MMHEALTHFAIR: EVALUATING FAIRNESS IN MULTIMODAL LEARNING APPROACHES FOR RISK PREDICTION (FEB - JUN 2025)

Konstantin Georgiev
University of Edinburgh
Centre for Cardiovascular Science
K.S.Georgiev@sms.ed.ac.uk

Jonathan Hope
NHS England
Transformation Directorate
jonathan.hope1@nhs.net

Jonathan Pearson
NHS England
Transformation Directorate
jonathanpearson@nhs.net

ABSTRACT

Innovations in digital health technologies continue to drive novel applications of Multimodal Artificial Intelligence (MMAI). Despite their superior discrimination ability over unimodal algorithms, MMAI approaches risk inheriting and amplifying hidden biases within routine healthcare data. Increasing understanding of fairness in model decisions is a crucial part of ensuring the equitable and safe use of AI-driven risk assessments in clinical practice. The **MM-HealthFair** framework provides a flexible end-to-end pipeline applicable to risk prediction models within the healthcare domain. It enables the detailed investigation of bias patterns induced by routine healthcare data, with applicability for quantifying and measuring bias in underrepresented and undermeasured individuals. It leverages the MIMIC-IV open dataset, allowing investigation of intermediate fusion models across tabular, time-series and text data.

In our fairness analysis, we used a statistically-grounded approach for quantifying fairness metrics through Bias-corrected and accelerated bootstrapping. We additionally provided a simple approach to adjust the multimodal algorithm for biases within sensitive groups using **deep adversarial mitigation**. Finally, we showcase the applicability of **SHAP** values for post-model examination of multimodal feature importance. This allowed us to compute the relative multimodal degrees of dependence in individual-level explanations, highlighting the effects of adversarial mitigation on the multimodal decision boundaries. This methodology is suitable for examining healthcare disparities and detecting patterns of attribution bias within sensitive groups. The output of this work aims to promote the dissemination of knowledge regarding fairness in MMAI algorithms, working towards ensuring transparent and equitable AI decisions. The **MM-HealthFair** framework is openly available on GitHub.

Contents

1 Introduction

Integration of MMAI into healthcare prognostics promises improved detection of disease progression and high-precision forecasting of treatment requirements. In modern healthcare research, these algorithms analyse and combine vast datasets incorporating health records, imaging, genomics, free-text, sensor data among many others [1]. The adaptability of Machine Learning (ML) and Deep Learning (DL) techniques to handle these data modalities opens a range of opportunities to apply patient data for predictive analytics, tailoring treatments and decision support [2].

Despite their increasing potential for revolutionising healthcare delivery, many of these algorithms suffer from biases driven by the learned data representations. In particular, attribution biases are driven by the lack of representation of minority groups in the data [3]. MMAI applications rely on granular indicators of patient health [4], a process that may be substantially limited for data collection in ethnic or social minorities or groups with limited access to health services (Figure 1). If not properly accounted for, these biases could promote dangerous and flawed judgments on these individuals, imposing a serious public health concern towards health equity. Moreover, reliance on general performance measures such as accuracy or Area-under-the-curve may flatter the model's clinical utility and conceal this behaviour [5]. To support responsible and safe use of MMAI algorithms, novel methodologies are required to understand, quantify and adjust for fairness in a multimodal setting.

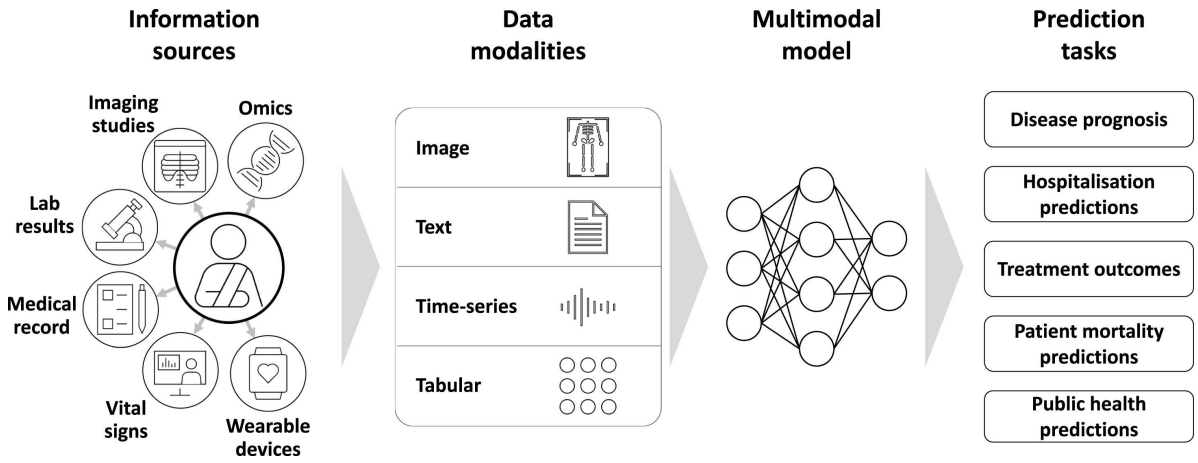

Figure 1: Broad workflow and use cases of MMAI in healthcare. Reproduced with permission from Krones et al. [4].

Understanding model fairness is essential for ensuring transparent and tractable AI decisions. On top of diverse data types, MMAI algorithms leverage hybrid architectures such as concepts from **Convolutional Neural Networks (CNNs)** for image inputs and **Transformer** components for text [6]. The complexity of multimodal fusion across these hybrid architectures poses a challenge for interpretability. Model-agnostic architectures, such as **SHAP** provide a way to examine reliability through measuring feature impact with post-hoc attribution scores [7]. SHAP provides bidirectional explanations for individual predictions, revealing which features are most influential in driving the model decision. While SHAP algorithms are sensitive to context in multimodal scenarios, they can aid in identifying incorrect or biased assumptions at the individual-level for patients assigned as high or low-risk. Other novel **Explainable AI (XAI)** techniques, such as **cross-modal attention**, can map correlations between different modalities [8]. However, they are less suitable for measuring healthcare disparity and studying what affects opportunities for healthcare. For example, understanding the impact of a patient belonging to an ethnic group or being under government-funded insurance on the prediction might expose limited opportunities for healthcare in these individuals.

The **MM-HealthFair** framework aimed to address the research gap in examining fairness within MMAI algorithms for predictive analytics. The primary objective was to develop a reproducible, scalable, and flexible toolkit for interpreting MMAI-driven risk assessments, model decisions, and relationships with attribution bias. This was tied to answering the following research questions:

1. What is the overall impact of multimodal fusion on healthcare disparity compared to unimodal approaches?
2. How does the interplay/inclusion of data modalities affect fairness in the model decisions?
3. What strategies can we leverage to impose fairness constraints on **MMAI** algorithms?

The framework currently supports multimodal evaluation in urgent care, leveraging secondary care data from the **MIMIC-IV** open dataset [9]. It enables the selection and fusion of multimodal components across three modalities:

tabular (EHR) count data, time-series (vital signs, lab tests), and free-text (discharge notes) data. These data are used to quantify fairness in prognostic outcomes, including prediction of **in-hospital death**, **extended stay**, **admission to ICU** and **non-home discharge**. This work is an extension of the baseline multimodal framework provided by Martin et al. [10], supporting multimodal learning objectives for time-series and tabular data: `https://github.com/nhsengland/mm-healthfair/tree/phase1`.

2 Background

This section will serve as a broad introduction to the main concepts in **MMAI**, **fairness** and **explainability** applicable to the conducted experiments using the **MM-HealthFair** framework. This will also cover some of the more prominent clinical applications of these mechanisms in the context of predictive analytics for healthcare.

2.1 Overview of Multimodal AI

2.1.1 MMAI Architectures

The flexibility of **MMAI** networks lies in their transformative potential for integrating multiple data sources to learn rich contextual representations. These algorithms provide a more comprehensive and granular view of patient health markers and their interactions. With the increasing adoption of multi-purpose Large Language Models (LLMs) for clinical tasks, the landscape of MMAI has gradually shifted towards **tokenization-based** algorithms [11] Most foundation models utilising tokenization are based on **Transformer** architectures [12]. The primary building blocks of Transformer architectures that allow development of multimodal representations consist of:

1. **Self-attention mechanism (SA)**: calculates the relevance between different positions in the input sequence to generate weighted representations. The goal of SA is to capture the interdependencies between different positions in the input sequence.
2. **Multi-head self-attention module (MHSA)**: each attention layer is computed in parallel and concatenated using a linear projection matrix, capturing mixed representations from the feature space (allows multimodal learning, through comparing similarities between representations across matrices, known as **cross-attention**).
3. **Feed-Forward Neural Network (FFN)**: adds non-linearity to the learned representations fed by the **MHSA** layers and produces the model weights.

According to a recent review by Wadekar and colleagues [11], **Transformers** can be categorised into four distinct structures:

1. **Type-A**: use of **cross-attention** between layers to integrate text and visual features.
2. **Type-B**: use of custom **cross-attention** projections to reduce computational overhead.
3. **Type-C**: use of modality-specific encoders (e.g. CNNs for images) to extract features, which are then concatenated or pooled for downstream tasks.
4. **Type-D**: all modalities use token representations (e.g. images as patches, text as language embeddings) into a unified input sequence processed by a single Transformer.

In **Type-A** and **Type-B** models, the inputs are integrated within the internal layers of the model, which leads to more fine-grained control over the flow of data representations across modalities (Figure 2). In practice, however, the decision boundaries of these models are often non-trivial to trace and interpret. On the other hand, **Type-C** and **Type-D** architectures leverage a large number of training parameters to create a unified multimodal network with combined embeddings across all data modalities. This setup can provide a tractable and reproducible way of capturing multimodal representations with rich contextual information. However, due to the fact that the modalities are fused at the input layer, this process can be computationally demanding.

2.1.2 Fusion mechanisms

Model design choices in MMAI networks largely depend on the strategy for fusing the input modalities. As suggested by a review from Huang and colleagues [13], the MMAI taxonomy typically consists of three types of multimodal fusion that can have variable impact on the captured data representations: **early**, **intermediate** (or joint) and **late** fusion. In a more recent survey by Li and Tang, these are also referred to as **data-level**, **feature-level** and **model-level** fusion [14], respectively matching the same concepts. While these boundaries are still not clearly defined in MMAI research, this report will use the terminologies defined by Huang et al. as they reflect the depth of learned contextual information across modalities. As categorised by the authors, there are four generalist multimodal setups, applicable to any **ML** or **DL** components (Figure 3). To simplify the distinction between components, they have been labelled as:

- **MM-A** (early fusion);
- **MM-B1** (intermediate fusion; single model fused with pre-trained embeddings);
- **MM-B2** (intermediate fusion, fusing multiple models);

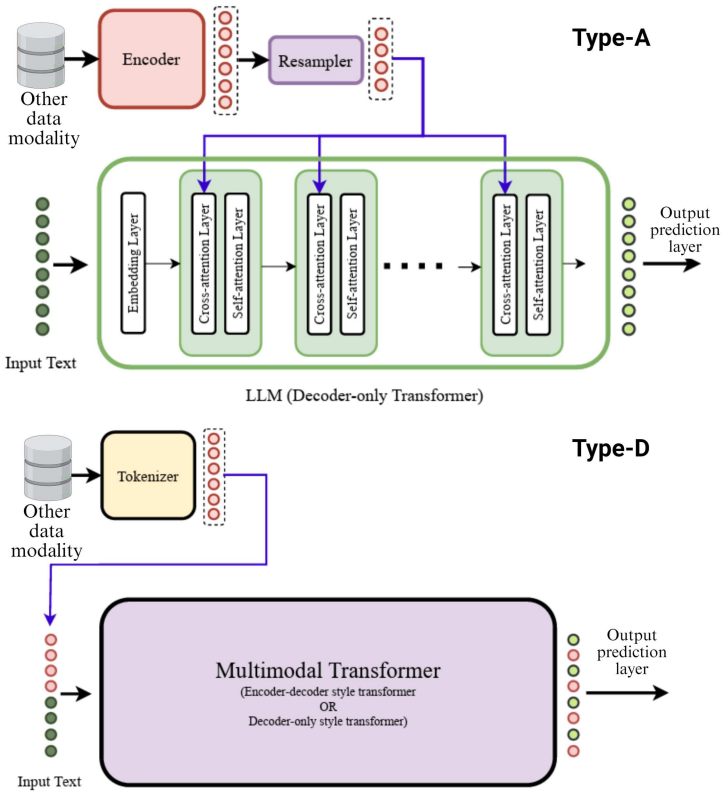

Figure 2: Comparison between **Type-A** and **Type-D** MMAI architectures leveraging Transformers. Modified with permission from Wadekar et al. [11].

- **MM-C** (late fusion);

MM-A type setups (equivalent to **Type-D Transformers**) typically include a form of feature-level integration within the input layer. This means that low-level features (e.g., word embeddings from text and individual time-series measures) are merged into a unified representation. This is typically done through concatenating feature embeddings, pooling or application of gated units [15], [16]. The advantage of this strategy lies in the ability to capture cross-modal correlations early, which is beneficial for tasks like visual question answering (VQA) or sentiment analysis [14]. It additionally reduces redundancy by processing the entire fused feature set once to generate and train the model weights. This setup is particularly effective when modalities are temporally aligned and their interactions are critical for task performance. On the other hand, it can have limited scalability for highly heterogeneous data and increased computational cost from processing high-dimensional inputs. Linear combinations of features may also struggle to capture more complex interdependencies between modalities. The previous iteration of this project adopted two **MM-A**-type mechanisms using concatenation and multi-adaptation gates [10].

Meanwhile, **MM-B1** and **MM-B2** strategies integrate data from different modalities during the model's processing pipeline, balancing the strengths of early (data-level) and late (decision-level) fusion. This relies on **DL** components specifically, due to their ability to backpropagate the output of the loss function to adjust the feature extraction layer [13]. This allows multiple networks to be trained and adjusted simultaneously (**MM-B2**), or these updates can be applied to a single network (**MM-B1**) and the second modality can be inserted within a hidden layer (e.g. cross-attention layers in Type-A Transformers) via pre-trained data embeddings. In both cases, this involves mapping modalities to a shared embedding space. In Transformer architectures, this is done via cross-modal attention layers to weigh and realign segments from these representations [8]. The advantages of intermediate fusion strategies are in their ability to tolerate missing or noisy modalities better than early fusion [14]. They additionally have the flexibility of achieving deeper contextual learning (e.g., using MRI, clinical and genomic data for disease progression detection) [17].

Finally, **MM-C** type strategies leverage isolated trainable model components representing each data modality. In classification tasks, the output layer is aggregated across models to produce a unified probability set. The choice of the aggregation is usually empirical [13], and it can vary depending on the context. Common aggregation methods include:

- Average pooling [18];
- Voting schemes (e.g. majority votes with distinct thresholds);
- Meta-learners (final classifier learning from the outputs of each individual model to produce a single prediction layer [19]);

Late fusion is the more flexible approach, as each modality can be modelled and optimised independently to fit the clinical context. This modularity makes it easier to add or remove components without retraining the entire multimodal network [13]. A late fusion framework also allows training with missing modalities, which can lead to transferrable approaches across health settings with highly heterogeneous data. One limitation, however, is that since modalities are processed independently, this strategy does not fully exploit inter-correlations between modalities, potentially missing important clinical context [20]. For some health settings, this may be overly simplistic if the task requires deep contextual understanding.

Each type of fusion can be suitably applicable for a risk prediction task, depending on the available routine data and clinical context. **MM-A** type models are typically suitable when modalities are tightly coupled (e.g., image + text captions, audio-visual data) [21]. These tasks typically require fine-grained cross-modal interactions. The challenge here is that of memory and computational cost due to the use of raw feature representations. Early fusion over multiple modalities can also be susceptible to the curse of dimensionality, especially with smaller samples. These methods can be highly sensitive to noise in any modality as noise can propagate through the fused representation. Furthermore, the joint feature space is complex and hard to disentangle and it typically requires all modalities to be present and temporally aligned [22].

MM-B1 and **MM-B2** models are useful when modalities have complementary, but not strictly dependent contextual information. Example include tasks like multimodal sentiment analysis and medical diagnostics (e.g. EHR + imaging data [13]). Computational cost is comparably lower than early fusion as they can leverage pre-trained unimodal variants. This scales better for approaches that handle moderate to large datasets. Each modality can also be denoised before fusion during the data preprocessing, and noise can also be controlled within the training objective. It arguably has better interpretability than early fusion, as these setups enable methods such as **SHAP** for modality-specific contribution analysis.

Finally, **MM-C** type strategies are preferred when modalities are largely independent from each other, useful for decision-level tasks such as ensemble predictions under the presence of missing data. In terms of parallel processing, they are the most efficient architecture compared to other fusion types, with low compute for the fusion itself. Despite this, they may require multiple full models, which may not be readily available from pre-trained algorithms. They are typically most robust in handling noise, as representations in one modality do not affect the others. They can also ignore missing or noisy modalities at the decision stage. Interpretability is high as the decision logic is transparent (e.g., through simple majority voting, averaging). However, it faces a major drawback, which is the very limited potential for cross-correlation analysis between data modalities [23].

2.2 Algorithmic Fairness and Bias

The concept of **fairness** in AI for healthcare relates to ethical principles, in areas where model decisions can directly impact patient health, access to services and outcomes. Unfair AI predictions generally refer to those that can perpetuate or even amplify existing health disparities, leading to misdiagnosis and missed opportunities to support underrepresented individuals [24]. In practice, this is not just about equal treatment (equality) but about **equitable** decisions, considering the different needs and contexts of patient subgroups [25]. Thus, to yield public benefits, AI algorithms should adjust for differences that are clinically justified rather than enforcing uniformity that could be inappropriate in medical contexts. For example, balancing discrimination for ethnicity in an algorithm using pulse oximeter data to predict respiratory disease progression could ignore the influence of skin tone possibly affecting light absorption on such devices.

In the healthcare domain, the taxonomy of fairness is still not clearly defined. Most of the literature typically describes these biases either on a **group-level** or **individual-level** (Figure 4). Group fairness ensures that outcomes are balanced across demographic groups (e.g. demographic parity), while individual fairness focuses on treating similar individuals similarly (e.g. counterfactual mechanisms in AI) [26]. Both perspectives are important, but individual fairness is often under-addressed in current healthcare AI research [27]. The difficulty there lies in determining what makes individuals "similar", often requiring expert input and careful selection of clinical features relevant to the task. Optimising for individual fairness can also conflict with group fairness objectives, and vice versa, requiring careful trade-offs to support informed decision-making.

While healthcare research in MMAI aims to improve prognostication across diverse populations, it also introduces novel challenges for understanding impact of fairness across multiple modalities. This is because MMAI algorithms

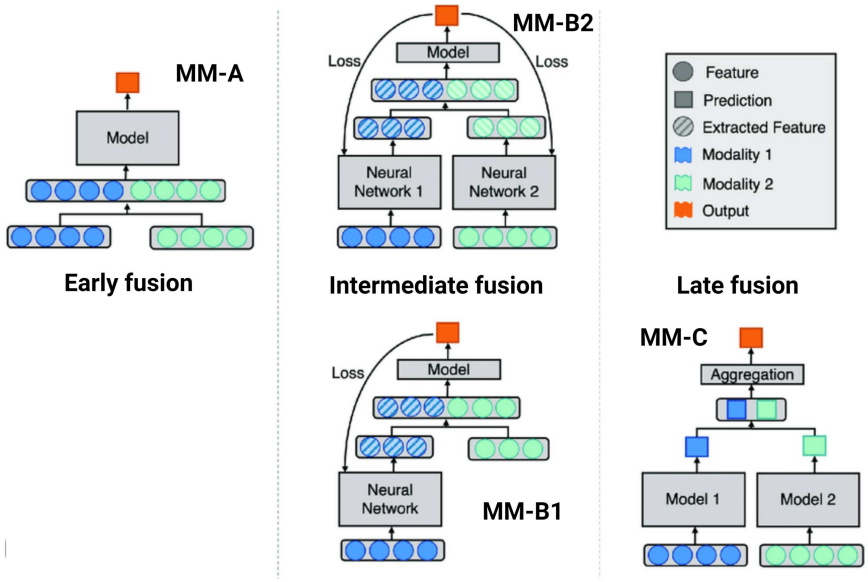

Figure 3: Multimodal fusion types categorised using the taxonomy provided by Huang et al. [13]. **MM-A** describes an **early fusion** strategy using concatenation at the input layer. **MM-B1** (single network fused with pre-trained embeddings) and **MM-B2** (fusion across two networks) describe fusion at an intermediate model layer, with a feedback loop to the original weights. **MM-C** strategies simply aggregate the output layers of separately trained networks. Figure modified with permission from the authors.

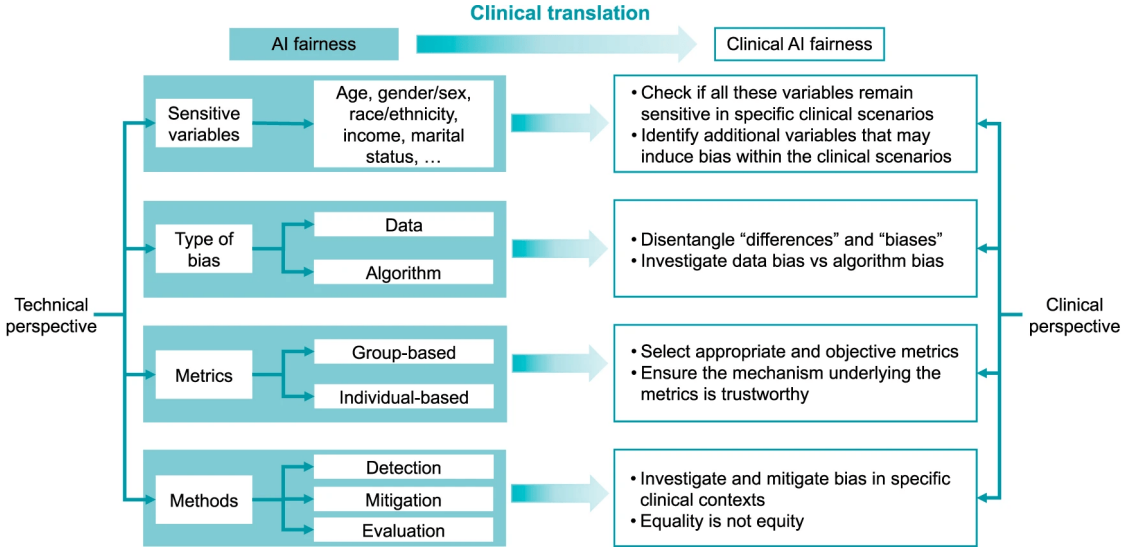

Figure 4: Taxonomy of fairness concepts in the context of clinical translation. Reproduced with permission from Liu et al. [25].

risk amplifying biases present in any single data modality or introducing new biases when integrating data sources with different heterogeneity, availability, or demographic representation [28]. For example, detailed imaging (e.g. mammograms) data may be more readily available for well-resourced, urban hospitals, serving populations with better access to healthcare. This may initiate potential harm in marginalised groups.

