed Cent. Trials* **22**, 1–15 (2021).

29. Habehh, H. & Gohel, S. Machine learning in healthcare. *Curr. Genomics* **22**, 291 (2021).

30. Haider, S. A. *et al.* The algorithmic divide: A systematic review on AI-driven racial disparities in healthcare. *J. Racial Ethn. Heal. Disparities* 1–30 (2024).

31. Yu, Z. *et al.* Iguevara2024largedentifying social determinants of health from clinical narratives: A study of performance, documentation ratio, and potential bias. *J. Biomed. Informatics* **153**, 104642 (2024).

32. Guevara, M. *et al.* Large language models to identify social determinants of health in electronic health records. *NPJ Digit. Medicine* **7**, 6 (2024).

33. Gao, Y., Sharma, T. & Cui, Y. Addressing the challenge of biomedical data inequality: An artificial intelligence perspective. *Annu. Rev. Biomed. Data Sci.* **6**, 153–171 (2023).

34. University of Toronto family medicine report. Tech. Rep., Department of Family and Community Medicine at the University of Toronto, Toronto, ON, Canada (2019).

35. OntarioMD. Provincial EMR-integrated access (2025).

36. OntarioMD. From foundation to integration: Annual report 2016-2017 (2017).

37. Canadian Insitute for Health Information. Guidance on the use of standards for race-based and indigenous identity data collection and health reporting in canadas (2022).

38. Lybarger, K., Ostendorf, M. & Yetisgen, M. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. *J. Biomed. Informatics* **113**, 103631 (2021).

39. Figueroa, R. L., Zeng-Treitler, Q., Ngo, L. H., Goryachev, S. & Wiechmann, E. P. Active learning for clinical text classification: Is it better than random sampling? *J. Am. Med. Informatics Assoc.* **19**, 809–816 (2012).

40. Chen, Y., Lasko, T. A., Mei, Q., Denny, J. C. & Xu, H. A study of active learning methods for named entity recognition in clinical text. *J. Biomed. Informatics* **58**, 11–18 (2015).

41. Yang, Z., Dehmer, M., Yli-Harja, O. & Emmert-Streib, F. Combining deep learning with token selection for patient phenotyping from electronic health records. *Sci. Reports* **10**, 1432 (2020).

42. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A. & Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization. *J. Mach. Learn. Res.* **18**, 1–52 (2018).

43. Caton, S. & Haas, C. Fairness in machine learning: A survey. *Assoc. for Comput. Mach. Comput. Surv.* **56**, 1–38 (2024).

44. Han, J., Kamber, M. & Pei, J. *Data Mining: Concepts and Techniques* (Morgan Kaufmann, Boston, 2012), 3rd edn.

45. Hardt, M., Price, E. & Srebro, N. Equality of opportunity in supervised learning. *Adv. Neural Inf. Process. Syst.* **29** (2016).

46. Khalili, M. M., Zhang, X. & Abroshan, M. Loss balancing for fair supervised learning. In *International Conference on Machine Learning*, 16271–16290 (PMLR, 2023).

47. Lai, Y. & Guan, L. Flexible fairness-aware learning via inverse conditional permutation. *arXiv preprint arXiv:2404.05678* (2024).

48. Liu, M. *et al.* FAIM: Fairness-aware interpretable modeling for trustworthy machine learning in healthcare. *Patterns* **5** (2024).

49. Lee, G. & Sayer, S. Exploring equality: An investigation into custom loss functions for fairness definitions. *arXiv preprint arXiv:2501.01889* (2025).

50. Stemerman, R. *et al.* Identification of social determinants of health using multi-label classification of electronic health record clinical notes. *J. Am. Med. Informatics Assoc. Open* **4**, ooaa069 (2021).

51. Grandini, M., Bagli, E. & Visani, G. Metrics for multi-class classification: An overview. *arXiv preprint arXiv:2008.05756* (2020).

52. Shaphiro, S. & Wilk, M. An analysis of variance test for normality. *Biometrika* **52**, 591–611 (1965).

53. Scheffe, H. *The analysis of variance*, vol. 72 (John Wiley & Sons, 1999).

54. Tukey, J. W. Comparing individual means in the analysis of variance. *Biometrics* 99–114 (1949).

55. Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* **32**, 675–701 (1937).