To help understand and mitigate attribution biases within AI-driven decisions, open-source projects, such as Fairlearn [29] and OxonFair [30] were developed as reproducible toolkits for bias-related analytics. These tools perform disaggregated evaluation by looking at specific groups that could be disproportionately impacted by an AI decision. This includes measures, such as demographic parity, equalised odds and equal opportunity, defined in Section 3.2.1. They support additional tools for mitigating attribution bias via enforcing fairness constraints which guarantee non-

discriminatory behaviour against a set group of individuals [31]. However, the problem of linking unfair decisions back to the raw data input to understand the causes of bias is not as clear-cut. This often requires integration of XAI-specific strategies to further investigate the decision boundaries.

2.3 XAI Paradigms for Detecting Bias

XAI refers to AI systems designed to provide clear, interpretable, and human-understandable explanations for their decisions and predictions [32]. XAI algorithms are designed to address the "black-box" nature of large prediction models by explaining the thought process behind the produced model outputs. For example, when an AI model predicts a high risk of cardiovascular disease, XAI can explain how factors like age, cholesterol levels, and lifestyle contribute to the prediction. If an AI model disproportionately predicts higher risks for certain demographic groups, the XAI module should be able to pinpoint the biased features driving these outcomes. An XAI algorithm should also be able to visualise links between historical data and predicted risk, encouraging preventative actions. These individual elements are known as model **explanations**.

There is a wide range of classifications on XAI subtypes in the literature, as highlighted by the recent review of Speith [33]. This report will focus on the distinction between **model-agnostic** and **model-specific** subtypes, more relevant to use cases within the MMAI domain. More specifically, model-agnostic methods are independent of the internal architecture of the AI model. These algorithms typically include perturbing inputs and observing changes in the output layer, or using surrogate models to approximate the distance to the original decision [34]. Examples include SHAP (SHapley Additive exPlanations) [7], LIME (Local Interpretable Model-agnostic Explanations) [35] or other permutation-based importance mechanisms. Meanwhile, model-specific algorithms are tailored to concrete architectures. They access and interpret the inner workings of the model to generate explanations. Examples include Grad-CAM [36] for CNNs or self-attention layers previously described in **Transformers**. Despite the fact that model-specific **XAI** algorithms are often more efficient since they directly exploit the model structure, model-agnostic algorithms are generally preferred. This is because they can be applied uniformly across different model types and data modalities without restricting the components used within each modality.

SHAP is the most-prominently used model-agnostic algorithm for post-hoc analysis in predictive analytics [7]. It is based on Shapley values from cooperative game theory and provides consistent and interpretable feature importance scores for predictions. These traits make it particularly useful for multimodal tasks as it can handle diverse input types. DeepSHAP is a variant designed for DL architectures, combining concepts from Shapley values with backpropagation with contribution weights [37]. For example, DeepSHAP can explain how specific pixel regions in medical images contribute to diagnoses (Figure 5A). TreeSHAP is an optimised variant for tree-based models like Random Forest, Gradient Boosting Machines (GBMs), and XGBoost [38], [39]. This is particularly useful for explaining individual-level predictions in static EHR data (Figure 5B). Similarly, DeepSHAP can also be repurposed to support free-text explanations in **Transformers**, linking tokenized segments from clinical notes to positive or negative contribution weights. More recent modifications of **SHAP** proposed by Parcalabescu and Frank [40] have also managed to leverage text and imaging data to produce attention-based image-text alignment scores (**MM-SHAP**) based on popular Transformer architectures like CLIP [41]. It emphasizes the utility of aggregating Shapley values across modalities to produce relative degrees of dependence – the degree to which modalities are useful across model predictions. These concepts are covered in Section 3.3.2.

2.4 MMAI Applications for Predictive Analytics

There have been notable reported benefits of multimodal learning algorithms over single-model approaches for various diagnostic and prognostic tasks. A recently developed **MMAI** framework by Soenksen et al. evaluated over 14,000 model variants across 11 unique data sources and 12 prediction tasks (Holistic Medicine in AI) [44]. The framework consistently produced models that outperformed single-source approaches by 6-33%. This demonstrated the potential of **MMAI** approaches for improving pathology detection, healthcare analytics, and precision medicine recommendations. A systematic review by Huang et al. highlighted the potential of fusing medical imaging and electronic health records (EHR) to enhance diagnostic accuracy and patient outcomes [13]. This integration is crucial as it mirrors the real-world practice where clinicians use both imaging, text and clinical data for informed decision-making.

These approaches are largely driven by the availability of open-source datasets to support experiments on multimodal fusion. The most prominent example is the **MIMIC-IV** database [9], covering a range of modalities driven from **EHR** data in general hospitalisations, covering patient measurements, orders, diagnoses, procedures, treatments, and de-identified free-text clinical notes. An emerging alternative is the recently developed **INSPECT** dataset, containing longitudinal data specifically focused on patients at risk of pulmonary embolism [45]. Although it is a more select

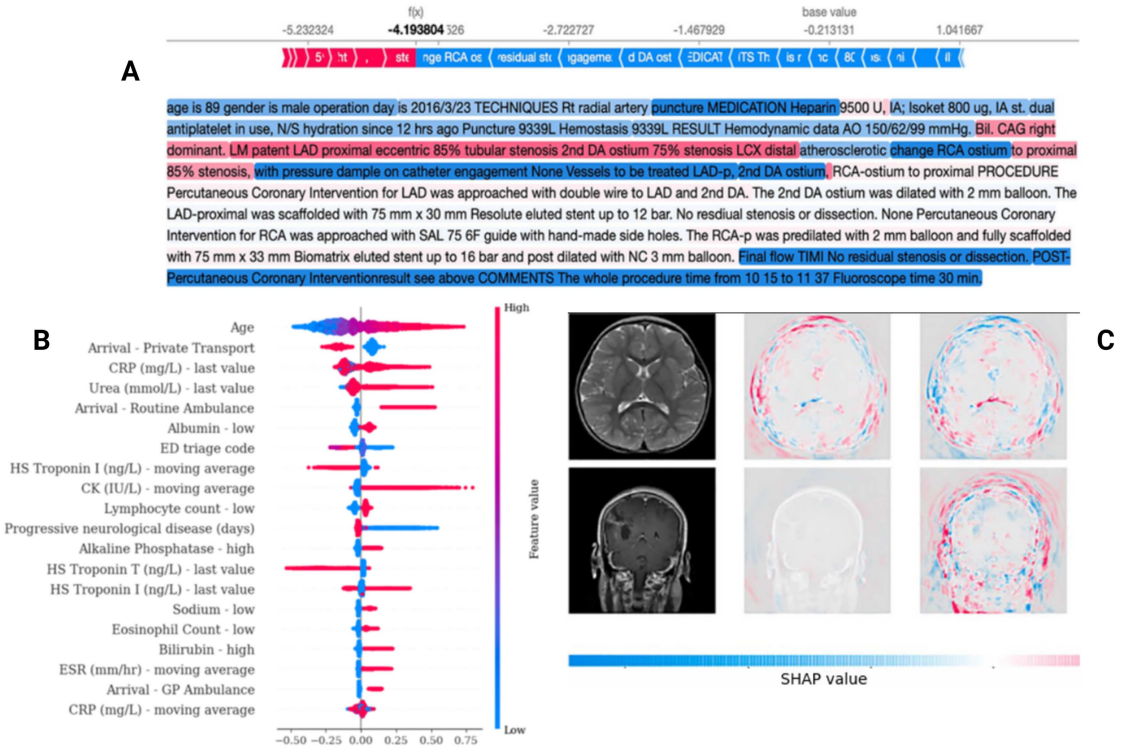

Figure 5: The various interfaces of **SHAP** representing feature importances in structured and unstructured data. A - patient-level evaluation on a discharge note segment (reproduced with permission from Li et al. [42]); B - global-level evaluation on tabular data (original work); C - patient-level evaluation on brain MRI scans in oncology (reproduced with permission from Ladbury et al. [43]).

population, if offers the ability to predict diagnostic and prognostic tasks, combining 3D medical imaging data with radiology reports and patient health records, making it highly flexible for capturing detailed contextual information.

Due to the variability of reporting of fairness, real-world examples of these metrics are few and far between. Jun et al. linked longitudinal EHR data along with social determinants of health to examine healthcare disparity in infection-related 30-day mortality [46]. It is one of the rarer studies applying statistical analysis to discover elements of bias by mapping real-world socio-demographic characteristics. Zhou et al. used a correlation ranking mechanism to optimise for imbalanced data for detecting fetal brain age using MRI scans [47]. They impose fairness-specific constraints to recalibrate feature representations by age to account for underrepresented individuals in routine data.

Applications of **XAI** in multimodal scenarios are also limited due to the complexity of integrating multiple explainable interfaces fine-tuned for each modality. Mortanges et al. propose the concept of using a LLM as an '**XAI orchestrator**' [48] trained on model explanations and fine-tuned to a target knowledge base of clinical concepts (e.g. BioBERT) [49]. This could be applied to model-agnostic methods, such as **SHAP**, to transcribe and efficiently communicate multimodal explanations to clinicians. SHAP itself has been widely adopted in many deep unimodal architectures for diagnostics of long-term conditions, such as Parkinson's disease [50], glioblastoma [51] or heart disorders [52]. By highlighting abnormal alterations in low-level inputs in imaging, textual or tabular datasets, SHAP values have increased potential in acute care, such as capturing high-risk segments to aid in patient referral and triaging procedures.

3 Methods and Software

This section will overview the building blocks of the **MM-HealthFair** framework, covering its **DL** components, fusion approach and evaluation procedures for quantifying fairness and explainability. This will also cover some of the essential implementation tools for processing data modalities, developing the fusion approach, evaluating multimodal fairness and explaining the model predictions. In this case study, the framework focuses on validating predictions at the patient-level using the MIMIC-IV database [9]. Predictions are assessed at the point of hospitalisation, where each patient has a linked ED attendance (recorded in MIMIC-IV-ED) associated with the episode (later described in Chapter 4). A broad summary of the workflow and framework components is provided in **Figure 6**.

In the previous project iteration, data pipelines were created to process demographics, lab testing and vital signs data to support fusion across tabular and time-series data. The latest version of **MM-HealthFair** extends this functionality by enriching the existing EHR data and integrating the discharge summaries as a third modality. This pipeline now supports model assessment over four distinct outcomes: **(i) in-hospital death**, **(ii) extended hospital stay**, **(iii) admission to intensive care or high-dependency unit** and **(iv) non-home discharge (transfer to institution)**. Engineered features are collated from previous hospitalisation history, covering EHR data on demographics, diagnoses, outpatient visits, prescriptions, provider orders, vital signs and discharge summaries. Collected sensitive attributes in the demographics history are used to describe healthcare disparities by **gender**, **ethnicity**, **health insurance** and **marital status**.

In the latest implementation, a pre-trained **BioBERT** (Bidirectional Encoder Representations from Transformers) tokenizer [49], leveraging information from biomedical corpora within discharge summaries, is used to construct token embeddings describing the hospital course of each patient. Modalities across the three **DL** components are fused at the **feature-level**, producing a unified output prediction layer describing the probabilities of the target outcome. The multimodal **fairness** evaluator then provides metrics to evaluate disparities across sensitive groups as a result of multimodal fusion. Finally, the **SHAP** estimator measures feature importance between the three feature modalities and provides options to investigate biases in high and low-risk groups.

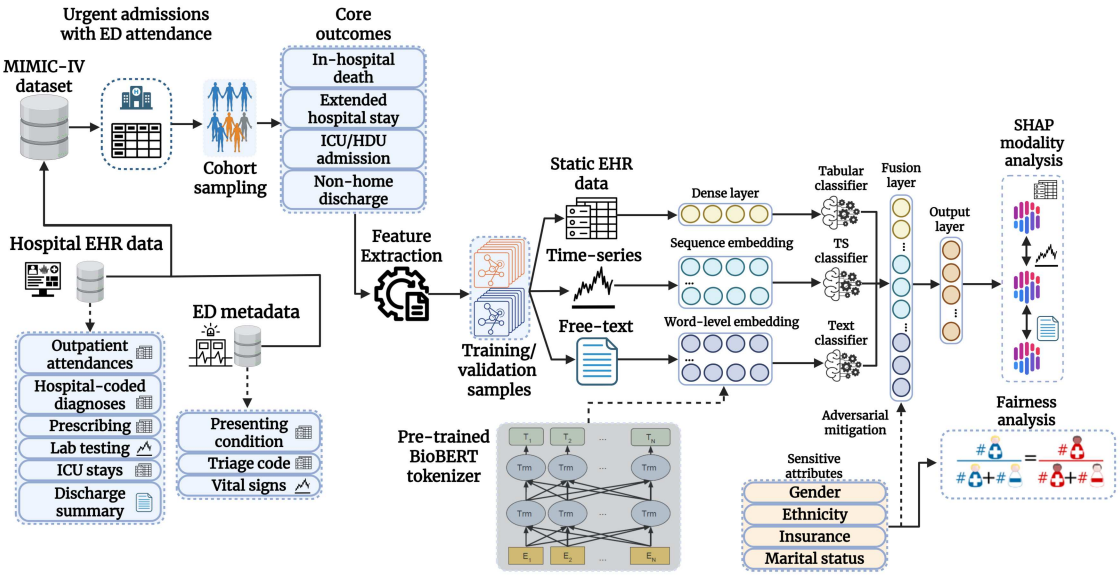

Figure 6: Conceptual overview of the main components of the **MM-HealthFair** framework, covering multimodal fusion across three modalities: tabular, time-series and text. Figure generated with BioRender.

3.1 Multimodal Architecture

The framework currently leverages an **intermediate fusion** approach, reflecting **MM-B2** type fusion across three deep neural network components representing each modality. This approach allows for data-efficient customisation, supporting independent **DL** architectures suitable for tabular, time-series and text-based classification tasks. Although more computationally demanding, this strategy achieves a reasonable middle ground between learning deep contextual representations and the ability to select and discard specific modalities, if not readily available. To understand how the fusion mechanism operates, we must first look into the building blocks of the three unimodal learning components.

3.1.1 Unimodal Components

The three **DL** architectures **MM-EHR**, **MM-TS** and **MM-NT** form the basis of the multimodal learning framework (Figure 7), where each classifier can also produce an isolated set of class probabilities to support unimodal learning.

MM-EHR is represented by a simplified Multi-layer Perceptron (MLP) classifier developed for tabular data with normalisation layers as regularisers to reduce risk of overfitting. This includes training on point measurements and count features covering demographic history, prescriptions, provider orders and diagnoses. Batch normalisation and dropout in particular, are established foundational algorithms to reduce the impact of weight initialisation on the learning rate, contributing to better model stability [53], [54]. As per the previous iteration of the project, **MM-TS** uses Long-short Term Memory (LSTM) networks to learn from sequential measurements in vital signs and lab tests measured during the initial stages of hospitalisation [55]. Each data component passes through its own LSTM branch and is concatenated before its final dense layer. This is required because each data source has variable sampling points and differences in the intervals of measurement. To concatenate the sequential representations across the two data sources, the representations are packed together and padded to a fixed length, as a way of handling time-wise missingness.

Finally, the **MM-NT** segment uses a **Transformer-encoder** classifier (two-unit encoder framework with self-attention) to classify each outcome using the discharge summaries data [12]. The raw text tokens are first extracted from the **Brief Hospital Course (BHC)** segments of the MIMIC-IV discharge letter and split into sentences [56]. To integrate domain knowledge, the BioBERT tokenizer is used to construct coded embeddings, capturing nuances and terminology of biomedical language, also collated from discharge letters [49]. Like traditional BERT, the tokenizer inserts special tokens such as [CLS] (for classification) and [SEP] (to separate segments), which are essential for the model to interpret the inputs correctly. This streamlines the preprocessing pipeline for the notes data, as the tokenizer directly handles splitting, truncating, and padding sequences to the required length for the **Transformer-encoder**.

After the sentence embeddings are constructed, they are converted into dense vector representations and fed into the model. Since **Transformers** lack inherent sequence order awareness [57], positional encodings are added to the embeddings to provide information about token positions in the sequence. The input passes through two identical encoder layers. Each layer consists of a:

- **Multi-head self-attention module**: Enables the model to focus on different parts of the input sequence simultaneously.
- **Feedforward neural layers**: Processes the non-linear output of the attention mechanism, which is then normalised and fed to the next component.

Each component contains a standardised logistic loss function with binary cross-entropy (BCE), with the ability to adjust class weights to handle the imbalance in clinical outcomes. In general, BCE loss functions assume that all samples contribute equally to the loss, which can cause the model to favor the majority class and underperform on the minority class. By introducing class weights, we modify the loss so that misclassifying a minority class sample incurs a larger penalty. In practice, this weight is typically inversely proportional to the number of subjects with the outcome [58].

3.1.2 Intermediate Fusion Approach

The main multimodal component follows a simple intermediate fusion approach, through a **concatenation** layer (Figure 8). Each modality is first processed independently by its own neural network branch, extracting high-level representations up to their last hidden layer. The feature vectors from each modality's last hidden layer are squeezed into dense linear layers and then concatenated to form a single, fused representation. This fused vector is then passed through one fully-connected (dense) layer, enabling the model to learn interactions between modalities before making the final prediction. This approach is chosen for its simplicity and computational efficiency, allowing fusion across more than two modalities.

However, using concatenation to fuse modalities does come with a few drawbacks. In particular, it may be less transparent regarding each modality's influence on the output. As this approach relies on the learned representations of the unimodal components, it is not flexible enough to emphasize or suppress modalities based on context. It is also not possible to measure modality importance using the framework itself, requiring use of SHAP or other model-agnostic techniques [7]. For high-dimensional inputs it may also perpetuate overfitting if one of the unimodal frameworks is unstable.

This issue was addressed by Martin et al. [10] in the previous iteration of this project. They applied an alternative approach using Multi-adaptation Gates (MAGs) enabling dynamic weighting of the multimodal features before fusing modalities [59]. MAGs operate by selecting a primary modality for constructing a set of gating vectors, highlighting

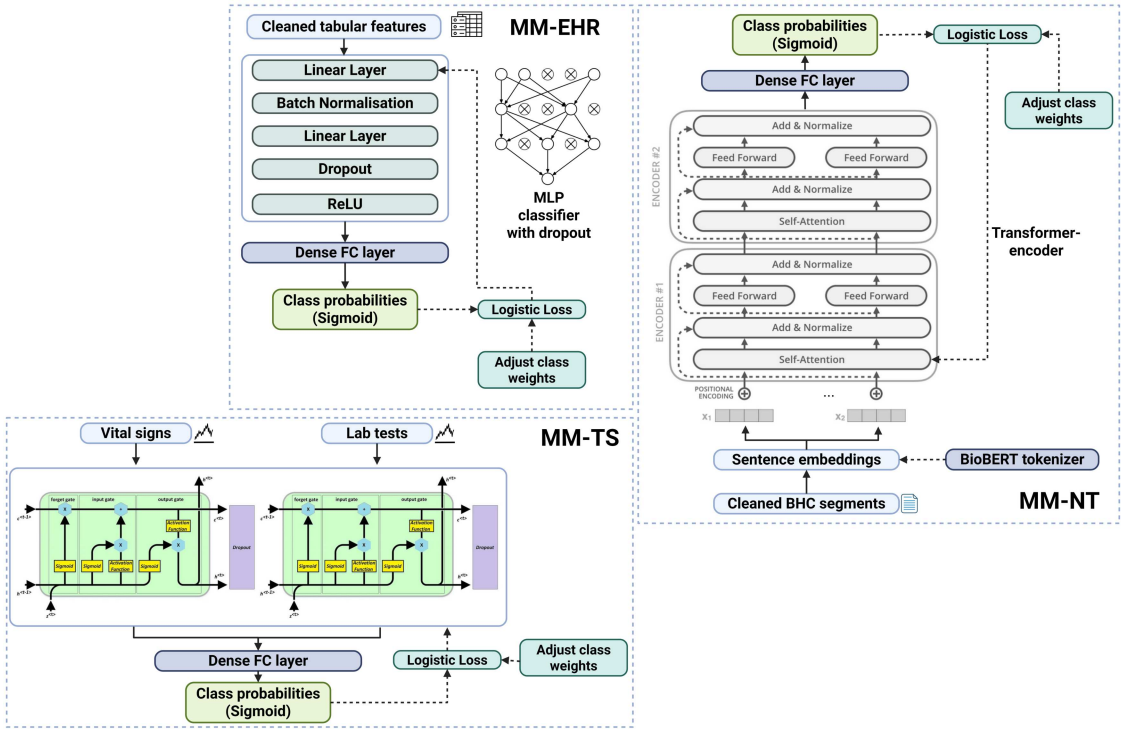

Figure 7: Unimodal Deep Learning architectures used to later develop a feature-level fusion framework. **MM-EHR** describes the tabular classifier developed for count features (demographics, prescriptions, provider orders and diagnoses). **MM-TS** builds a unified **LSTM** framework with a concatenated layer across the vital signs and lab test measurements data. The **MM-NT** framework leverages a **Transformer-encoder** network with self-attention layers to learn sentence embeddings from the BHC (Brief Hospital Course) segments of each discharge letter. Figure generated with BioRender.

the relevant information in the accompanying modalities whilst being conditioned on the primary one. A control parameter α can then be used to control how much shifting is applied to the primary modality to modify the decision boundary. While this approach is currently available for fusion between **MM-EHR** and **MM-TS** models, a more efficient implementation is required to control for adaptivity against the larger **MM-NT** network.

3.2 Quantifying and Adjusting for Fairness

This section looks into the algorithms for measuring fairness across sensitive groups and proposes ways to enforce fairness constraints to control for these attributes. These mechanisms generally do not depend on the model setup and can be empirically tested across different unimodal and multimodal architectures, as long as the output prediction layer and the target loss function remain consistent across the networks (e.g. solving a binary classification task).

3.2.1 Fairness metrics

As listed by a recently conducted survey by Mehrabi and colleagues [60], there is a wide range of studied definitions of fairness in the literature. This section will re-iterate some of the most widely used definitions in the context of a **ML** classification problem.

Perhaps the simplest definition is that of **Demographic Parity (DP)**. It states that perfect statistical parity is achieved when a set of predictions $\hat{Y}$ for a positive outcome remains the same regardless of whether the subject is in the group A or not (e.g. 30% male and 30% female patients at risk of in-hospital death) [61]:

$$P(\hat{Y}|A=0) = P(\hat{Y}|A=1) \tag{1}$$

DP is one of the simplest measures of patient equality, but it does not consider whether groups differ in actual outcome rates. Thus, depending on the context, enforcing it can lead to unfair scenarios when opportunities for treatment are naturally disproportionate between individuals (e.g. chronic disease prevalence).

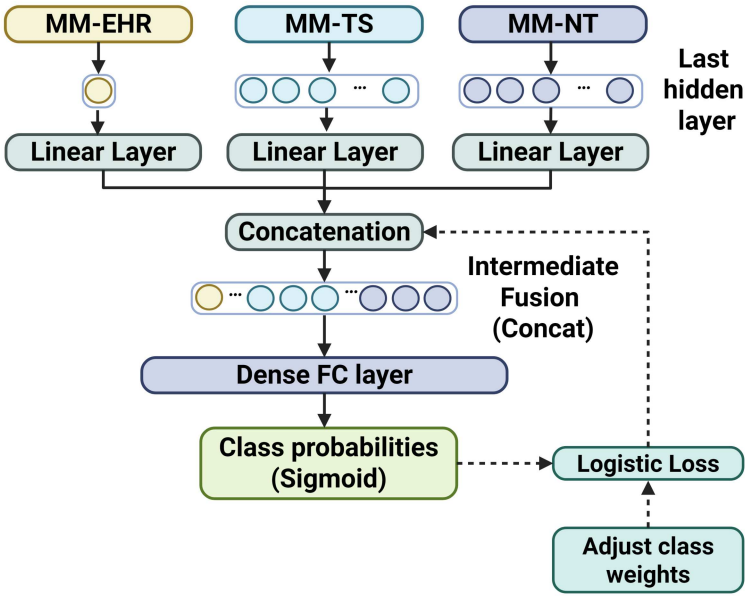

Figure 8: Multimodal Deep Learning framework used to incorporate **intermediate (feature-level) fusion** across the three data modalities. Figure generated with BioRender.

This can be taken into account by measuring **Equalised Odds (EQO)** and **Equal Opportunity (EOP)**. To guarantee **EQO** for a predictor $\hat{Y}$ with respect to an attribute A and an outcome Y, $\hat{Y}$ and A must be independently conditional on Y [62]. This means that the probability of a person in the *positive* class being *correctly* assigned as having the outcome and the probability of a person in a *negative* class being *incorrectly* assigned a positive outcome should both be the same across groups. Thus, each group should have equal rates for true positives (TPR) and false positives (FPR) (e.g., an area-under-the-curve (ROC-AUC) of 0.8 in both male and female patients):

$$P(\hat{Y} = 1|A = 0, Y = y) = P(\hat{Y} = 1|A = 1, Y = y), y \in \{0, 1\} \tag{2}$$

In cases where the FPR is not relevant to the classification task (e.g. guaranteeing equal precision across genders to ensure diversity for a clinical trial), the **EOP** can be used instead. It states that the probability of an individual with the outcome being assigned as positive should be equal across group members:

$$P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1) \tag{3}$$

A worked example of how these measures are estimated in a patient population can be seen in Figure 9. It is also worth noting that there are other more treatment-specific measures that can be used to test group fairness, such as **conditional statistical parity** [63], **treatment equality** [64] and **test equality** [65]. Fairness can also be measured at the individual-level by approaches such as **Fairness through Awareness** (giving similar predictions to similar individuals based on distance scores) [61], **Fairness through Unawareness** (excluding attributes from the decision-making process) [66], or counterfactual fairness (if representations can remain fair under any context of the input data) [26].

3.2.2 Statistical Validation with Bootstrapping

One efficient way of estimating confidence intervals in classification tasks is through Bias-corrected and accelerated intervals (BCa) [68]. The main advantage of this method is the ability to correct for bias and skewness in the distribution of bootstrap estimates [69]. This yields more accurate and reliable confidence intervals, especially when the underlying statistic is not symmetrically distributed or biased, which is a likely scenario in imbalanced classification. The BCa algorithm operates by repeatedly resampling the data and recalculating the fairness metric to build a bootstrap distribution (Algorithm 1). To achieve this, it computes two terms:

- Bias term z_0: adjusts for systemic bias in the samples by considering the proportion of bootstrap estimates less than the observed statistic (typically the mean);

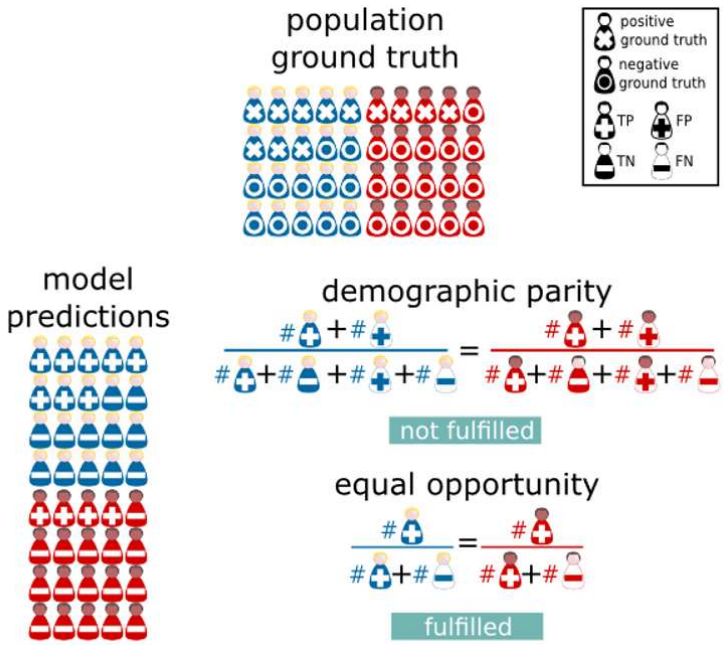

Figure 9: Case example for algorithmic fairness, reproduced with permission from Lara et al. [67]. In this example, two sub-populations characterised by different sensitive attributes (red and blue) have varying prevalence for a prognosis (e.g. in-hospital death). Positive predictions are marked with '+', while negative predictions are marked with a '-' sign. In this case, the requirement for full **demographic parity** is not fulfilled as the positive prediction rates vary between sub-groups: 40% (8 positive predictions over 20 cases) for the blue sub-group vs. 20% (4 positive predictions over 20 cases) for the red sub-group. Meanwhile, if we observe the population ground-truth ('x' indicating subjects with the outcome, and 'o' indicating subjects without it), the model achieves perfect **equal opportunity**, as true positive rates match for both sub-groups (8 true positives out of 8 positive ground truth cases for the blue sub-group and 4 true positives out of 4 positive ground truth cases for the red sub-group).

- Acceleration term a: adjusts for skewness in the bootstrap distribution;

The acceleration term a can be approximated using the **jackknife** method (also known as leave-one-out resampling) [70]. The idea is to estimate a test statistic for a target metric θ (e.g. **DP**) based on a set of $n-1$ **jackknife** replicates, leaving each individual sample out once during the computation. Then, the global jackknife statistic θ_{jack} will be the mean of all replicates θ_i. The inference values u_i for the acceleration factor a will then be the absolute differences between θ_{jack} and each sub-statistic:

$$\bar{\theta}_{jack} = \frac{1}{n}\sum_{i=1}^{n}\theta_{(i)} \tag{4}$$

$$u_i = \bar{\theta}_{jack} - \theta_{(i)} \tag{5}$$

The acceleration factor can then be defined from the skewness of these differences, quantifying how asymmetrically individual data points influence the statistic:

$$a = \frac{\sum_{i=1}^{n} u_i^3}{6(\sum_{i=1}^{n} u_i^2)^{3/2}} \tag{6}$$

A detailed demonstration by Erik Drysdale [71], implemented in Python shows how BCa draws empirically stable samples when assessing model precision compared to other resampling methods.

Algorithm 1: Estimating BCa Confidence Intervals with **jackknife** resampling for a fairness metric.

Data: Dataset of size n; Fairness metric θ; Bootstrap samples B; Confidence interval range (e.g. 95% CI)
; **Result:** CI upper and lower bounds α_1, α_2
for $b = 1$ *to* B **do**
  Sample n data points with replacement from the dataset to form bootstrap sample b.
  Compute θ_b^*.
end
Store θ_b^* values. **for** $i = 1$ *to* n **do**
  Remove the *i*-th observation to form the jackknife sample.
  Compute each $\theta_{(i)}$.
end
Compute the mean of jackknife estimates (Equation 4).
Compute the acceleration parameter a (Equations 5 and 6).
Compute the bias-correction parameter z_o as the standard normal quantile of the proportion of bootstrap estimates less than the observed statistic θ, where Φ^{-1} is the inverse standard normal cumulative distribution function (CDF):

$$z_0 = \Phi^{-1}\Big(\frac{\#\{\theta_b^* < \bar{\theta}\}}{B}\Big)$$

Compute the adjusted percentiles α_1, α_2, e.g. for a confidence interval of level $1 - \alpha = 95\% : \alpha = 0.05$, where z_p is the p-th quantile of the standard normal distribution, we define:

$$\alpha_1 = \Phi\Big(z_0 + \frac{z_0 + z_{\alpha/2}}{1 - a(z_0 + z_{\alpha/2})}\Big)$$

$$\alpha_2 = \Phi\Big(z_0 + \frac{z_0 + z_{1-\alpha/2}}{1 - a(z_0 + z_{1-\alpha/2})}\Big)$$

3.2.3 Adversarial Mitigation

The ability to compute fairness metrics with confidence intervals is already a powerful mechanism to explore biases in multimodal algorithms. However, performing fairness analysis on a post-hoc basis might be too late in a model is already in clinical use. For this reason, it is preferable to have the option to adjust the model for biases at training time. One way to achieve this is through **adversarial mitigation** (also known as debiasing) algorithms [72], which aim to enforce fairness constraints on specific protected attributes. This is a form of **in-process** mitigation.

The goal is to feed the input features X, class labels Y and a held-out set of sensitive features Z into a classification estimator θ, predicting Y from X, while enforcing fairness constraints based on **DP**, **EQO** or **EOP** metrics (as described by the authors of the **Fairlearn** framework [29]). The main mechanism of an adversarial network consists of two components:

- The predictor model (θ_p): trained to predict the outcome in a standard classification setting.
- The adversary model (θ_a): trained to predict the sensitive attributes Z (e.g., ethnicity group) using the prediction layer from $\hat{Y} = f_\theta(X)$.

The predictor is penalised if the adversary can accurately infer the sensitive attribute, forcing it to remove bias, corresponding to the fairness measure. This will reduce performance estimates, such as **ROC-AUC**, while increasing the score target fairness metric. Thus, when training an adversarial mitigation algorithm, we want to decrease the adversary's ability to predict the sensitive features from the predictor's outputs (case of **DP**), or jointly from the predictor's outputs and true labels (case of **EQO** or **EOP**). To achieve this, we split the target loss function into two parts: L_p (loss for the main prediction task, equivalent to binary cross-entropy) and L_a (loss for the adversarial network A_ϕ, equivalent to cross-entropy from predicting Z), enforcing the fairness-accuracy tradeoff with a control hyperparameter λ. Then, we can compute the objective function targeting the minimisation of L_p and maximisation of L_a:

$$\min_\theta \max_\phi \; \mathcal{L}_{\text{p}}(Y, \hat{Y}) - \lambda \mathcal{L}_{\text{a}}(Z, A_\phi(\text{input})) \tag{7}$$

The adversary input will depend on the enforced fairness constraint. In the case of enforcing **DP**, the adversary receives only the predictor's output $\hat{Y}$ and attempts to predict Z from it:

$$\min_{\theta} \max_{\phi} \; \mathcal{L}_{\mathrm{p}}(Y, \hat{Y}) - \lambda \mathcal{L}_{\mathrm{a}}(Z, A_{\phi}(\hat{Y})) \tag{8}$$

In the case of **EQO** the adversary receives both the predictor's output $\hat{Y}$ and the true label Y and attempts to predict Z from $(Y, \hat{Y})$:

$$\min_{\theta} \max_{\phi} \; \mathcal{L}_{\mathrm{p}}(Y, \hat{Y}) - \lambda \mathcal{L}_{\mathrm{a}}(Z, A_{\phi}(\hat{Y}, Y)) \tag{9}$$

Finally, when optimising for **EOR**, the adversary will receive both the predictor's output $\hat{Y}$ and the true label Y but is trained only on samples where Y is positive ($Y = 1$):

$$\min_{\theta} \max_{\phi} \; \mathcal{L}_{\mathrm{p}}(Y, \hat{Y}) - \lambda \mathcal{L}_{\mathrm{a}}(Z, A_{\phi}(\hat{Y}, (Y \mid Y = 1))) \tag{10}$$

It is worth noting that adversarial learning is inherently difficult because of various issues, such as mode collapse, divergence, and diminishing gradients [29]. Mode collapse is the scenario where the predictor learns to produce one output, and because it does this relatively well, it will never learn any other output. Diminishing gradients could be introduced as a result of an adversary that is trained too well in comparison to the predictor. To remedy risks of overfitting and instability, other empirical frameworks have been proposed, such as **OxonFair**, leveraging inference methods based on the Pareto principles [30].

3.3 Multimodal explanations using Shapley values

The original **SHAP** framework developed by Lundberg and Lee [7], was derived from cooperative game theory, specifically the Shapley value concept of feature attribution. In this context, the ML algorithm is the "game". The input features are the "players" and the predicted probability score is the "outcome". By accumulating the contributions of each feature to the predicted outcome, SHAP values can be used to estimate a unified measure of feature importance. This allows both instance-level (local) explanations and overall (global) feature importances to be computed for a more comprehensive understanding of biases. Predictions of risk in classification models are typically presented through density plots showcasing bidirectional (positive or negative) impact scores measured across each patient sample (Figure 10).

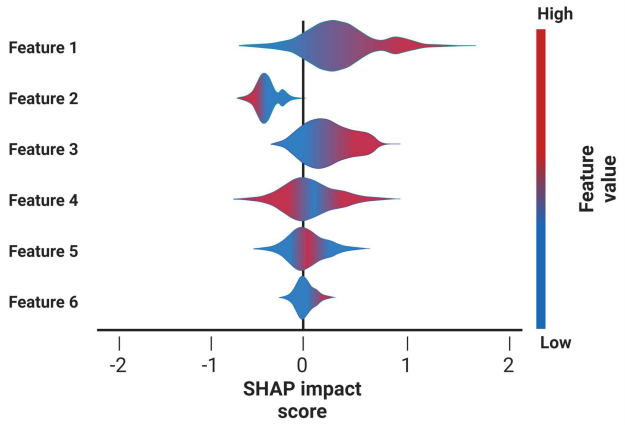

Figure 10: Overview of a **SHAP** density plot across six example model features (can also be presented as a beeswarm or violin plot). Denser regions represent larger sample sizes with these respective SHAP impact scores. Positive and negative impact scores indicate the direction of risk with respect to the outcome. Figure generated with BioRender.

In this example, Feature 1 (the variable with highest importance) highlights a positive linear association between the feature value and the impact score, increasing the risk of the target outcome (e.g. age as predictor of mortality). However, impact scores might not always highlight a linear association. In Feature 2, the presence of the variable in the

dataset is enough to indicate decrease in risk and high values further decrease this risk (e.g. sections covered in mobility assessment as predictor of inpatient falls). The relationship may be non-linear but still have an important effect on the outcome, as in Features 4 and 5. For example, in the case of Feature 4, high blood pressure might decrease or increase risk of mortality depending on the interactions with age. To investigate these types of interactions further, we can then plot individual scatterplots of the SHAP interactions across feature combinations.

3.3.1 DeepSHAP

As this work utilises **DL** architectures, the focus is on the **DeepSHAP** algorithm [7], designed for computing attribution in Deep Neural Nets (DNNs). **DeepSHAP** combines ideas from game theory and the previously developed DeepLIFT algorithm [37], which uses backpropagation through DNN layers to measure cumulative attribution scores. To fit the **DeepSHAP** algorithm we require a trained **DNN** network f. We additionally define a set of input instances (e.g. the validation set) $\mathbf{x}$ and a set of background samples $\mathbf{x}^b$ within it, balanced across a set of sensitive attributes (A) that are descriptive of the data distribution. **DeepSHAP** will then seek an explanation model $g(f(\mathbf{x}'))$, where $\mathbf{x}'$ is a binary vector indicating feature presence. For each observed sample j, ϕ_j will be the attribution score of each feature k_j with size M. The objective feature explanation model $g(f(\mathbf{x}'))$ can be expressed as follows:

$$g(f(\mathbf{x}')) = \phi_0 + \sum_{j=1}^{M} \phi_j k'_j \tag{11}$$

To optimise this model, we need to compute the local accuracy (efficiency) over the chosen background samples, where we define $E[f(x^{(b)})]$ as the expected output for these samples:

$$f(\mathbf{x}) = \mathbb{E}[f(\mathbf{x}^{(b)})] + \sum_{j=1}^{M} \phi_j \tag{12}$$

Each Shapley value for a feature ϕ_j will then be computed using the marginal contribution across the feature set F, where $f_s(\mathbf{x})$ is the model output where only the feature set S is present, iterating over each possible coalition of features k_j:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(M - |S| - 1)!}{M!} \left[f_{S \cup \{j\}}(\mathbf{x}) - f_S(\mathbf{x}) \right] \tag{13}$$

After formalising the DeepSHAP objective, we can compute these values by estimating a baseline (averaged contribution across all samples), adding a linear normalisation term and propagating the attributions via the chain rule for differentiation (Algorithm 2).

Thus, the sum of **SHAP** values explains the difference between each model output for $\mathbf{x}$ and the expected output over the background $\mathbf{x}^b$. Computing these values in a multimodal setting, however, is slightly more computationally challenging and requires us to consider the magnitude of contribution across each data modality.

3.3.2 Aggregating SHAP Values in a Multimodal Scenario

Researchers Parcalabescu and Frank recently proposed a mechanism to leverage Shapley values in order to explore multimodal contributions, that is, quantifying modality usage using SHAP aggregations [40]. While it was developed strictly for explaining outputs in Vision-Language models and Visual Question Answering (VQA) tasks [41], its simplicity indicates that it may be applicable to higher dimensionalities, including tabular and time-series data, as long as the information can be fused within a unified embedding layer, as per **Type-A** or **Type-D** Transformer architectures.

In this case, suppose we have a set of tabular n_{tb}, time-series n_{ts} and textual n_{tx} tokens. Using an intermediate fusion approach (**MM-B2**), we can theoretically compute the attribution scores independently for each modality by defining a set of modality-specific background samples $\mathbf{x}^{(b)_{tb}}$, $\mathbf{x}^{(b)_{ts}}$, $\mathbf{x}^{(b)_{tx}}$. Then we sum over the propagated contributions within the appropriate network for the chosen modality. The contribution sets can then be defined as:

$$; \phi_{ts} = \frac{1}{|B_{ts}|} \sum_{(b)_{ts}=1}^{|B_{ts}|} \phi_j^{(b)_{ts}}; \phi_{tx} = \frac{1}{|B_{tx}|} \sum_{(b)_{tx}=1}^{|B_{tx}|} \phi_j^{(b)_{tx}}; \tag{14}$$

Algorithm 2: Summary of the **DeepSHAP** algorithm for computing gradient-based attribution scores as model explanations.

Data: Deep Neural Net f; Data samples $\mathbf{x}$; Background samples drawn from a data batch $\{\mathbf{x}^b\}$ with size B; Feature set k with size M.
Result: Output SHAP vector $\phi = (\phi_1, \phi_2, ..., \phi_M)$
. Compute the baseline attribution ϕ_0, averaging the background samples $\{\mathbf{x}^b\}$.

$$\phi_0 = \mathbb{E}[f(\mathbf{x}^{(b)})]$$

Linearise each network component.
for *each layer (linear, activation, pooling), defined as* $m_{i\to j}$ **do**
Define the gradient Δy_j as the change in output neuron j due to a change in input x_i, relative to the baseline.
Compute the local attributions:

$$m_{i\to j} = \frac{\Delta y_j}{\Delta x_i}$$

end
Multiply and sum multipliers through the network layers to propagate feature contributions from output to the input data using the product over multipliers along each path from feature k_j:

$$\phi_j = \sum_{\text{paths}} \prod_{l \in \text{path}} m_l \cdot \Delta k_j$$

Aggregate over the background samples $\{\mathbf{x}^{(b)}\}$, computing each feature attribution $\phi_j^{(b)}$:

$$\phi_j = \frac{1}{|B|} \sum_{b=1}^{|B|} \phi_j^{(b)}$$

We can then consider the magnitude of a token contribution, as we are interested in measuring whether a token is active in a modality, regardless of the direction it pushes the prediction into. **MM-SHAP** defines this magnitude as the proportion of modality contributions, or their relative tabular (TB), time-series (TS) or textual (TX) degrees:

$$TB = \frac{\phi_{tb}}{\phi_{tb} + \phi_{ts} + \phi_{tx}}; TS = \frac{\phi_{ts}}{\phi_{tb} + \phi_{ts} + \phi_{tx}}; TX = \frac{\phi_{tx}}{\phi_{tb} + \phi_{ts} + \phi_{tx}}; \tag{15}$$

This way **MM-SHAP** can theoretically be extended to support any number of modalities. On the individual-level, this enables fine-grained analyses of the relevance of multimodal inputs, which can be used to examine edge cases of unfair scenarios. On the global-level we can average individual-level **MM-SHAP** scores, because of the additivity property of Shapley values. Hence, it can help link fairness metrics from Section 3.2 to dependence on specific modalities. One potential limitation is the difficulty of computing all possible coalitions between input tokens for multimodal Shapley aggregates. This may require approximation methods, such as random Monte Carlo sub-sampling [40], which may lead to less exact explanations. It may also be more expensive for pre-trained models with more extensive masking configurations. As **MM-SHAP** is yet to be implemented for tabular and time-series modalities, the process of fine-tuning this framework for binary classification tasks could require additional regularisation techniques to account for noisy outputs within the Shapley distributions.