56. Nemenyi, P. B. *Distribution-free multiple comparisons.* (Princeton University, 1963).

57. Wilcoxon, F. *Individual comparisons by ranking methods* (Springer, New York, NY, USA, 1992).

58. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *The Annals Math. Stat.* 50–60 (1947).

59. Kruskal, W. H. & Wallis, W. A. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* **47**, 583–621 (1952).

60. Pearson, K. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, Dublin Philos. Mag. J. Sci.* **50**, 157–175 (1900).

61. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32** (2019).

62. Wolf, T. *et al.* Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2020).

63. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. Assoc. Comput. Linguistics*, 4171–4186 (2019).

64. Liu, Y. *et al.* RoBERTa: A robustly optimized bert pretraining approach. *Clin. Orthop. Relat. Res.* (2019).

65. He, P., Liu, X., Gao, J. & Chen, W. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations* (2021).

66. Alsentzer, E. *et al.* Publicly available clinical BERT embeddings. presented at the Proceedings of the 2nd clinical natural language processing workshop (2019).

67. Gichoya, J. W. *et al.* AI recognition of patient race in medical imaging: A modelling study. *The Lancet Digit. Heal.* **4**, e406–e414 (2022).

68. Sun, M., Oliwa, T., Peek, M. E. & Tung, E. L. Negative patient descriptors: Documenting racial bias in the electronic health record. *Heal. Aff.* **41**, 203–211 (2022).

69. Wen, D. *et al.* Characteristics of publicly available skin cancer image datasets: A systematic review. *The Lancet Digit. Heal.* **4**, e64–e74 (2022).

70. Adam, H. *et al.* Write it like you see it: Detectable differences in clinical notes by race lead to differential model recommendations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 7–21 (2022).

71. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623 (2021).

72. Webster, K. *et al.* Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032* (2020).

73. Kaneko, M. & Bollegala, D. Unmasking the mask–evaluating social biases in masked language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 11954–11962 (2022).

74. Gallifant, J. *et al.* Peer review of GPT-4 technical report and systems card. *PLOS Digit. Heal.* **3**, e0000417 (2024).

75. Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V. & Daneshjou, R. Large language models propagate race-based medicine. *NPJ Digit. Medicine* **6**, 195 (2023).

76. Zack, T. *et al.* Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: A model evaluation study. *The Lancet Digit. Heal.* **6**, e12–e22 (2024).

77. Labban, M. *et al.* Disparities in travel-related barriers to accessing health care from the 2017 national household travel survey. *J. Am. Med. Assoc. Netw. Open* **6**, e2325291–e2325291 (2023).

78. Yang, J., Soltan, A. A., Eyre, D. W., Yang, Y. & Clifton, D. A. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ Digit. Medicine* **6**, 55 (2023).

79. Tsai, T. C. *et al.* Algorithmic fairness in pandemic forecasting: Lessons from COVID-19. *NPJ Digit. Medicine* **5**, 59 (2022).

80. Dunkelau, J. & Leuschel, M. Fairness-aware machine learning. *An Extensive Overv.* 1–60 (2019).

81. van de Sande, D., van Bommel, J., Fung Fen Chung, E., Gommers, D. & van Genderen, M. E. Algorithmic fairness audits in intensive care medicine: Artificial intelligence for all? *Critical Care* **26**, 315 (2022).

82. Liu, X. *et al.* The medical algorithmic audit. *The Lancet Digit. Heal.* **4**, e384–e397 (2022).

83. Hassija, V. *et al.* Interpreting black-box models: A review on explainable artificial intelligence. *Cogn. Comput.* **16**, 45–74 (2024).

84. Nizam, T. & Zafar, S. Explainable artificial intelligence (XAI): Conception, visualization and assessment approaches towards amenable XAI. In *Explainable Edge AI: A Futuristic Computing Perspective*, 35–51 (Springer, 2022).

85. Ghai, B. & Mueller, K. D-bias: A causality-based human-in-the-loop system for tackling algorithmic bias. *IEEE Transactions on Vis. Comput. Graph.* **29**, 473–482 (2022).

86. Albert, S. M. *et al.* Do patients want clinicians to ask about social needs and include this information in their medical record? *BMC Heal. Serv. Res.* **22**, 1275 (2022).

87. Yelton, B. *et al.* Assessment and documentation of social determinants of health among health care providers: Qualitative study. *J. Med. Internet Res. Form. Res.* **7**, e47461 (2023).