3.4 Implementation Tools

The **MM-HealthFair** framework was designed to support flexible classification pipelines, providing an end-to-end pipeline for multimodal fusion, evaluation and fairness investigation. The framework was built on **Python** version 3.10.11 and tested on a local Windows 11 machine. Additionally, model training and evaluation were performed on a Microsoft Azure machine using a Windows 10 Server with the following specifications: 1 x NVIDIA Tesla T4 GPU and 4 x vCPUs (28 GiB memory). Some of the important Python packages used to create the multimodal pipeline are described below.

The *polars* library (0.20.23) was used to support efficient reading and processing of time-series and notes data (>110M lab test and vital signs events and >300K discharge summaries) [1]. It builds memory-efficient computation graphs to support multithreading during the feature extraction process, through its parallel execution engine. The **DL** frameworks were developed using *PyTorch* (2.5.1) with its *Lightning* interface (2.5.0.post0) [2], supporting re-usable components, such as data loaders, training schedulers, and optimisers. For tokenizing the text inputs, the *SpaCy* library (3.7.5) with its *en-core-sci-md* corpus containing biomedical data with a vocabulary and 50K pre-trained word vectors was used [3]. Each sentence from the discharge summary was fed into the tokenizer and embedded via the pre-trained **BioBERT** model [73] by freezing the model weights and feeding the tokenizer input up to the last hidden layer of the model. The **BioBERT** model is available for download on the HuggingFace platform [4].

To train, validate and checkpoint the models, the *Weights & Biases (W&B)* (0.17.0) library [5] was used to create downloadable model artifacts. It streamlines the MLOps lifecycle, making it easier to build and deploy models in a reproducible manner. It also integrates seamlessly with PyTorch Lightning components providing the ability to track experiments, create logging scripts and version control models. To use the service within the **MM-HealthFair** framework, we have to setup an account and create an example project space. Then we modify the *args.project* argument that sets the training scheduler (located in the *train.py* script), as follows:

```
if use_wandb:
    logger = WandbLogger(
    log_model=True,
        project=args.project,
        save_dir="logs",
    )
    # store config args
    logger.experiment.config.update(config)
```

From then on, we simply set our training parameters in the *model.toml* config file and run the training. After the first model run, we will be prompted to set up an access token. The W&B service will automatically create a model artifact and generate dynamic performance plots as the training progresses. After the training is done and if the run was successful, a coded checkpoint will be stored inside an online artifact object, which we can be downloaded and added to a target folder:

```
import wandb
run = wandb.init()
artifact = run.use_artifact("/nhs-mm-healthfair/model-31357dly:v0", type="model")
artifact_dir = artifact.download()
```

To quantify fairness and perform fairness-specific error analyses, the **MM-HealthFair** framework uses components from the *Fairlearn API* (0.12.0) [6]. It supports a range of fairness metrics for quantifying model disparity and a number of utilities for weighing performance across sensitive groups. It also supports adversarial mitigation algorithms for retraining models to optimise **DP**, **EQO**, and **EOP**, as described in Section 3.2.3.

[1] Polars API: https://docs.pola.rs/user-guide/getting-started/.

[2] Pytorch Lightning: https://lightning.ai/docs/overview/getting-started.

[3] SpaCy pipelines containing the biomedical corpus (originally built with SciSpaCy, but does not require it to download and use the word vectors): https://allenai.github.io/scispacy/.

[4] BioBERT (Bio+Discharge Summary) pre-trained clinical note embeddings.

[5] W&B web service: https://wandb.ai/.

[6] Fairlearn API: https://fairlearn.org/v0.12/user_guide/index.html.

4 Data and Preprocessing Steps

This section will describe the relevant parts of the data extraction, cleaning and preprocessing procedures for multimodal learning and fairness evaluation. It covers the key data sources extracted from the **MIMIC-IV** database, used to create features describing the tabular, time-series and free-text modalities.

4.1 Data Sources

To construct the multimodal learning pipeline we use the **MIMIC-IV** open database consisting of retrospectively collected medical data in secondary care [9]. The main database consists of a large deidentified dataset of patients admitted to the emergency department (ED) or an intensive care unit (ICU) at the Beth Israel Deaconess Medical Center in Boston, MA. This allows the linkage of hospital episodes with previous ED attendance and prediction of hospital-related outcomes as early as the point of ED arrival. The **MM-HealthFair** framework uses four **MIMIC-IV** data modules:

- *hosp*: measurements recorded during hospital stay for training, including demographics, lab tests, prescriptions, diagnoses and care provider orders
- *icu*: records individuals with associated ICU admission during the episode with additional metadata (used mainly for measuring the ICU admission outcome)
- *ed*: records metadata during ED attendance in an externally linked database (**MIMIC-IV-ED**); primarily used to link the target population and extract vital sign measurements for the time-series modality
- *note*: records deidentified discharge summaries as long form narratives which describe the reason for admission to the hospital, the hospital course of the patient, and any relevant discharge instructions

This framework currently uses the latest available version of **MIMIC-IV** 3.1, which includes an extended population compared to the previously used version 2.2, covering 364627 patients, 546028 total admissions and 94458 stays in ICU.

The framework additionally uses a curated clinical notes dataset built on top of the *note* component, called **MIMIC-IV-Ext-BHC** [74]. **MIMIC-IV-Ext-BHC** is a labeled clinical notes dataset originally adopted for hospital course summarisation tasks with LLMs. However, it is also applicable for supervised learning tasks, as it has rigorously cleaned and standardised segments of the Brief Hospital Courses (BHC) from the discharge letter. The preprocessing pipeline follows a number of key steps to extract and clean the original letters using the *note* module:

- **Separation and Type Conversion**: BHC sections separated from the "text" column with regular expressions used to filter and extract the entire substring under the heading "Brief Hospital Course".
- **Whitespace and Formatting Cleanup**: removed extraneous whitespace with consistent formatting across the dataset.
- **Section Identification and Headers Standardisation**: section headers standardised to uppercase within angle brackets to facilitate easier parsing and extraction.
- **Further Cleanup**: removed line breaks, extraneous symbols, and reformatting the BHC segments for uniformity.

4.2 Data Curation

This part will briefly describe the data collection and extraction procedures, relevant to cleaning and preparing the **MIMIC-IV** datasets for multimodal learning. An overview of the components and outputs from the multimodal data extraction procedure can be seen in Figure 11. The *extract_data.py* module uses the base **MIMIC-IV** database and the **MIMIC-IV-Ext-BHC** extension as raw data inputs. This feeds into a series of sequential operations for extracting and cleaning the individual data modalities. Thus, the data sources used for each modality include:

- **Tabular (Static EHR) modality**: demographics, prescriptions, previous diagnoses and provider orders
- **Time-series modality**: ED vital signs, in-hospital lab tests and measurements
- **Text modality**: BHC segments from clinical notes

To ensure consistent multimodal representation, we restrict the population to individuals with at least one prior significant hospitalisation record (with discharge note containing a BHC segment, and with any vitals or lab tests measured during

the episode). The main unit of analysis is subject-level, and the final ED attendance per eligible patient is used as the prediction timepoint. The output of the script will generate three independent data files to be used for ML-specific preprocessing and developing the training data.

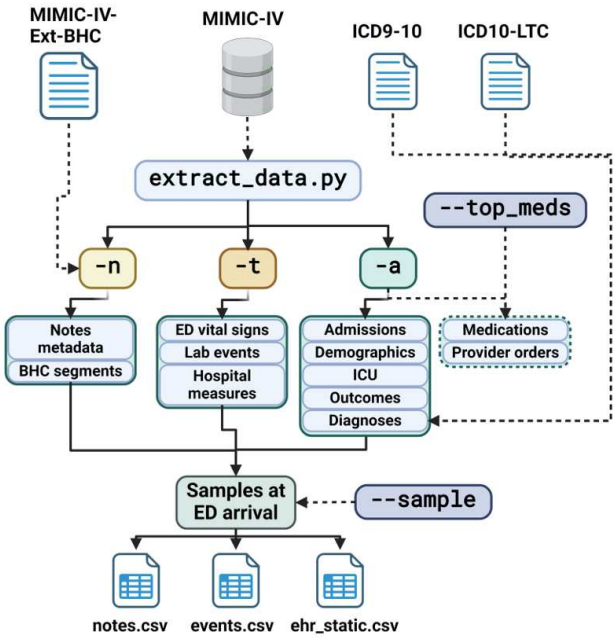

Figure 11: Overview of the **MM-HealthFair** data curation and extraction procedures, leveraging the base **MIMIC-IV** dataset and the **MIMIC-IV-Ext-BHC** extension for curating the notes segments. Figure generated with BioRender.

4.2.1 Tabular Data Pipeline

The feature collection pipeline for tabular data (static **EHR** features) underwent through a number of extensions to enrich the existing data representations. These primarily included engineering of count features, describing the patient's medical history. The static data also included information about the four sensitive attributes used in the fairness analysis: gender, insurance, marital status and ethnicity.

Clinical diagnoses were first converted from ICD-9 (International Classification of Diseases) to ICD-10 mapping wherever possible, using an existing data pipeline provided by Gupta and colleagues [75]. This was a required step to then map the ICD-10 codes to specific long-term condition groups. These mappings were reproduced from previous work studying in-hospital activity and relationships with multimorbidity [76]. The codes were grouped into 25 common chronic condition groups (22 physical and 3 related to mental health). The input features then included presence of these groups and days since last diagnosis.

Prescriptions were parsed using the online administration record data described by order date. Eligible prescriptions were labelled with status "Administered", "Started" or "Confirmed". The top 50 most common drug-level features were retained for the training data. Provider orders were extracted from specialty-grouped data from specialties described in the treatment order history. Included specialties were: nutrition, cardiology, radiology, respiratory medicine, neurology and hemodialysis.

4.2.2 Time-series Data Pipeline

The measurements for vital signs included temperature, heart rate, respiratory rate, oxygen saturation (SpO2) and blood pressure. The lab tests included the test item, value and metric across the 50 most commonly found tests in the population. Both datasets were cleaned to include valid test ranges and test names. Records with outliers greater or less than 2 standard deviations from the mean were removed. Then, the measurements were linked in a unified long format describing the measurement date, item label, value and unit of measurement. Due to the different sampling frequencies between vital signs (routinely collected every 1-4 hours) and lab tests data (recorded 1-3 times daily), we used a fixed collection interval for each data source independently, with a 30 minute interval for vitals and 5-hour interval for lab tests.

4.2.3 Notes Data Pipeline

To extract the prepared BHC segments for training, we linked the original MIMIC-IV *note* table with the **MIMIC-IV-Ext-BHC** table by hospitalisation identifier. To aggregate multiple clinical notes across the full hospitalisation history, we inserted an "<ENDNOTE> <STARTNOTE>" token in-between each BHC segment and extracted the total number of input tokens. To improve training efficiency, we then removed any de-identified fields from the note (marked as "____") and replaced any additionally introduced whitespaces.

4.3 Data Preprocessing

The curated data files from the **MIMIC-IV** pipeline are then fed into the *prepare_data.py* script for modality-specific preprocessing and training set generation (Figure 12). Additional metadata, such as the target outcome (e.g. in-hospital death), the target sensitive attributes and the time-series frequencies are pulled from the configuration file *targets.toml*. In this setup, we used a stratified random split on the patient-level, balanced across the four attribute groups. The data was split for training, validation (in-model batch assessment) and testing (overall performance and fairness reporting), with 80% of the data used for training, 10% for validation and 10% for testing. The patient identifiers belonging to each set were exported to separate files, accessed by the training and evaluation scripts.

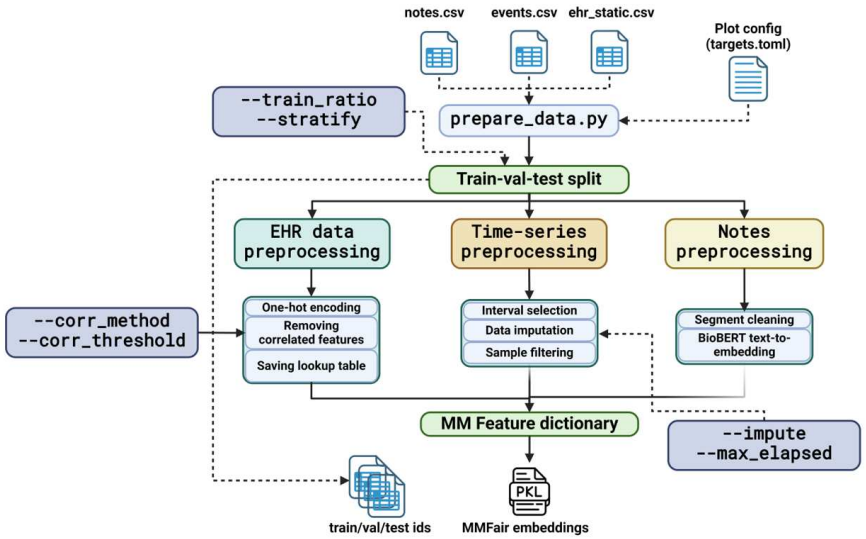

Figure 12: Overview of the **MM-HealthFair** data preprocessing pipeline per modality and split procedure for multimodal training. Figure generated with BioRender.

The tabular EHR data underwent basic feature preparation procedures, including one-hot-encoding (binarising) non-ordinal categorical variables and removing correlated features with a Pearson correlation coefficient (PCC) over 0.9. Date and text fields not suitable for prediction were saved into a separate lookup table. Missing temporal features measuring days since an event were replaced with a meaningless value '-1'.

As per the implementation in the previous project iteration by Martin et al. [10], the time-series data pipeline supports various strategies for filtering, resampling and imputation. The resampling function uses upsampling to 1-minute intervals and forward filling, before downsampling to a desired frequency with a window-wise average. Any remaining missing values are then imputed with a second forward filling step. To maintain a population with consistent measurements, we additionally restricted the population to have at least 2 measurements within the first 72 hours of admission. Due to varying lengths between the vital signs and lab tests, the input sequences were padded to the longest length across the two time-series sub-modalities, with a batch-wise procedure ensuring better information retention through dynamic lengths for each training batch.

To prepare the text features, we used token embeddings on the sentence-level, generated using the pre-trained **BioBERT** model [73]. Tokens were filtered and padded to a fixed length of 768. The Transformer-encoder network was trained using these inputs as a 2-layer encoder, with each having 8 attention-head units and a final feed-forward layer length of 256. The final set of multimodal features are saved into a binarised dictionary (.pkl file) containing the embedded features per modality and mapped to a unique patient identifier as the primary key. This file is then used for the remainder of the multimodal training and evaluation procedures.

5 Results

This chapter will overview the key findings from the multimodal fusion experiments built using the MIMIC-IV open dataset. It will use **intermediate fusion** as the primary approach, studying the effects of multimodal and unimodal combinations on performance, fairness and feature importance, for predicting hospital outcomes at point of ED attendance. Section 5.5 will additionally introduce fairness constraints through **adversarial mitigation** [72], and explore related challenges in accounting for biases in a multimodal setting.

5.1 MIMIC-IV Cohort Summary

To develop a representative cohort for risk prediction in a multimodal scenario, a number of exclusion criteria were used during linkage of the individual MIMIC-IV datasets. Due to the availability of secondary care data only, the point of prediction was the final relevant hospitalisation with recorded ED attendance within the MIMIC-IV-ED database. This allowed the collection of historical **EHR** data from previous hospitalisations as multimodal training features. The unit of analysis was based on individual-level measurements to accurately measure fairness and identify person-level variation in biases.

The exclusion criteria were minimal but required complete historical information across the tabular, time-series and free-text modalities (Figure 13). The initial cohort contained 143130 patients across **289353** confirmed admissions with prior ED attendance. Out of these individuals, 100357 had at least one discharge note with a Brief Hospital Course segment, linked via the **MIMIC-IV-Ext-BHC** dataset. Out of these patients, 36% had more than one prior hospitalisation with a complete discharge note. To bring further focus on urgent medical pathways and utilise the time-series measures, we required at least 2 recorded measurements in either the vital signs or the lab tests data within 72 hours of admission. Thus, the final extracted population consisted of **20596** unique patients, consistent across the training data over the four core outcomes.

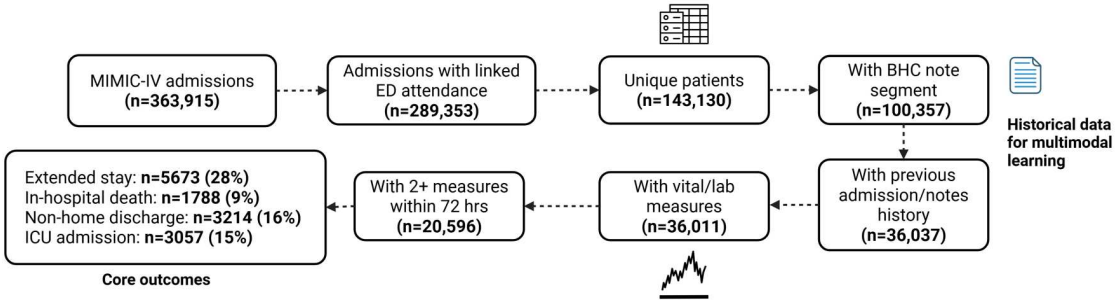

Figure 13: Overview of the **MIMIC-IV** cohort selection criteria and overall population used for the risk prediction tasks. BHC - Brief Hospital Course note segment, used as free-text training data from **MIMIC-IV-Ext-BHC**.

The final population included relatively rare hospital outcomes, reflective of a whole population approach in urgent care settings (**Table 1**). Additional statistical testing was performed via the Kruskal-Wallis H test for continuous data, and the Chi-squared test for categorical data, adjusted for multiple testing using Bonferroni correction. All listed variables were significant at $p < 0.001$ in at least one outcome group. Unsurprisingly, the cohort included an older population, as at least one prior hospitalisation was required.

There notable imbalances in the protected attributes. The cohort consisted of a predominantly white population, with Asian and Hispanic or Latino groups being in the recorded ethnic minority. As institutionalised individuals were notably older, they were also less likely to be single, and more likely to be widowed. Additionally, fewer individuals with non-home discharge were under a private health insurance, as opposed to federal health insurance. However, in-hospital death and non-home discharge were less common for Medicaid insurance types, which typically cover lower income individuals. These outcome groups were also more likely to include multimorbid patients, with over two-thirds of individuals having at least 4 long-term chronic conditions. The average number of discharge notes used for training was similar across the groups, but patients with in-hospital death had relatively longer discharge notes and BHC segments within their hospitalisation history.

5.2 Data Exploration

To understand why a model produces fairer outputs in one attribute group as opposed to another, we must also understand their interactions with other input features. It is unsurprising that underrepresented groups are likely to be favored

Table 1: Baseline characteristics and outcomes in the extracted MIMIC-IV population.

Variable	**All** (n=20596)	**IHD** (n=1788)	**EXT-ST** (n=5673)	**ICU** (n=3057)	**NHD** (n=3214)
Age, median [IQR]	64 [51,76]	70 [60,80]	66 [55,77]	66 [55,78]	74 [64,83]
Women, n (%)	10,544 (51%)	864 (48%)	2,765 (49%)	1,439 (47%)	1,652 (51%)
Ethnicity, n (%)					
Asian	814 (4%)	92 (5%)	244 (4%)	123 (4%)	121 (4%)
Black	3,499 (17%)	329 (18%)	971 (17%)	495 (16%)	479 (15%)
Hispanic/Latino	1,188 (6%)	87 (5%)	283 (5%)	128 (4%)	104 (3%)
White	14,327 (70%)	1,218 (68%)	3,962 (70%)	2,186 (72%)	2,408 (75%)
Other	768 (4%)	62 (3%)	213 (4%)	125 (4%)	102 (3%)
Marital status, n (%)					
Divorced	1,799 (9%)	153 (9%)	497 (9%)	253 (8%)	285 (9%)
Married	8,603 (42%)	770 (43%)	2,302 (41%)	1,239 (41%)	1,252 (39%)
Single	6,850 (33%)	484 (27%)	1,930 (34%)	1,012 (33%)	801 (25%)
Widowed	3,344 (16%)	381 (21%)	944 (17%)	553 (18%)	876 (27%)
Insurance, n (%)					
Medicaid	3,620 (18%)	218 (12%)	977 (17%)	520 (17%)	323 (10%)
Medicare	12,162 (59%)	1,293 (72%)	3,580 (63%)	1,944 (64%)	2,514 (78%)
Private	4,437 (22%)	251 (14%)	1,012 (18%)	525 (17%)	336 (10%)
Other	377 (2%)	26 (1%)	104 (2%)	68 (2%)	41 (1%)
Comorbidity history, n (%)					
$\geq$2 conditions	17,048 (83%)	1,682 (94%)	4,998 (88%)	2,674 (87%)	3,022 (94%)
$\geq$4 conditions	9,914 (48%)	1,216 (68%)	3,220 (57%)	1,687 (55%)	2,149 (67%)
$\geq$1 physical and $\geq$1 mental	8,092 (39%)	751 (42%)	2,371 (42%)	1,308 (43%)	1,373 (43%)
Clinical notes metadata					
# discharge notes	2 [1, 3]	2 [1, 5]	2 [1, 3]	2 [1, 3]	2 [1, 4]
# raw tokens[†]	4057 [2417, 7890]	6082 [3210, 12858]	4849 [2732, 9471]	4662 [2608, 9140]	5760 [3033, 11240]
# processed tokens[‡]	1037 [541, 2101]	1723 [860, 3530]	1318 [681, 2652]	1256 [645, 2539]	1535 [824, 3134]

IHD: In-hospital death; EXT-ST: Extended Hospital stay ($\geq$7 days); ICU: Admission to intensive care or related high-dependency unit; NHD: Non-home discharge (transfer to institution).
[†]Raw input tokens refer to the original word embedding counts from **MIMIC-IV-Note**.
[‡]Processed input tokens refer to token counts after preprocessing, and extraction of the Brief Hospital Courses, taken from **MIMIC-IV-Ext-BHC**.

less in an imbalanced classification scenario, but their decision boundary can also be affected by other measures that could point to increased health risk. Age is the most typical non-modifiable risk factor for hospital outcomes. Here, patients with a hospital outcome vary by ethnic groups, but also across their age at ED arrival (Figure 14). There are notable gaps in less-represented groups across age, such as groups coded as 'other', while the age distribution is fairly consistent across white individuals. For in-hospital deaths, those coded as 'other' are primarily older individuals, which may make the model prone to type I errors (false positives) for that group. Meanwhile, as the hispanic and latino population with extended hospital stays or ICU admissions is more likely to be younger, that might make these groups more prone to type II errors (false negatives). By knowing this, we can adjust the target loss function of our model using fairness constraints (Equations 7-10), we can customize the optimiser to target false positive rates or false negative rates, depending on the characteristics of our population.

Of course, in a multimodal scenario these constraints are more complex to enforce as different modalities can produce non-linear interactions that could affect biases inferred from more fine-grained levels of contextual information. In this case, the tokens recorded in the discharge notes could provide additional bias based on expressions that could be associated with underlying health risks. While the BHC segments have been de-identified and redacted to hide this personal information, the model could still amplify attribution biases by favoring a patient profile that is aligned with a certain ethnic group or gender. The embedding lengths could also provide useful information, as they are linked to increased visits, and typically tied to health risks (Figure 15).

After adjusting for the skewness in the BHC segment lengths, it was particularly clear that these may vary across insurance types and certain ethnicities and marital groups. Individuals under private health insurance contained significantly less expressions on average, compared to those with federal and state insurance, possibly pointing to decreased health risks. This may also be affected by socioeconomic status, which is not routinely available in MIMIC-IV. Widowed individuals may have a slightly longer hospitalisation history, possibly influenced by their average age at presentation. To fully observe these interactions, an unsupervised clustering may be required. This approach could

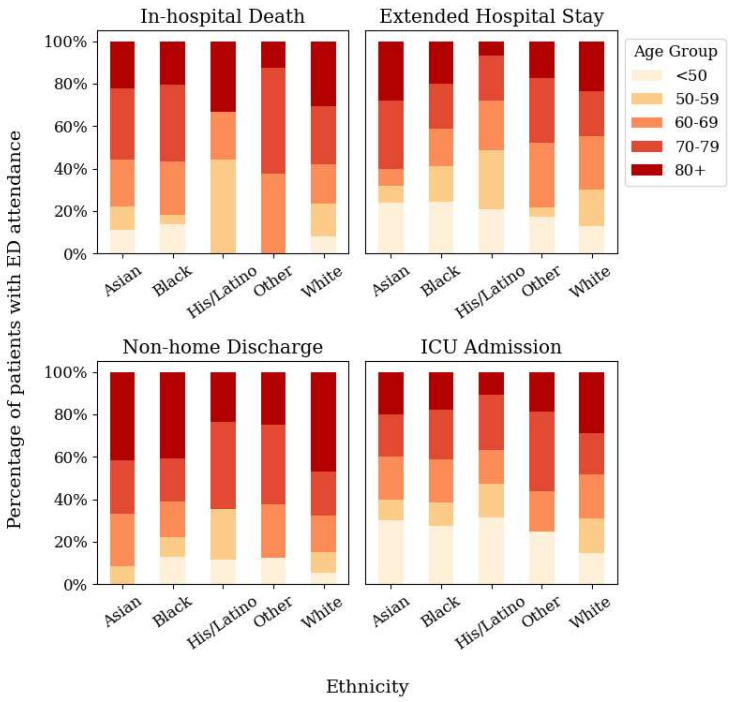

Figure 14: Age distribution across ethnic groups in patients with any of the four hospital outcomes.

leverage techniques, such as Gaussian Mixture Models (GMMs) or Hierarchical Clustering for profiling potentially bias-prone patient subgroups. However, this is beyond the scope of this framework.

5.3 Performance Summary

We evaluated performance on combinations of multimodal and unimodal learning approaches across the four outcomes. All training, validation and testing sample splits were stratified across outcome prevalence and the used sensitive attributes. The validation set (n=2060) was used for batch-wise evaluation at training time. The testing set (n=2060) was held-out for the final performance, fairness and explainability analysis. All models were consistently trained on the same input features across all modalities. All models were trained for 60 epochs, with a batch size of 256 and an early stopping patience of 10 epochs, which interrupted training once the validation loss stopped improving. The training procedure for the fully-fused models can be seen in Figure 16, showcasing the training and validation losses and the change in Receiver Operating Characterstics (ROC) and their **Area-under-the-curves (AUROC)**. During the training phase, the fully-fused model recorded the lower validation loss for extended hospital stay prediction. While the training **AUROC** scores were comparable between in-hospital death and extended hospital stay, the in-hospital death models achieved better discrimination over the validation set. Despite the stratified split, this effect did not translate over to the testing set (Table **??**), and the discrimination quality there was comparable across the two outcomes.

The intermediate fusion approach leveraging concatenation across the three modalities consistently showed the best predictive performance. Overall prognostic quality was assessed using the AUROC, with additional metrics specific for imbalanced data describing **Area under the precision-recall curve (AUPRC)**, **Sensitivity (true positive rate)**, **Specificity (true negative rate)**, **Positive Predictive Value (precision on positive class, PPV)** and **Negative Predictive Value (precision on negative class, NPV)**. Confidence intervals were produced using a weighted DeLong test, optimised for large sample sizes.

While the multimodal fusion approaches only increased discrimination ability by fine margins, the fully-fused model (IF-EHR+TS+NT) achieved the best generalisation overall across the four outcomes. In the best case scenario, the IF-EHR+TS+NT model AUROC ranged between 0.69 for non-home discharge prediction and 0.73 for in-hospital death and extended stay prediction. Detection rate for the positive class was relatively weak, as portrayed by the PPV and AUPRC scores, indicative of an imbalanced classification scenario. As portrayed by the NPV, the models also performed well in identifying the negative class. In-hospital death and extended stay models were more efficient in ruling-out patients without the outcome (Specificity), while non-home discharge and ICU admission models were more efficient in ruling-in (Sensitivity) patients that would have these outcomes at point of ED arrival. It is worth noting that

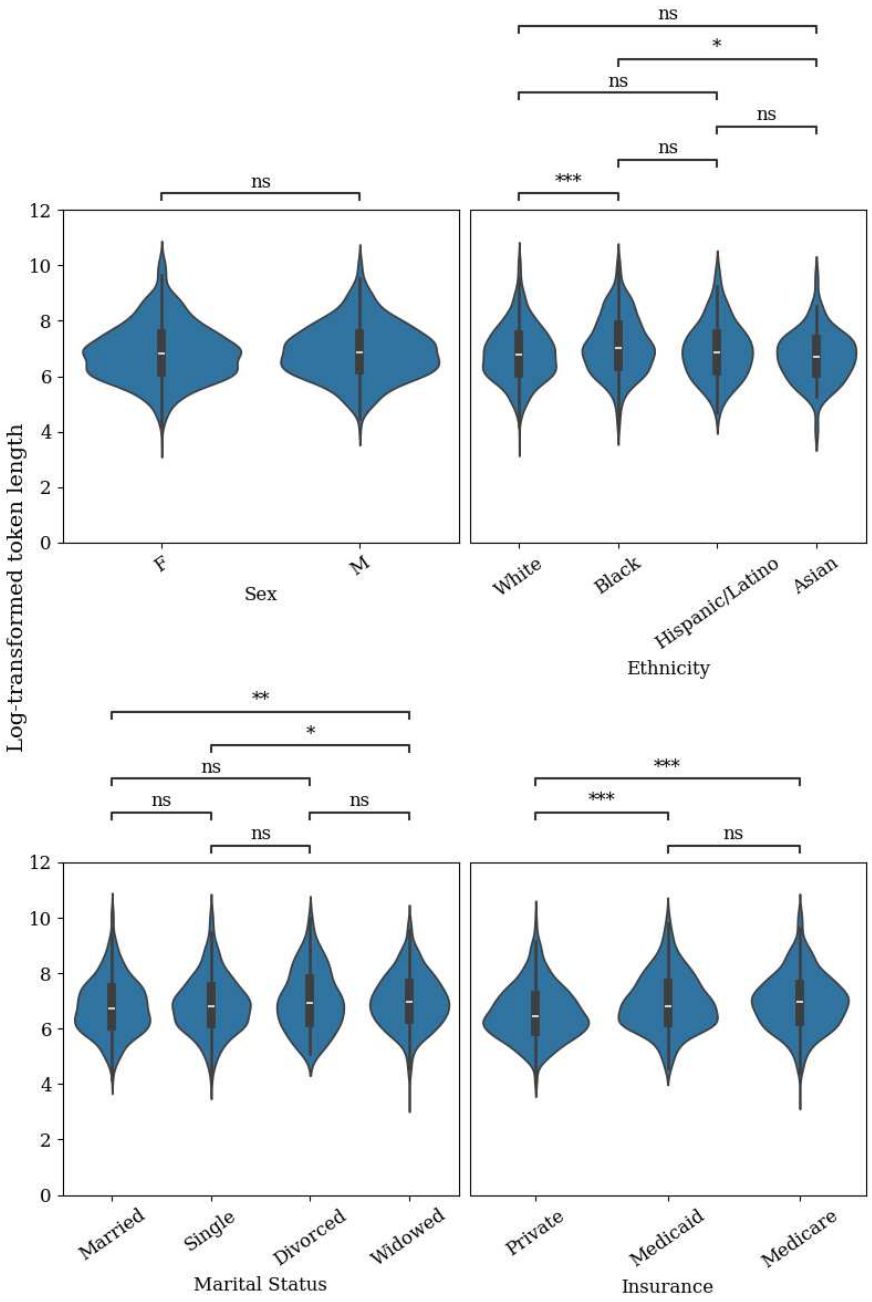

Figure 15: Spread of Brief Hospital Course word embedding lengths within the discharge notes, grouped by protected attribute. Lengths were log-transformed to account for skewness. The transformed values were tested using the Welch's t-test for unequal variances, adjusted for multiple testing using Bonferroni correction. P-value mapping - ***ns***: $p \geq 0.01$; *: $p < 0.1$; **: $p < 0.01$; ***: $p < 0.001$.

the probability thresholds for estimating these metrics were selected using the Youden's J-statistic (maximum of the sum between Sensitivity and inverse Specificity). This is treated as the optimum point of discrimination for a prognostic test. Other thresholds, such as the maximum F1-score may produce better calibrated thresholds for the positive class, in the presence of class imbalance. As the IF-EHR+TS+NT model predicting extended hospital stay achieved the best overall discrimination (via AUROC and AUPRC), it was selected for downstream analysis of fairness and explainability alongside its unimodal subtypes.

Figure 16: Summary of the multimodal training and evaluation for the fully-fused models using intermediate fusion (IF-EHR+TS+NT) over the four core outcomes. Top-left and top-right panel portray the binary cross-entropy loss over epochs, while the bottom-left and bottom-right panel highlights the change in training and validation AUROC scores. IHD: In-hospital death; EXT-ST: Extended Hospital stay ($\geq$7 days); ICU: Admission to intensive care or related high-dependency unit; NHD: Non-home discharge (transfer to institution). Plots were generated using the Weights and Biases web service.

5.4 Fairness Summary

Before evaluating the models based on fairness metrics, it might be useful to observe the spread of sensitive attributes across the probability thresholds produced by the model. This can be achieved using risk stratification by quantile-based model responses. In this experiment, the output probabilities of the outcome are sorted and discretised into ten equally-sized groups (1 - indicating least likely patients to require extended stay, 10 - indicating highest risk of extended stay). We can profile the patients recorded in the test set within each risk group, by grouping them based on their sensitive attributes. Figure 17 indicates the profiles ranked by the unimodal tabular model (**EHR-MLP**), whereas Figure 18 shows the profiles ranked by the fully-fused model (**IF-EHR+TS+NT**). Additional risk stratification patterns within the other model variants are shown in Appendix Figures A.1, A.2 and A.3.

We observe that, while the fully-fused model still contains bias patterns, it clearly achieves better parity across risk quantiles. Meanwhile, the unimodal algorithm is significantly more prone to selecting individuals as high or low-risk of extended stay based on certain background characteristics. Around 75% of individuals in the top risk decile were female, more than 90% were under Medicare insurance and a substantial amount were from non-white backgrounds. Meanwhile, <1% at the lowest risk decile were widowed, <3% were under Medicare insurance and >60% were women. Although the AUROC scores between the two models are not substantially different, the unimodal approach is clearly less ideal for ensuring health equity. In this case the multimodal approach reduces some of these biases, offering better parity across genders, ethnic groups, divorced and married individuals. However, the model is still prone to selecting widowed, non-single and patients under Medicare insurance with a higher likelihood of extended stay. Although there are likely substantial interactions with older age, this may warrant further investigation of these underlying biases.

While we can visually observe the benefits of the multimodal fusion approach in balancing patient selection rates, we still do not know whether these results are statistically significant. To effectively quantify these biases, we use the **BCa method with jackknife resampling** (Algorithm 1). We reuse this method to produce the fairness metrics and

Table 2: Model performance comparison across outcomes and modality combinations reported with 95% confidence intervals.

		Unimodal			Multimodal	
Outcome (prevalence)	**Metric**	**EHR (MLP)**	**TS (LSTM)**	**NT (TF-E)**	**IF-EHR+TS**	**IF-EHR+TS+NT**
IHD (9%)	AUROC	0.69 [.65–.72]	0.72 [.68–.75]	0.70 [.66–.74]	0.68 [.65–.72]	0.73 [.69–.76]
	AUPRC	0.16 [.02–.30]	0.17 [.01–.33]	0.17 [.01–.33]	0.16 [.01–.31]	0.17 [.01–.34]
	Sensitivity	0.77 [.70–.83]	0.75 [.68–.81]	0.64 [.56–.70]	0.70 [.63–.78]	0.56 [.48–.63]
	Specificity	0.52 [.50–.54]	0.60 [.58–.62]	0.66 [.64–.68]	0.59 [.56–.61]	0.79 [.77–.80]
	PPV	0.13 [.11–.15]	0.15 [.13–.18]	0.15 [.13–.18]	0.14 [.12–.16]	0.20 [.17–.23]
	NPV	0.96 [.95–.97]	0.96 [.95–.97]	0.95 [.94–.96]	0.95 [.94–.96]	0.95 [.94–.96]
EXT-ST (28%)	AUROC	0.70 [.67–.73]	0.67 [.64–.71]	0.70 [.67–.74]	0.69 [.65–.73]	*0.73 [.70–.77]*
	AUPRC	0.19 [.07–.32]	0.19 [.04–.34]	0.20 [.04–.35]	0.20 [.05–.36]	0.21 [.05–.37]
	Sensitivity	0.78 [.72–.83]	0.45 [.38–.52]	0.74 [.68–.80]	0.48 [.41–.55]	0.63 [.61–.66]
	Specificity	0.52 [.50–.54]	0.81 [.79–.83]	0.58 [.56–.60]	0.81 [.79–.83]	0.76 [.70–.81]
	PPV	0.16 [.13–.18]	0.21 [.17–.25]	0.17 [.15–.19]	0.23 [.19–.27]	0.19 [.17–.22]
	NPV	0.95 [.94–.97]	0.93 [.91–.94]	0.95 [.94–.96]	0.93 [.92–.94]	0.96 [.95–.97]
ICU (15%)	AUROC	0.69 [.65–.73]	0.67 [.63–.71]	0.69 [.65–.73]	0.69 [.65–.73]	0.71 [.67–.75]
	AUPRC	0.14 [.01–.30]	0.14 [.01–.30]	0.17 [.01–.33]	0.15 [.01–.31]	0.17 [.01–.33]
	Sensitivity	0.74 [.66–.80]	0.65 [.57–.72]	0.68 [.60–.75]	0.63 [.55–.70]	0.80 [.73–.85]
	Specificity	0.54 [.52–.56]	0.62 [.60–.64]	0.62 [.60–.64]	0.65 [.63–.67]	0.52 [.50–.55]
	PPV	0.12 [.10–.14]	0.13 [.10–.15]	0.13 [.11–.16]	0.13 [.11–.16]	0.12 [.11–.15]
	NPV	0.96 [.95–.97]	0.96 [.94–.97]	0.96 [.95–.97]	0.96 [.94–.97]	0.97 [.96–.98]
NHD (16%)	AUROC	0.63 [.59–.67]	0.65 [.61–.70]	0.67 [.63–.71]	0.67 [.63–.71]	0.69 [.66–.73]
	AUPRC	0.11 [.01–.27]	0.16 [.01–.31]	0.14 [.01–.30]	0.16 [.01–.32]	0.16 [.01–.32]
	Sensitivity	0.64 [.56–.71]	0.81 [.75–.86]	0.49 [.42–.57]	0.57 [.49–.64]	0.81 [.75–.86]
	Specificity	0.59 [.57–.61]	0.42 [.40–.44]	0.77 [.75–.79]	0.69 [.67–.71]	0.48 [.46–.50]
	PPV	0.12 [.10–.14]	0.11 [.09–.13]	0.16 [.13–.19]	0.14 [.11–.17]	0.12 [.10–.14]
	NPV	0.95 [.94–.96]	0.96 [.95–.97]	0.95 [.93–.96]	0.95 [.94–.96]	0.97 [.95–.98]

Outcomes: IHD - In-hospital death; EXT-ST: Extended Hospital stay ($\geq$7 days); ICU - Admission to intensive care or related high-dependency unit; NHD - Non-home discharge (transfer to institution);
Model subtypes: EHR (MLP) - Tabular modality, Multi-layer Perceptron classifier; TS (LSTM) - Time-series modality, Long Short-term Memory network; NT (TF-E) - Notes modality, Transformer-encoder network; IF-EHR+TS - Intermediate fusion with tabular and time-series networks; IF-EHR+TS+NT - Intermediate fusion over all three modalities. Best-performing model overall based on the **AUROC** and **AUPRC** scores (highlighted in italics) was selected for downstream analysis of explainability and fairness.

their confidence intervals across all algorithms for extended hospital stay prediction (Table **??**). Despite the adjusted boundaries across risk quantiles, the results for **DPR**, **EQO** and **EOP** can vary substantially, with substantial overlap especially in heavily imbalanced attribute groups, such as ethnicity and marital status.

The fully-fused model achieved the best balance across gender, but interestingly, the **EOP** and **EQO** scores are much lower on average for insurance and marital status, compared to the unimodal tabular model. This suggests that both False Positive Rates (**FPR**) and False Negative Rates (**FNR**) deviated more across these groups. We can observe that this is the case in Appendix Figures A.4 and A.5, particularly for **FNR**, where divorced patients or those under Medicaid insurance are more likely to be missed by the **IF-EHR+TS+NT** model. This indicates potential signs of amplified bias from other modalities affecting predictions for these minority groups. This bias may be induced from the time-series modality as it had the poorest **EOP** and **EQO** scores, possibly introduced by the variation in measurements for lab and vital signs testing within certain individuals. Following these findings, we will attempt to reduce these biases by incorporating **adversarial mitigation** for the **IF-EHR+TS+NT** model at training time.

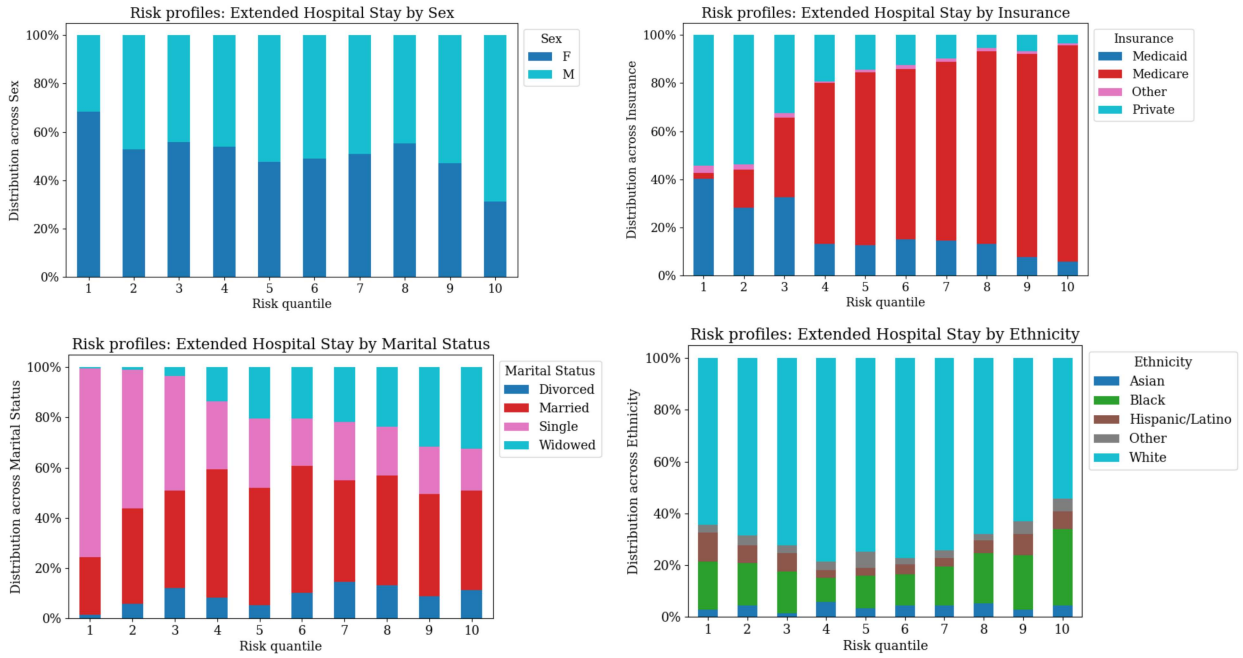

Figure 17: Risk stratification results in extended hospital stay prediction using the unimodal tabular model (**EHR-MLP**). The stacked barchart highlights the patient profiles per risk group, grouped by sensitive attribute.

5.5 Adversarial Mitigation

To perform adversarial mitigation, we include a two-layer internal adversarial classifier composed of adversarial head modules, matching the input dimensions of the **tabular** part of the **IF-EHR+TS+NT** model. We included one adversarial head module per sensitive attribute value to adjust across all possible values within the four sensitive attributes. We then used a **Gradient Reversal Layer (GRL)** [77] to maximise the objective function for the adversarial classifier (second part of Equation 7), while minimising the objective function for the main classifier. This results in the option to penalise the adversarial classifier's ability to predict the sensitive attributes from the output layer, while maintaining the discrimination utility of the classifier for detecting extended hospital stay. The effects of debiasing can then be controlled using the parameter λ, where $\lambda = 0$ indicates no debiasing (normal training procedure) and $\lambda > 0$ indicates increasingly stronger debiasing effects. The adversarial loss and the subsequent effects of debiasing on the validation set **AUROC** can be seen in Figure 19.

Here, we observe that $\lambda < 2$ is sufficient for maintaining the **AUROC** score on the validation set. Meanwhile, $\lambda = 2$ and $\lambda = 5$ introduce moderate to considerable reduction in the global classifier's discrimination ability. To test the utility of this deep adversarial debiasing approach with regard to fairness, we repeat the BCa resampling approach to generate the fairness metrics and their confidence intervals for the debiased variants (Figure 20). The full comparison between performance and fairness metrics across the different debiasing thresholds is shown in Appendix Table **??**. Interestingly, we observe that enforcing a stronger debiasing constraint did not improve the fairness metrics in most cases, with the exception of **EQO** and **EOP** in ethnic groups. In marital status groups, it also worsened the balance between **FPR** and **FNR** on average.

This may be due to a number of reasons. Firstly, as we only train the adversarial classifier to detect the attributes within the tabular modality, we do not know how the time-series and free-text modalities interact with the adjusted weights. In Table **??**, we observed that the time-series modality produces the poorest **EQO** and **EOP** scores, which may be the main source of bias in this scenario. On the other hand, the tabular modality scored higher for these fairness metrics, which may indicate limited potential for bias reduction with this approach. An alternative method could leverage Pareto frontiers (as demonstrated by **OxonFair**) [30], defining a multimodal objective for optimisation. As an in-model approach, this would be non-trivial as we would either require domain knowledge to appropriately weigh the effects on each modality, or require a way to trace non-linear interactions between coarse and fine-grained features within the deeper hidden layers of the network. Thus, a post-model approach that adjusts the output probability layer for fairness might be more realistic, similarly to calibration algorithms.

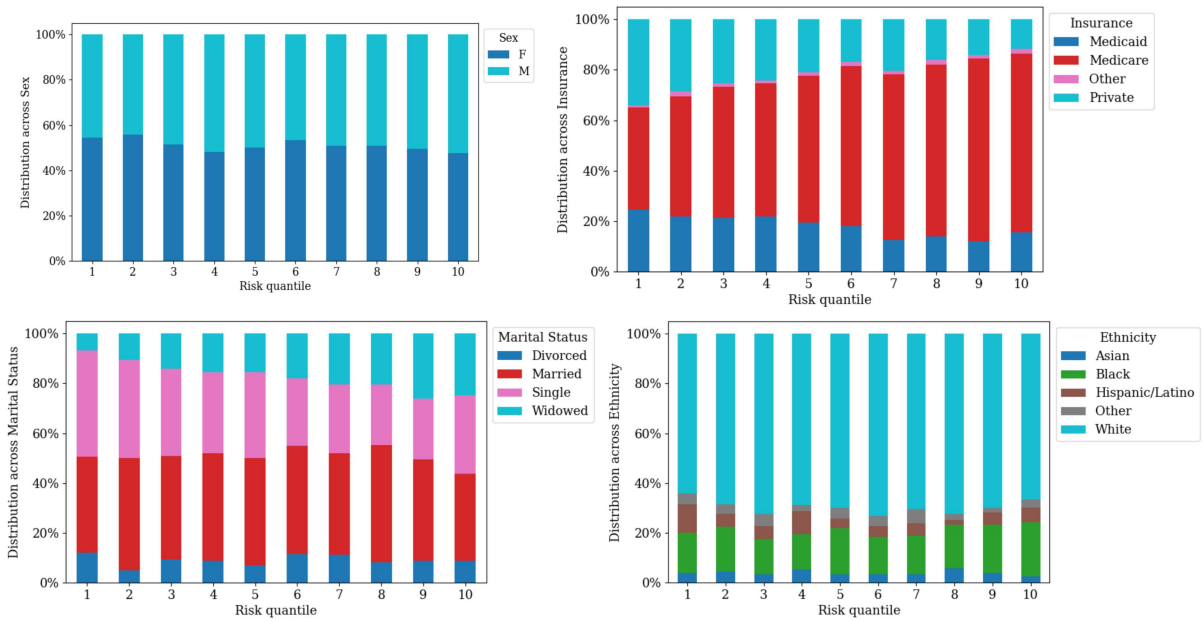

Figure 18: Risk stratification results in extended hospital stay prediction using the IF-EHR+TS+NT model. The stacked barchart highlights the patient profiles per risk group, grouped by sensitive attribute.

Table 3: Multimodal and unimodal fairness summary quantifying bias using the BCa method for bootstrapping-based 95% CI. The models were tested for prediction of extended hospital stay across 1000 bootstrap iterations using **jackknife** resampling.

Attribute	**Fairness Metric**	**Unimodal**			**Multimodal**	
		EHR (MLP)	**TS (LSTM)**	**NT (TF-E)**	**IF-EHR+TS**	**IF-EHR+TS+NT**
Gender	DPR	0.96 [.88–1]	0.89 [.77–.98]	0.96 [.87–1]	0.89 [.77–1]	0.95 [.86–1]
	EQO	0.85 [.70–.95]	0.75 [.50–.91]	0.92 [.83–.98]	0.82 [.62–.93]	0.92 [.81–.98]
	EOP	0.86 [.70–.96]	0.75 [.50–.93]	0.93 [.83–1]	0.86 [.62–.97]	0.93 [.81–1]
Insurance	DPR	0.90 [.83–.98]	0.88 [.72–.96]	0.87 [.74–.97]	0.85 [.69–.95]	0.92 [.79–.98]
	EQO	0.81 [.63–.89]	0.58 [.25–.83]	0.79 [.59–.91]	0.61 [.37–.83]	0.68 [.46–.86]
	EOP	0.82 [.65–.94]	0.59 [.27–.91]	0.82 [.57–.95]	0.62 [.37–.91]	0.69 [.48–.87]
Marital Status	DPR	0.87 [.77–.95]	0.71 [.46–.86]	0.73 [.61–.88]	0.63 [.41–.81]	0.80 [.62–.93]
	EQO	0.77 [.57–.86]	0.31 [.01–.62]	0.54 [.24–.78]	0.32 [.01–.65]	0.52 [.24–.81]
	EOP	0.79 [.57–.90]	0.31 [.01–.62]	0.54 [.29–.81]	0.32 [.01–.65]	0.53 [.24–.81]
Ethnicity	DPR	0.76 [.59–.89]	0.60 [.36–.84]	0.78 [.61–.91]	0.59 [.39–.80]	0.75 [.53–.91]
	EQO	0.55 [.01–.72]	0.16 [.01–.56]	0.53 [.01–.76]	0.31 [.01–.60]	0.52 [.04–.75]
	EOP	0.56 [.01–.75]	0.16 [.01–.58]	0.55 [.01–.83]	0.33 [.01–.68]	0.54 [.01–.80]

Fairness metrics: DPR - Demographic Parity; EQO: Equalised Odds Ratio; EOP - Equal Opportunity;
Model subtypes: EHR (MLP) - Tabular modality, Multi-layer Perceptron classifier; TS (LSTM) - Time-series modality, Long Short-term Memory network; NT (TF-E) - Notes modality, Transformer-encoder network; IF-EHR+TS - Intermediate fusion with tabular and time-series networks; IF-EHR+TS+NT - Intermediate fusion over all three modalities.

Secondly, it might be impractical to adjust across four sensitive attributes simultaneously. Although the adversarial losses and validation losses converged correctly for the **IF-EHR+TS+NT** model variants, there may be substantial limitations in finding the optimal solution. Due to the model's heterogeneous optimisation space, which aims to refine thresholds across gender, marital status, ethnicity, and insurance, an ideal solution may not be achievable. This is because there is likely significant overlap between the sensitive groups. Additionally, many of these attributes can be associated with ageing-related risk factors, which might be important to include within the model decisions. Thus, a reduced **EQO** or **EOP** cannot be treated as universal indicators of poorer health equity. To understand whether

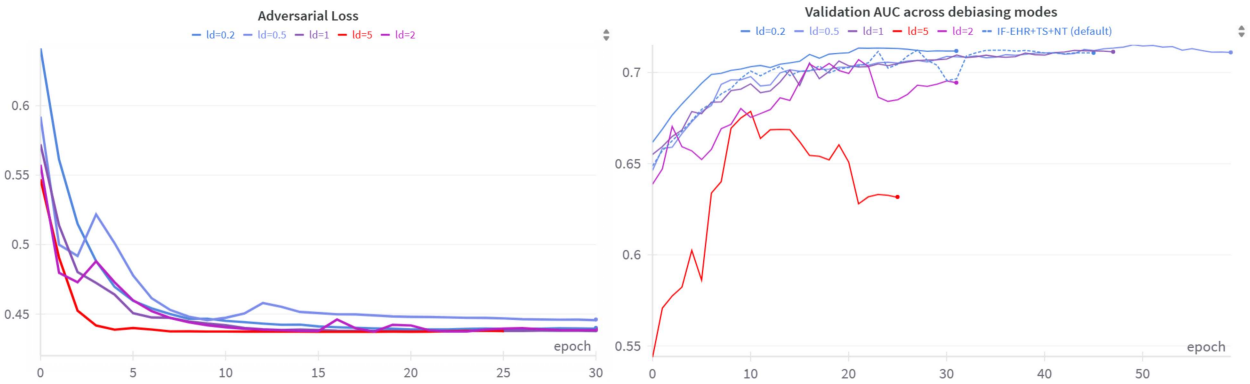

Figure 19: Debiasing model training summary with adversarial loss (left) and AUROC scores (right) using different control parameters, tested using the **IF-EHR+TS+NT** model for prediction of extended stay. $\lambda = 0.2; 0.5$ introduce a slight to moderate debiasing effect; $\lambda = 1; 2$ introduce a strong debiasing effect and $\lambda = 5$ introduces an abnormally high debiasing effect, leading to increased noise.

this behavior is equitable, we would need to further investigate and deconstruct the decision boundaries of the model, through explainable methods such as **SHAP** and **MM-SHAP**.

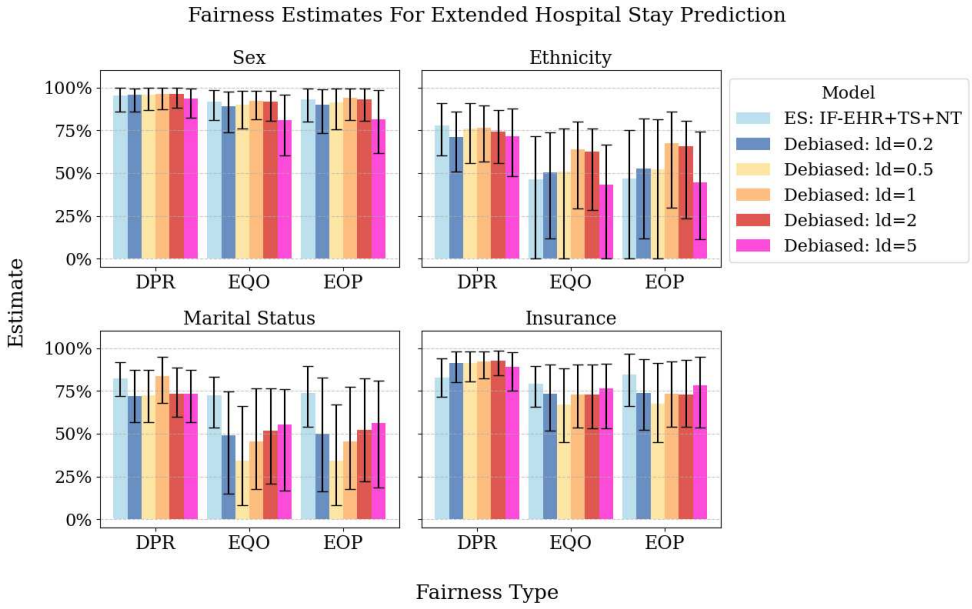

Figure 20: Fairness metrics in the fully-fused model (**IF-EHR+TS+NT**) across different levels of λ, controlling for debiasing effects targeting the four sensitive attributes in the tabular data. **Fairness metrics:** DPR - Demographic Parity; EQO: Equal Opportunity; EQO - Equalised Odds Ratio; ES: Extended Stay outcome; $\lambda = 0.2; 0.5$ introduce a slight to moderate debiasing effect; $\lambda = 1; 2$ introduce a strong debiasing effect and $\lambda = 5$ introduces an abnormally high debiasing effect (leading to increased noise).

5.6 Leveraging SHAP for Multimodal Explanations

To display the feature importance within the fully-fused model, we first execute DeepSHAP (Algorithm 2) over the held-out test set, acquiring the attribution scores across each individual network component. As the models use batch-wise computation ($n = 256$) for any training or validation procedure, we set the background samples $\mathbf{x}^b$ as the patient features within each batch. Thus, for local-level explanations, the reference (average) SHAP value will be computed using the batch belonging to the individual. To aggregate the SHAP values acquired for the time-series modality, we additionally use mean pooling over each timestep containing a valid measurement. For the global-level

plots, we rank the overall feature importance, using SHAP's *heatmap* utility: simultaneously displaying direction of importance and feature interaction. The *heatmap* plot uses a **hierarchical clustering** approach to group similar explanations together, measuring the change in the global average **SHAP** value. This results in patients that have the same model output for the same reasons mapped closer to each other within the heatmap. In addition, this also gives us a view of feature activation in a multimodal scenario, as modalities that do not affect the model decision for specific individuals will have a feature-level average **SHAP** value close to 0. As the text modality is too fine-grained for this approach, we use a standard bar plot over the letter segments to simply display the top-ranking segments and their direction of importance.

5.6.1 Global-level Explanations

We first observe the global feature importances and their distribution across patients in the test set, as shown in Figure 21. The heatmap plots suggest that the tabular data has substantial interaction effects between features, while the time-series modality has only a small portion of significant interactions that affect outcome risk. This highlights the diversity of risk interactions across different individuals in the tabular modality, which is more common in the presence of categorical data. In particular, the lab test measurements only affected the decision in $< 10\%$ of the population and the vital signs in close to 25% of individuals. Important lab tests affecting length of stay were linked to full blood count or electrolyte testing, while oxygen saturation and body temperature were indicative of some higher-risk cases among the vital signs. In the tabular modality, age was linked to an increase in the average **SHAP** risk scores with important associations across all four sensitive attribute groups. There were also relative links between long-term conditions and risk, such as arthritis reducing risk of extended stay or presence of malignant tumors increasing risk for specific individuals with similar explanations. The text barplot indicated that patients on certain interventions, medications, or specific screenings had a higher impact on reducing cumulative risk. Meanwhile, some indicators of disease severity, specialised testing and end-of-life care seemed to increase risk. These effects were not specifically tied to the sensitive attributes. However, as the tabular modality indicated important associations across all four sensitive groups, the model may hold substantial risk of amplified bias in the final decisions.

Thus, it may be useful to compare whether the adversarial debiasing was able to remedy these effects on the tabular modality (Figure 22). To perform the comparison we used the debiased variant of **IF-EHR+TS+NT** with a control parameter $\lambda = 1$, indicating a stronger fairness constraint. This was the variant enforcing the strongest constraint without limiting the **AUROC** score, as observed in Figure 19. We observe that a stronger constraint does in fact change a solid proportion of the interaction effects across **SHAP** explanations. Most prominently, the debiased variant highlights a wider region of high-risk individuals with increased age, not specifically dependent on the sensitive groups. However, we observe that the solution is not as clear-cut as this does not fully reduce the feature rankings for these attributes.

The debiased model ultimately became more dependent on white and black ethnicities, as well as on Medicaid and private insurance types. Meanwhile, it mitigated the effects of married and single marital status. Interestingly, this contradicts the fairness estimates for **EQO** and **EOP** in Figure 20, which are poorer for marital status and better for ethnicity after debiasing. This may indicate a reverse effect between feature importance ranking and health equity for these types of individuals. Thus, limiting non-linear interaction effects between these attributes might counter-intuitively result in harmful effects on these patients. However, more detailed analyses are required to capture the interaction effects introduced by the other modalities.

5.6.2 Local-level Explanations

MM-SHAP provides a more fine-grained mechanism for detecting biases within sensitive groups, as we can leverage **SHAP** aggregates to validate decision influence on the modality level. To test the effect of debiasing on the **MM-SHAP** scores, we randomly sample an individual assigned to the top decile of risk. In addition, we restrict the sample to target some of the sensitive attributes affecting the debiased model variant in Figure 22, setting an example patient profile as: woman, black ethnicity, married and under medicare insurance, selected from the top decile of risk. This allowed us to explore whether the adjusted **SHAP** rankings significantly affected the model decision for the individual after multimodal fusion. We use SHAP's *decision plot* tool to highlight the changes in the decision boundaries between tabular and time-series modalities, and we individually highlight the importance of the note segments. The SHAP reference values are estimated using the batch-wise means across the data batch belonging to the individual, with a sample size $n = 256$. We then aggregate the **SHAP** scores, as per Equations 14 and 15, computing the **MM-SHAP** dependence scores for each modality. The decision plots along with **MM-SHAP** degrees of dependence can be seen in Figure 23 (using the original fully-fused model) and Figure 24 (using the debiased model with $\lambda = 1$) for a high-risk case in the test set. Meanwhile, Figure A.6 and A.7 show the multimodal decision summary for the same profile, sampled from the lowest decile of risk.

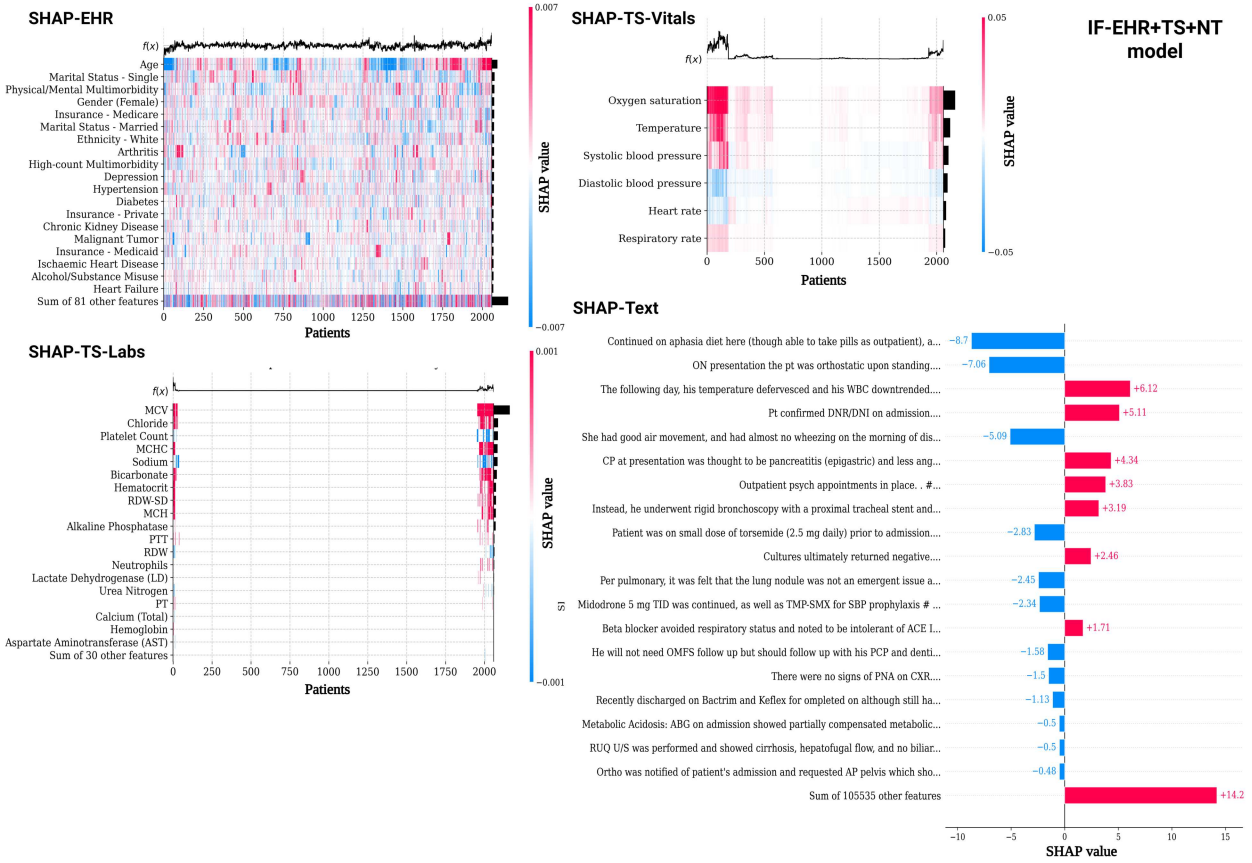

Figure 21: Global SHAP feature importance within each individual data modality for the fully-fused model (**IF-EHR+TS+NT**) predicting extended stay. **SHAP-EHR** and **SHAP-TS** represent heatmap plots that highlight clustered individuals based on explanation similarity, alongside their attribution score indicating risk. $f(x)$ represents the change in the average **SHAP** score across features per individual. **SHAP-TEXT** represents a **SHAP** bar plot of the top-ranking **BHC** segments across any individual in the test set. Blue bars indicate higher association with the negative class (no extended stay), red bars indicate higher association with the positive class (extended stay), while no bars indicates no interaction effect on outcome risk.

Unsurprisingly, the **MM-SHAP** scores suggest that the free-text modality contributed to most of the model decision for this individual (Figure 23). This was likely due to their detailed hospitalisation history containing 3 prior hospital episodes. In addition, the **Transformer-encoder** model might be more robust in detecting patterns linked to risk, compared to the **MLP** and **LSTM** classifier, which may have more limited ability to learn efficient representations. In this case, the higher-risk segments likely point to a diagnosis for acute myeloid leukemia in the latest discharge note ("10% blast count on a blood differential"). Although the discharge segment points to a number of long-term conditions, it does not directly expose any sensitive attributes other than gender. The time-series modality is affected by an elevated heart rate, high blood pressure, slightly high chloride and low platelet count within this episode. The tabular modality suggests that the patient's history of peripheral arterial disease and ischaemic heart disease is critical to the decision. However, her married status, insurance and ethnicity also seem to increase risk of extended hospital stay.

Upon validating the debiased variant on the same individual (Figure 24), the model dependence on the text modality is clearly amplified, with a 7% increase in *TX-SHAP*, a 5% reduction in *TB-SHAP* and a 3% reduction in *TS-SHAP*. The adjusted model clearly shifts the decision boundaries within the text modality, flagging many of the segments as contributors for risk reduction, while focusing on the single segment that suggests a blood cancer diagnosis. Meanwhile, the interaction effects driven by the tabular modality have almost completely diminished. Even though, some of the sensitive attributes have shifted higher in ranking, we see that their overall effects on the decision are negligible. This may be treated as a favorable outcome resulting in improved health equity. However, the debiasing algorithm may also be strongly diminishing the overall representational power that can be gathered from the tabular modality. In this instance, it removes the interaction effects imposed by peripheral arterial disease, ischaemic heart disease and

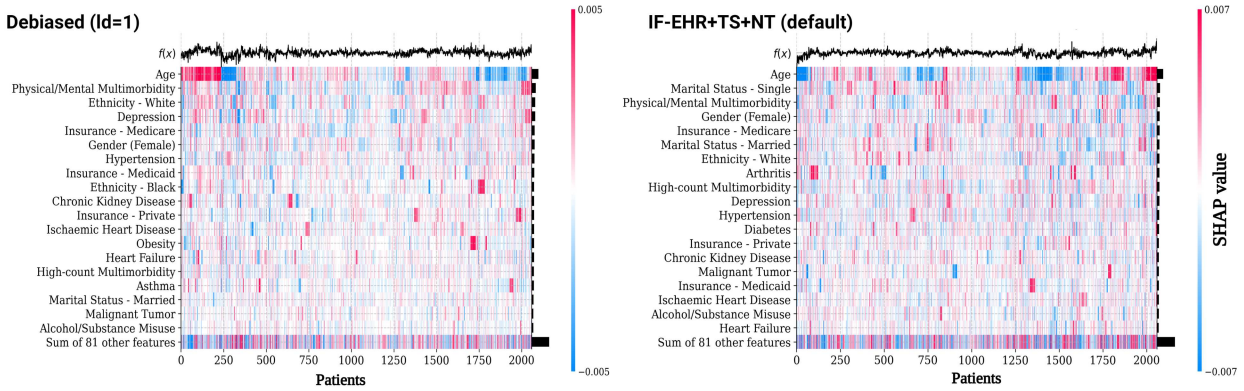

Figure 22: Global SHAP feature importance comparing the tabular data explanations between the fully-fused model (**IF-EHR+TS+NT**) and its debiased variant (stronger debiasing constraint $\lambda = 1$). The plots represent heatmap plots that highlight clustered individuals based on explanation similarity, alongside their attribution score indicating risk. $f(x)$ represents the change in the average **SHAP** score across features per individual.

multimorbidity. This may be detrimental to the application of this model if these patterns cannot be inferred from the text data. Despite this, it likely means that we can fine-tune the control parameter λ to balance between bias induction for sensitive groups and interaction effects for important risk factors.

5.7 Summary

The **MM-HealthFair** framework was developed to provide a clear evaluation of multimodal fairness and explainability, focusing on understanding driving factors behind MMAI decisions in healthcare. It leverages the **MIMIC-IV** open database for risk prediction in urgent care, supporting use of tabular, time-series and free-text data. The fairness toolkit provides a statistically-grounded validation of established fairness metrics for risk prediction, and an in-model adversarial mitigation mechanism to correct for biases in underrepresented groups. It additionally provides aggregates of **Shapley Additive eXplanations (SHAP)** estimates for interpreting decisions in a multimodal scenario and determining relative modality dependence.

This case study showcases a number of novel elements, contributing to the existing knowledge around quantifying and mitigating biases in **MMAI** algorithms. We provide a reproducible multimodal validation pipeline across four hospital outcomes, leveraging the available routine secondary care data from the **MIMIC-IV** dataset. We then provide a unique comparison between multimodal and unimodal prediction algorithms, highlighting the effects on patient selection rates after risk stratification. To measure fairness, we use the **BCa** approach to provide confidence intervals on **DPR**, **EOP** and **EQO**, useful for quantifying level of bias between sensitive groups and its significance. The adversarial mitigation then provides a way to control for these biases at training time using inputs from the tabular modality. Finally, by computing the **MM-SHAP** aggregated scores, we are able to trace the change in modality dependence before and after mitigation. This allows users to define their own fairness optimisation objective and conduct a detailed exploration on how it impacts multimodal decision boundaries and changes in feature importance.

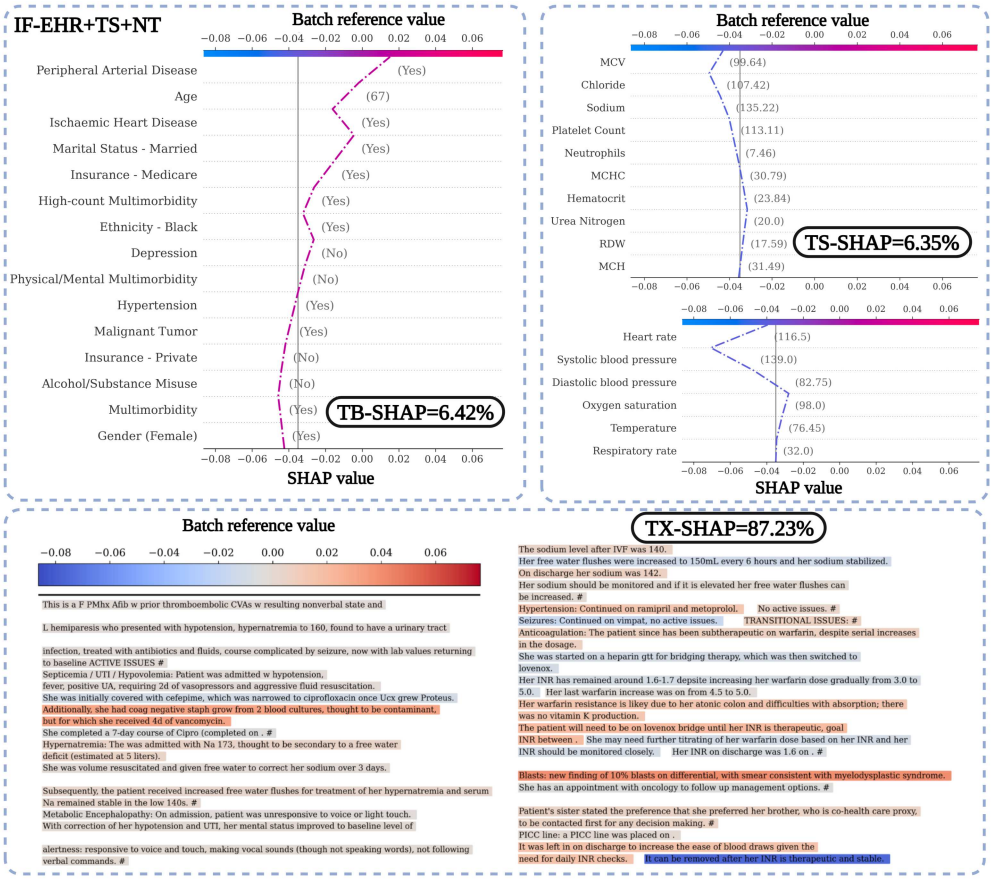

Figure 23: MM-SHAP feature importance summary for a randomly sampled individual on the test set, using the **IF-EHR+TS+NT** model for prediction of extended stay. Top-left: **SHAP** decision plot for the tabular modality and the *TB-SHAP* percentage of dependence; Top-right: **SHAP** decision plots for the time-series modality (vitals and lab tests) and the *TS-SHAP* percentage of dependence; Bottom-left and right: First and last segments from the **SHAP** text plots for the free-text modality (discharge note) and the *TX-SHAP* percentage of dependence;

6 Discussion

6.1 Strengths and Limitations

The **MM-HealthFair** framework provides a number of key utilities for examining bias patterns in risk prediction algorithms, induced by routine healthcare data. We introduced measures of group fairness, coupled with risk stratification and a bootstrapping mechanism (BCa) [68], to quantify fairness in the presence of uncertainty. Incorporating risk stratification allowed us to observe the change in distributions of demographic parity by multimodal probability thresholds. This functionality can allow stakeholders and clinicians to observe global healthcare disparities in risk predictions over heterogeneous populations, as well as track how these disparities are affected by fusion across different data modalities. The algorithm for **deep adversarial mitigation** [72] provides an in-model approach for adjusting the influence of sensitive attributes present in the routine data. This allows for the creation of custom multi-objective constraints for debiasing within the multimodal learning pipeline to reduce model dependence on specific sensitive attributes. As thresholds for performance and fairness highly depend on the clinical context and use cases of the risk prediction tool, we provide the utility to fine-tune the control parameter affecting the MMAI decision boundaries until a sufficient **performance-fairness** tradeoff is achieved. We believe that this could be useful in driving more robust reporting in research around risk prediction using MMAI, which could subsequently drive AI guidelines around 'safe' thresholds in these tools, ensuring both precision and health equity. We additionally used **SHAP** [7] and **MM-SHAP** [40] to examine modality-specific feature importance and generate individual-level decision summaries. Our toolkit provides a novel multimodal explanation wrapper using **DeepSHAP** across three deep neural network components fused at the intermediate layer. This approach makes it easy to remove and add deep neural components to support

Figure 24: MM-SHAP feature importance summary for the same randomly sampled individual on the test set, using the debiased **IF-EHR+TS+NT** variant with a control $\lambda = 1$ for prediction of extended stay. Top-left: **SHAP** decision plot for the tabular modality and the *TB-SHAP* percentage of dependence; Top-right: **SHAP** decision plots for the time-series modality (vitals and lab tests) and the *TS-SHAP* percentage of dependence; Bottom-left and right: First and last segments from the **SHAP** text plots for the free-text modality (discharge note) and the *TX-SHAP* percentage of dependence;

different experimental setups and risk explanations across other modalities. Patient risk profiling with **MM-SHAP** (Figures 22 and 23) additionally provides ways to investigate failure modes in potentially vulnerable individuals. This is crucial for ensuring trust and transparency in **MMAI** decisions and an important element of clinical decision-support.

However, it is also worth noting a few limitations that warrant attention. The deep adversarial debiasing mechanism inserts a simple optimisation objective within the multimodal pipeline for enforcing fairness constraints at training time. However, it currently relies on the target attributes for optimisation to be recorded within the tabular dataset. As we are unable to adjust the gradient layer within the fused representation directly, the model can theoretically still infer biases from other data modalities. Furthermore, the task of ensuring **multi-objective** fairness is difficult because most underrepresented individuals will have significantly overlapping attributes, persisting in multiple minority groups. By having all individual values as optimisation targets, this drastically reduces the chance of good convergence, as there will be many conflicting trade-offs between time-steps. For example, in Figure 22 the debiased model reduced dependence on marital status, but amplified the importance of ethnic groups and insurances. Thus, optimising for one attribute can result in harmful effects in another, which might not be ideal based on the clinical context. It is also worth noting that group fairness measures often produce conflicting tradeoffs as well (e.g. DPR vs EQO) [78]. For example, in critical interventions like ICU transfer, equalised odds might be more useful to prioritise accurate identification of at-risk patients. Meanwhile, for the detection of rare diseases, if disease prevalence across groups is uncertain or influenced by biased data, demographic parity could avoid perpetuating underdiagnosis. Thus, the application of these tools does require clinical insight as to selecting the appropriate metric for the task. Moreover, neither **DPR** nor **EQO** can guarantee health equity, when it is beyond the bounds of statistical parity. This is particularly important when

optimising for attributes that can change over time (e.g. married to widowed). In these scenarios, correlation might be insufficient, requiring additional analysis of causal effects.

In Figures 23 and 24, A.6 and A.7, we can also observe some potential constraints within the **MM-SHAP** scoring mechanism. As the textual inputs here had the lowest data granularity, they tend to be highly influential on the model decision, even when there are fewer embeddings present in the samples. This is partially due to the data preprocessing choices, as the notes modality includes the full patient hospitalisation history, while the time-series modality only includes the previous hospitalisation, with measurements within the first 72 hours of admission. This shows that **MM-SHAP** scores are sensitive to data preprocessing and augmentation choices. While these choices were made to avoid the computational overhead that these modalities introduce, future approaches could aim to balance these representations across modalities to ensure unbiased attribution scoring. Another constraint is **DeepSHAP**'s dependence on background (reference) values to define the baseline risk threshold. Here, we used the batch-wise average SHAP values in each modality independently to compute the background values. However, this can be influenced by the samples in the test set, the chosen batch size and the randomised selection of samples within the batch. Unfortunately, other model-agnostic approaches, such as **LIME** [35] and **PFI** [79], suffer from the same constraints and integrate less well with multimodal risk predictors. It is difficult to determine which background samples would be a representation of the true risk baseline. Future efforts could look into methods for stratified sampling of Shapley values, to ensure the background samples are based on cohorts that capture underrepresented individuals.

6.2 Clinical Relevance

The **MM-HealthFair** framework has good implications and a number of utilities that could be beneficial for bias assessment in **MMAI** algorithms. Given the UK's diverse patient population and equity-focused healthcare mandate, there are a number of potential applications of this framework within NHS trusts and related health services. In terms of mitigating healthcare disparities, this tool can help prevent misdiagnosis and underdiagnosis in AI algorithms for deprived areas or minority groups. Incorporating this pipeline into a live data feed could help automatically adjust the **MMAI** decision boundaries based on regional prevalence differences. This could enable the customisation of fairness constraints and target metrics, allowing the risk prediction algorithms to recalibrate based on target performance and equity criteria. The **SHAP** and **MM-SHAP** explainability components could then be used to generate patient audit trails. This could be a clinician-in-the-loop procedure, where signs of bias affecting the multimodal contribution scores are fed back to the model to readjust its fairness constraints. These pipelines can then position NHS systems to provide more robust guidelines around equitable AI deployment with validated thresholds to meet fairness requirements around health equity and opportunity.

The ability to reliably quantify biases with measures of group fairness in **MMAI** algorithms is crucial from a public health perspective. This allows stakeholders and officials to understand global healthcare disparities within these algorithms that are induced by routine healthcare data. More research incorporating these tools can serve as a call for action to prioritise data collection efforts and resource allocation towards underrepresented groups that have an evident disadvantage in being selected as high-risk. NHS Trusts and strategic partners can then begin working on follow-up protocols for recommendations regarding assessment of MMAI-driven tools. This could ensure substantial progress towards applications of equitable triage and resource allocation in these tools. At the same time, this would improve data quality protocols, driving robust analytical pipelines around recommendations for fairness-performance tradeoffs and thresholds for clinical use. This is also in line with bias mitigation mechanisms, useful for driving AI guidelines around bias control and acceptable thresholds for diversity and precision. These tools can allow stakeholders to refine NHS AI guidelines to mandate subgroup performance monitoring for specific populations. These could necessitate model recalibration if disparities are evidenced in previous work.

In addition, providing explainability mechanisms in **MMAI** applications can have important implications for understanding and deconstructing the complex interactions between data modalities. This brings an important level of transparency and trust in **MMAI** decisions, but also provides a toolkit for pinpointing **failure modes** in decision-support systems. The NHS employs a variety of risk prediction systems to anticipate adverse health events and target preventive interventions [80] [81]. These tools combine routine clinical data and demographic risk factors to identify high-risk patients. These are all scenarios that are prone to being influenced by ongoing healthcare disparities, and it is very challenging to trace these on the individual-level. The **MM-HealthFair** framework and its follow-up iterations can aid in clarifying which modality tipped the risk score above/under the acceptance criteria. If the system consistently underestimates risk in patients with missing measurements, it can also flag these cases as 'failure modes', recommending alternative assessment or further data collection. This enables more transparent auditing of multimodal behaviour. We believe this would enhance the utilities of decision-support interfaces and their capabilities in examining risk direction and allowing clinicians to consider the possible harmful effects of disparities.

6.3 Future Work

As research around Explainable AI methods and bias mitigation continues to evolve, this will likely lead to new avenues for project iteration. The current framework is still in its early experimentation phase and provides a baseline end-to-end pipeline that can be modified in various areas. These may include:

- **Data Curation and Extraction**: incorporating imaging and other modalities, adding support for prediction on community-wide cohorts;
- **Data Preprocessing**: modality-specific feature selection and engineering, clustering to identify underrepresented patient profiles;
- **Multimodal Learning**: refining the individual network components, incorporating other fusion mechanisms;
- **Quantifying bias and fairness**: causal approaches for quantifying bias (e.g. counterfactual examples), recalibrating predictions based on unfair causal effects;
- **Explainability**: validation of **MM-SHAP** attribution scoring, refining **MM-SHAP** to correct for differences in data granularity and measurement points;

Regarding data curation, secondary validation on another multimodal dataset may also expose the framework to unseen biases. One alternative is the recently released **INSPECT** dataset, which introduces **CTPA** imaging as a potential fourth modality [45]. **INSPECT** further provides the ability to perform diagnostic risk predictions which may be better suited for testing applications for clinical trial recruitment. However, it is worth noting that the population will also be more restricted to individuals with these measurements. Therefore, it might also be worth testing this framework on community-wide cohorts that cover individuals from deprived areas and limited access to healthcare services. This would allow for a more robust analysis of biases from a population health perspective.

The **MM-HealthFair** framework does not currently provide robust preprocessing procedures for feature selection and engineering, common in most state-of-the-art approaches. It is important to consider which preprocessing techniques could be worth adding to a specific modality, without limiting the general purpose of the framework and its ability to provide multimodal explanations. Model-based approaches, such as Gaussian Mixture Models and Latent Class Analysis are popular in healthcare research for their flexibility in modelling complex, heterogeneous populations [82]. These approaches could be applicable for downstream fairness modelling to audit the change in fairness in risk predictions, highlighting disparities across groups. Incorporating other fusion mechanisms can also provide more flexibility for testing different levels of learned contextual representations and how these may affect biases. Fusion architectures at the input layer (early fusion) and output layer (late fusion) are well-documented in **MMAI** literature [13]. Incorporating different fusion methods would also affect the level of aggregation for the **SHAP** values and may require redesign of the adversarial mitigation components of the multimodal network. In addition, testing more complex methods for fusing representations, other than feature concatenation could also be used to test robustness to biases. An approach adopted by Martin et al. in the previous iteration of the project used multi-adaptation gates (MAGs), adding in components for cross-modal attention [59].

One of the main identified challenges in **adversarial debiasing** was the ability to define a suitable **multi-objective** optimisation task within a **multimodal** scenario. There are a few ways this could be handled using emerging approaches for unsupervised learning. Pre-model approaches, such as **multi-view** clustering could serve as an efficient method for cross-modal clustering to infer suitable characteristic in underrepresented groups [83]. This approach could be designed as a label learning task targeting the categorisation of underrepresented patient identities in a single categorical variable. This could then be fed into the fairness objective function to ensure balanced mitigation across individuals and efficient model convergence. An alternative post-model approach would be fairness-specific recalibration, using the final probability layer of the fusion model to adjust and align outputs based on fairness criteria [84]. This could serve as an efficient way of aligning the model with fairness criteria without the need to retrain it with additional constraints. For example, after a model predicts the risk score, the fairness recalibration step can rescale probabilities to ensure that the likelihood of a positive prediction is consistent across groups. Such techniques could include **threshold adjustment** (setting decision thresholds to equalise outcome rates), **probability calibration** (transforming predicted probabilities, e.g. via Platt scaling variants [85] separately computed for each group) or **reweighting outputs** (assigning weights to predictions based on group membership to counteract biases).

Causal methods can potentially provide a more principled way to quantify and mitigate biases in **MMAI** algorithms. These approaches go beyond correlation-based metrics by leveraging counterfactual examples and causal graphs to identify and correct unfair influences. This is grounded in causal inference, where fairness is assessed by simulating a "counterfactual world" where only the sensitive attribute (e.g., race, gender) is altered, and observing if the model's decision changes. This is also important when accounting for sensitive attributes that may change over time, where correlation is not applicable. A recent review paper by Binkyte et al. argues that causality is key when balancing for

multiple competing fairness objectives [86]. In addition, it may be feasible to design conditional definitions of **DPR**, **EQO** and **EOP** based on the constraints of the training distribution [87]. In that sense, a model will be treated as counterfactually fair if, for every individual, the prediction is the same, provided the outcome itself does not change. This can be operationalised using generative models (e.g., variational autoencoders) to estimate counterfactuals in complex datasets. In the context of bias mitigation, counterfactual fairness frameworks could allow for bias correction that arises not just from data imbalance, but from causal pathways linking sensitive attributes to predictions.

Finally, there are many possible avenues for leveraging **SHAP** values for attribution scoring in a multimodal scenario. Here, we see that aggregating SHAP values by modality can help assess which data sources disproportionately influence predictions. A wider statistical or probabilistic analysis of the **SHAP** value distributions (e.g. KL-divergence) can allow for a multi-level comparison between contributions of sensitive groups and modalities. These analyses can also be helpful in driving new targets for fairness-aware learning. Other joint optimisation approaches (e.g. Pareto principles applied in OxonFair [30]) can also aid in seeking out solutions that can balance between performance, fairness metrics and divergence in **SHAP** contribution scores. There may be a need to adjust the SHAP scores to better account for the variability in data granularity. One way to achieve this would be a simple normalisation weight based on the proportion of embeddings in each modality. Further efforts may also be necessary to validate the variability of **SHAP** and **MM-SHAP** scores across stratified data splits.

7 Conclusion

In our experimental setup, we compared multimodal and unimodal algorithms for risk predictions using intermediate fusion across three data modalities within **MIMIC-IV**: tabular (count health record data), time-series (vitals and lab test repeated measures) and text (discharge summaries). We evaluated the model variants on four hospital outcomes in hospitalised patients with prior ED attendance: **in-hospital death**, **extended stay**, **non-home discharge** and **ICU admission**. The fully-fused algorithm only reached moderate discrimination across outcomes, but consistently outperformed single-modality approaches. While performance improved after fusion, overall effects on fairness varied. There were signs of amplified biases in the fully-fused algorithm, reducing **EQO** and **EOP** in insurance and marital status groups. This was reflected by a substantial increase in false negative rates within the model for the lowest prevalence attributes. Adversarial mitigation was then tested for rectifying these biases, enforcing a stronger constraint across all sensitive groups in the tabular data, while retaining the baseline performance. This led to an improvement in **EQO** and **EOP** for ethnicities but worsened the results for marital status groups. The **DeepSHAP** algorithm was then used to investigate modality activations, suggesting a smaller influence of the time-series data on the global-level. The debiasing mechanism effectively reduced feature importance but placed higher importance on ethnicity and insurance. **MM-SHAP** then highlighted potential over-dependence on the textual inputs after debiasing, limiting most of the contextual information provided by the tabular modality. The results highlighted the need for more refined multi-objective optimisation tasks in a multimodal scenario for risk prediction, as well as the need to define more consistent fairness representations suitable for integration with multimodal learning objectives.

The **MM-HealthFair** framework was developed to provide an extensible and reproducible framework for quantifying and mitigating biases in **MMAI** risk prediction algorithms for healthcare. Among its main components is the utility to enforce fairness constraints through deep adversarial mitigation, and examine the impact on healthcare disparities, targeting equality of selection and error rates within sensitive groups. Aggregation of multimodal explanations through **SHAP** further allows for testing the shifts in model behaviour before and after model adjustment. The current **MMAI** pipeline provides a baseline for future iteration, with possibilities to support other fusion mechanisms, modalities and deep learning mechanisms to support bias mitigation. Although it requires testing on diverse patient populations, this tool could have substantial wider implications for applications in NHS trusts and beyond. Toolkits of this nature could help establish clinical guidelines around target criteria for health equity and opportunity in AI risk prediction models, as well as novel approaches for correcting biases in **MMAI** algorithms.

References

[1] Diny Dixon, Hina Sattar, Natalia Moros, Srija Reddy Kesireddy, Huma Ahsan, Mohit Lakkimsetti, Madiha Fatima, Dhruvi Doshi, Kanwarpreet Sadhu, and Muhammad Junaid Hassan. Unveiling the Influence of AI Predictive Analytics on Patient Outcomes: A Comprehensive Narrative Review. *Cureus*, 16(5):e59954. ISSN 2168-8184. doi:10.7759/cureus.59954. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11161909/.

[2] Julián N. Acosta, Guido J. Falcone, Pranav Rajpurkar, and Eric J. Topol. Multimodal biomedical AI. *Nature Medicine*, 28(9):1773–1784, September 2022. ISSN 1546-170X. doi:10.1038/s41591-022-01981-2. URL https://www.nature.com/articles/s41591-022-01981-2. Publisher: Nature Publishing Group.

[3] Trishan Panch, Heather Mattie, and Rifat Atun. Artificial intelligence and algorithmic bias: implications for health systems. *Journal of global health*, 9(2):020318, 2019.

[4] Felix Krones, Umar Marikkar, Guy Parsons, Adam Szmul, and Adam Mahdi. Review of multimodal machine learning approaches in healthcare. *Information Fusion*, 114:102690, 2025.

[5] James L Cross, Michael A Choma, and John A Onofrey. Bias in medical ai: Implications for clinical decision-making. *PLOS Digital Health*, 3(11):e0000651, 2024.

[6] Sai Munikoti, Ian Stewart, Sameera Horawalavithana, Henry Kvinge, Tegan Emerson, Sandra E Thompson, and Karl Pazdernik. Generalist multimodal ai: A review of architectures, challenges and opportunities. *arXiv preprint arXiv:2406.05496*, 2024.

[7] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[8] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-Modal Self-Attention Network for Referring Image Segmentation, April 2019. URL http://arxiv.org/abs/1904.04745. arXiv:1904.04745 [cs].

[9] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, January 2023. ISSN 2052-4463. doi:10.1038/s41597-022-01899-x. URL https://www.nature.com/articles/s41597-022-01899-x. Publisher: Nature Publishing Group.

[10] Sophie Martin and Jonathan Pearson. UNDERSTANDING FAIRNESS AND EXPLAINABILITY IN MULTI-MODAL APPROACHES WITHIN HEALTHCARE.

[11] Shakti N Wadekar, Abhishek Chaurasia, Aman Chadha, and Eugenio Culurciello. The evolution of multimodal model architectures. *arXiv preprint arXiv:2405.17927*, 2024.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[13] Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P. Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digital Medicine*, 3(1):1–9, October 2020. ISSN 2398-6352. doi:10.1038/s41746-020-00341-z. URL https://www.nature.com/articles/s41746-020-00341-z. Publisher: Nature Publishing Group.

[14] Yihao Li, Mostafa El Habib Daho, Pierre-Henri Conze, Rachid Zeghlache, Hugo Le Boité, Ramin Tadayoni, Béatrice Cochener, Mathieu Lamard, and Gwenolé Quellec. A review of deep learning-based information fusion techniques for multimodal medical image classification. *Computers in Biology and Medicine*, page 108635, 2024. Publisher: Elsevier.

[15] Ming Hou, Jiajia Tang, Jianhai Zhang, Wanzeng Kong, and Qibin Zhao. Deep multimodal multilinear fusion with high-order polynomial pooling. *Advances in Neural Information Processing Systems*, 32, 2019.

[16] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

[17] Janani Venugopalan, Li Tong, Hamid Reza Hassanzadeh, and May D Wang. Multimodal deep learning models for early detection of alzheimer's disease stage. *Scientific reports*, 11(1):3254, 2021.

[18] Pitiwat Lueangwitchajaroen, Sitapa Watcharapinchai, Worawit Tepsan, and Sorn Sooksatra. Multi-level feature fusion in cnn-based human action recognition: A case study on efficientnet-b7. *Journal of Imaging*, 10(12):320, 2024.

[19] Yao Ma, Shilin Zhao, Weixiao Wang, Yaoman Li, and Irwin King. Multimodality in meta-learning: A comprehensive survey. *Knowledge-Based Systems*, 250:108976, 2022.

[20] Luis Manuel Pereira, Addisson Salazar, and Luis Vergara. On comparing early and late fusion methods. In *International Work-Conference on Artificial Neural Networks*, pages 365–378. Springer, 2023.

[21] Tianzhe Jiao, Chaopeng Guo, Xiaoyue Feng, Yuming Chen, and Jie Song. A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications. *Computers, Materials & Continua*, 80(1), 2024.

[22] Simon Dietz, Thomas Altstidl, Dario Zanca, Björn Eskofier, and An Nguyen. How intermodal interaction affects the performance of deep multimodal fusion for mixed-type time series. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2024.

[23] Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd international conference on information fusion (FUSION)*, pages 1–6. IEEE, 2020.

[24] Daiju Ueda, Taichi Kakinuma, Shohei Fujita, Koji Kamagata, Yasutaka Fushimi, Rintaro Ito, Yusuke Matsui, Taiki Nozaki, Takeshi Nakaura, Noriyuki Fujima, et al. Fairness of artificial intelligence in healthcare: review and recommendations. *Japanese Journal of Radiology*, 42(1):3–15, 2024.

[25] Mingxuan Liu, Yilin Ning, Salinelat Teixayavong, Mayli Mertens, Jie Xu, Daniel Shu Wei Ting, Lionel Tim-Ee Cheng, Jasmine Chiat Ling Ong, Zhen Ling Teo, Ting Fang Tan, et al. A translational perspective towards clinical ai fairness. *NPJ Digital Medicine*, 6(1):172, 2023.

[26] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[27] Joshua W Anderson and Shyam Visweswaran. Algorithmic individual fairness and healthcare: a scoping review. *JAMIA open*, 8(1):ooae149, 2025.

[28] Tosin Adewumi, Lama Alkhaled, Namrata Gurung, Goya van Boven, and Irene Pagliai. Fairness and bias in multimodal ai: A survey. *arXiv preprint arXiv:2406.19097*, 2024.

[29] Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and Improving Fairness of AI Systems, 2023. URL `http://jmlr.org/papers/v24/23-0389.html`. Publication Title: Journal of Machine Learning Research Volume: 24 original-date: 2018-05-15T01:51:35Z.

[30] Eoin Delaney, Zihao Fu, Sandra Wachter, Brent Mittelstadt, and Chris Russell. Oxonfair: A flexible toolkit for algorithmic fairness. *arXiv preprint arXiv:2407.13710*, 2024.

[31] Michael Lohaus, Michael Perrot, and Ulrike Von Luxburg. Too relaxed to be fair. In *International Conference on Machine Learning*, pages 6360–6369. PMLR, 2020.

[32] Nitin Rane, Saurabh Choudhary, and Jayesh Rane. Explainable Artificial Intelligence (XAI) in healthcare: Interpretable Models for Clinical Decision Support, November 2023. URL `https://papers.ssrn.com/abstract=4637897`.

[33] Timo Speith. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2239–2250, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi:10.1145/3531146.3534639. URL `https://dl.acm.org/doi/10.1145/3531146.3534639`.

[34] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[35] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[36] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, October 2017. doi:10.1109/ICCV.2017.74. URL `https://ieeexplore.ieee.org/document/8237336`. ISSN: 2380-7504.

[37] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR, 2017.

[38] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[39] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[40] Letitia Parcalabescu and Anette Frank. MM-SHAP: A Performance-agnostic Metric for Measuring Multimodal Contributions in Vision and Language Models & Tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4032–4059, 2023. doi:10.18653/v1/2023.acl-long.223. URL http://arxiv.org/abs/2212.08158. arXiv:2212.08158 [cs].

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

[42] Yu-Hsuan Li, I-Te Lee, Yu-Wei Chen, Yow-Kuan Lin, Yu-Hsin Liu, and Fei-Pei Lai. Using Text Content From Coronary Catheterization Reports to Predict 5-Year Mortality Among Patients Undergoing Coronary Angiography: A Deep Learning Approach. *Frontiers in Cardiovascular Medicine*, 9, February 2022. doi:10.3389/fcvm.2022.800864.

[43] Colton Ladbury, Reza Zarinshenas, Hemal Semwal, Andrew Tam, Nagarajan Vaidehi, Andrei Rodin, An Liu, Scott Glaser, Ravi Salgia, and Arya Amini. Utilization of model-agnostic explainable artificial intelligence frameworks in oncology: a narrative review. *Translational Cancer Research*, 11, 01 2021. doi:10.21037/tcr-22-1626.

[44] Luis R. Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M. Wiberg, Michael L. Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *npj Digital Medicine*, 5(1):1–10, September 2022. ISSN 2398-6352. doi:10.1038/s41746-022-00689-4. URL https://www.nature.com/articles/s41746-022-00689-4. Publisher: Nature Publishing Group.

[45] Shih-Cheng Huang, Zepeng Huo, Ethan Steinberg, Chia-Chun Chiang, Matthew P. Lungren, Curtis P. Langlotz, Serena Yeung, Nigam H. Shah, and Jason A. Fries. INSPECT: A Multimodal Dataset for Pulmonary Embolism Diagnosis and Prognosis, November 2023. URL http://arxiv.org/abs/2311.10798. arXiv:2311.10798 [cs].

[46] Inyoung Jun, Sarah E Ser, Scott A Cohen, Jie Xu, Robert J Lucero, Jiang Bian, and Mattia Prosperi. Quantifying health outcome disparity in invasive methicillin-resistant staphylococcus aureus infection using fairness algorithms on real-world data. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024*, pages 419–432. World Scientific, 2023.

[47] Ran Zhou, Yang Liu, Wei Xia, Yu Guo, Zhongwei Huang, Haitao Gan, and Aaron Fenster. Jocorank: joint correlation learning with ranking similarity regularization for imbalanced fetal brain age regression. *Computers in Biology and Medicine*, 171:108111, 2024.

[48] Aurélie Pahud de Mortanges, Haozhe Luo, Shelley Zixin Shu, Amith Kamath, Yannick Suter, Mohamed Shelan, Alexander Pöllinger, and Mauricio Reyes. Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging. *NPJ digital medicine*, 7(1):195, 2024.

[49] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36 (4):1234–1240, 2020.

[50] Anna Meldo, Lev Utkin, Maxim Kovalev, and Ernest Kasimov. The natural language explanation algorithms for the lung cancer computer-aided diagnosis system. *Artificial intelligence in medicine*, 108:101952, 2020.

[51] Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6):e200267, 2021.

[52] Atul Anand, Tushar Kadian, Manu Kumar Shetty, and Anubha Gupta. Explainable ai decision model for ecg data of cardiac disorders. *Biomedical Signal Processing and Control*, 75:103584, 2022.

[53] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[54] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[55] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[56] Asad Aali, Dave Van Veen, Yamin Arefeen, Jason Hom, Christian Bluethgen, Eduardo Pontes Reis, Sergios Gatidis, Namuun Clifford, Joseph Daws, Arash Tehrani, Jangwon Kim, and Akshay Chaudhari. MIMIC-IV-Ext-BHC: Labeled Clinical Notes Dataset for Hospital Course Summarization. URL https://physionet.org/content/labelled-notes-hospital-course/1.2.0/.

[57] Yunpeng Huang, Jingwei Xu, Junyu Lai, Zixu Jiang, Taolue Chen, Zenan Li, Yuan Yao, Xiaoxing Ma, Lijuan Yang, Hao Chen, et al. Advancing transformer architecture in long-context large language models: A comprehensive survey. *arXiv preprint arXiv:2311.12351*, 2023.

[58] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016.

[59] Xianbing Zhao, Yixin Chen, Wanting Li, Lei Gao, and Buzhou Tang. Mag+: An extended multimodal adaptation gate for multimodal sentiment analysis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4753–4757. IEEE, 2022.

[60] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[61] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[62] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[63] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

[64] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. arxiv, 2017.

[65] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[66] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, volume 1, page 11. Barcelona, Spain, 2016.

[67] María Agustina Ricci Lara, Rodrigo Echeveste, and Enzo Ferrante. Addressing fairness in artificial intelligence for medical imaging. *Nature Communications*, 13(1):4581, August 2022. ISSN 2041-1723. doi:10.1038/s41467-022-32186-3. URL `https://www.nature.com/articles/s41467-022-32186-3`. Publisher: Nature Publishing Group.

[68] Bradley Efron. Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397): 171–185, 1987.

[69] Vinícius Litvinoff Justus, Vitor Batista Rodrigues, and Alex Rodrigo dos Santos Sousa. Bootstrap confidence intervals: A comparative simulation study. *arXiv preprint arXiv:2404.12967*, 2024.

[70] Yves G Berger. A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika*, 94(4):953–964, 2007.

[71] Erik Drysdale. Implementing the bias-corrected and accelerated bootstrap in Python. URL `https://www.erikdrysdale.com/bca_python/`.

[72] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating Unwanted Biases with Adversarial Learning, January 2018. URL `http://arxiv.org/abs/1801.07593`. arXiv:1801.07593 [cs].

[73] Emily Alsentzer and Anne Kim. Extractive summarization of ehr discharge notes. *arXiv preprint arXiv:1810.12085*, 2018.

[74] Asad Aali, Dave Van Veen, Yamin Ishraq Arefeen, Jason Hom, Christian Bluethgen, Eduardo Pontes Reis, Sergios Gatidis, Namuun Clifford, Joseph Daws, Arash S Tehrani, Jangwon Kim, and Akshay S Chaudhari. A dataset and benchmark for hospital course summarization with adapted large language models. *Journal of the American Medical Informatics Association*, page ocae312, December 2024. ISSN 1067-5027, 1527-974X. doi:10.1093/jamia/ocae312. URL `https://academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocae312/7934937`.

[75] Mehak Gupta, Brennan Gallamoza, Nicolas Cutrona, Pranjal Dhakal, Raphael Poulain, and Rahmatollah Beheshti. An Extensive Data Processing Pipeline for MIMIC-IV. *Proceedings of machine learning research*, 193:311–325, November 2022. ISSN 2640-3498. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9854277/`.

[76] Konstantin Georgiev, Joanne McPeake, Susan D Shenkin, Jacques Fleuriot, Nazir Lone, Bruce Guthrie, Julie A Jacko, and Atul Anand. Understanding hospital activity and outcomes for people with multimorbidity using electronic health records. *Scientific Reports*, 15(1):8522, 2025.

[77] Yaroslav Ganin and V Lempitsky. Unsupervised domain adaptation by backpropagation. arxiv. *arXiv preprint arXiv:1409.7495*, 2014.

[78] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):4209, 2022.

[79] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[80] Julia Hippisley-Cox, Carol Coupland, and Peter Brindle. Development and validation of qrisk3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *bmj*, 357, 2017.

[81] Ash K Clift, Carol AC Coupland, Ruth H Keogh, Karla Diaz-Ordaz, Elizabeth Williamson, Ewen M Harrison, Andrew Hayward, Harry Hemingway, Peter Horby, Nisha Mehta, et al. Living risk prediction algorithm (qcovid) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study. *bmj*, 371, 2020.

[82] Caroline X Gao, Dominic Dwyer, Ye Zhu, Catherine L Smith, Lan Du, Kate M Filia, Johanna Bayer, Jana M Menssink, Teresa Wang, Christoph Bergmeir, et al. An overview of clustering methods with guidelines for application in mental health research. *Psychiatry Research*, 327:115265, 2023.

[83] Uno Fang, Man Li, Jianxin Li, Longxiang Gao, Tao Jia, and Yanchun Zhang. A comprehensive survey on multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12350–12368, 2023.

[84] Yuzhen Mao, Zhun Deng, Huaxiu Yao, Ting Ye, Kenji Kawaguchi, and James Zou. Last-layer fairness fine-tuning is simple and effective for neural networks. *arXiv preprint arXiv:2304.03935*, 2023.

[85] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[86] Ruta Binkyte, Ivaxi Sheth, Zhijing Jin, Mohammad Havaei, Bernhard Schölkopf, and Mario Fritz. Causality is key to understand and balance multiple goals in trustworthy ml and foundation models. *arXiv preprint arXiv:2502.21123*, 2025.

[87] Jacy Anthis and Victor Veitch. Causal context connects counterfactual fairness to robust prediction and group fairness. *Advances in Neural Information Processing Systems*, 36:34122–34138, 2023.

A Appendix

A.1 Performance and Fairness analysis across different levels of adversarial mitigation

Table A.1: Performance and fairness analysis using the **IF-EHR+TS+NT** model and its debiased variants for prediction of extended stay (prevalence 28%). The control parameter λ represents the strength of the enforced fairness constraint over all four attribute groups. $\lambda = 0.2; 0.5$ introduce a slight to moderate debiasing effect; $\lambda = 1; 2$ introduce a strong debiasing effect and $\lambda = 5$ introduces an abnormally high debiasing effect (leading to increased noise).

Type	Metric	Original (IF-EHR+TS+NT)	$\lambda = 0.2$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$	$\lambda = 5$
Performance	AUROC	0.73 [.70–.77]	0.72 [.69–.76]	0.72 [.68–.75]	0.73 [.70–.76]	0.71 [.67–.74]	0.64 [.60–.68]
	AUPRC	0.21 [.05–.37]	0.21 [.06–.37]	0.20 [.04–.36]	0.21 [.05–.36]	0.21 [.06–.37]	0.18 [.04–.32]
	Sensitivity	0.63 [.61–.66]	0.64 [.62–.66]	0.64 [.62–.66]	0.59 [.57–.61]	0.56 [.54–.59]	0.53 [.46–.60]
	Specificity	0.76 [.70–.81]	0.71 [.65–.77]	0.70 [.64–.76]	0.78 [.72–.83]	0.76 [.70–.81]	0.70 [.68–.72]
	PPV	0.19 [.17–.22]	0.18 [.16–.21]	0.18 [.16–.21]	0.18 [.15–.20]	0.17 [.14–.19]	0.17 [.14–.20]
	NPV	0.96 [.95–.97]	0.95 [.94–.96]	0.95 [.94–.96]	0.96 [.95–.97]	0.95 [.94–.97]	0.93 [.91–.94]
Fairness							
Gender	DPR	0.95 [.86–1]	0.96 [.86–.99]	0.96 [.86–.99]	0.96 [.88–.99]	0.97 [.90–1]	0.93 [.84–1]
	EQO	0.92 [.81–.98]	0.89 [.71–.98]	0.90 [.74–.98]	0.92 [.82–.97]	0.92 [.81–.98]	0.81 [.57–.94]
	EOP	0.93 [.81–1]	0.90 [.71–.99]	0.91 [.73–.99]	0.93 [.82–1]	0.93 [.81–1]	0.82 [.59–.99]
Insurance	DPR	0.92 [.79–.98]	0.91 [.80–.97]	0.91 [.82–.99]	0.93 [.84–.98]	0.92 [.82–.99]	0.88 [.76–.99]
	EQO	0.68 [.46–.86]	0.72 [.51–.90]	0.67 [.44–.88]	0.73 [.55–.91]	0.72 [.50–.89]	0.76 [.58–.92]
	EOP	0.69 [.48–.87]	0.72 [.51–.95]	0.67 [.44–.91]	0.73 [.55–.94]	0.72 [.50–.90]	0.77 [.59–.94]
Marital Status	DPR	0.80 [.62–.93]	0.73 [.57–.89]	0.73 [.59–.86]	0.85 [.68–.95]	0.74 [.61–.90]	0.75 [.62–.87]
	EQO	0.52 [.24–.81]	0.47 [.23–.70]	0.32 [.09–.63]	0.43 [.18–.69]	0.50 [.21–.73]	0.55 [.24–.76]
	EOP	0.53 [.24–.81]	0.49 [.23–.88]	0.32 [.09–.63]	0.43 [.18–.69]	0.50 [.25–.78]	0.56 [.25–.82]
Ethnicity	DPR	0.75 [.53–.91]	0.71 [.53–.89]	0.77 [.61–.91]	0.77 [.61–.89]	0.74 [.60–.86]	0.72 [.43–.85]
	EQO	0.52 [.04–.75]	0.51 [.20–.75]	0.53 [.14–.80]	0.64 [.36–.78]	0.62 [.24–.74]	0.46 [.21–.65]
	EOP	0.54 [.01–.80]	0.54 [.19–.78]	0.55 [.17–.81]	0.68 [.36–.83]	0.65 [.24–.77]	0.47 [.21–.72]

Fairness metrics: DPR - Demographic Parity; EQO: Equalised Odds Ratio; EOP - Equal Opportunity;

A.2 Additional risk stratification plots

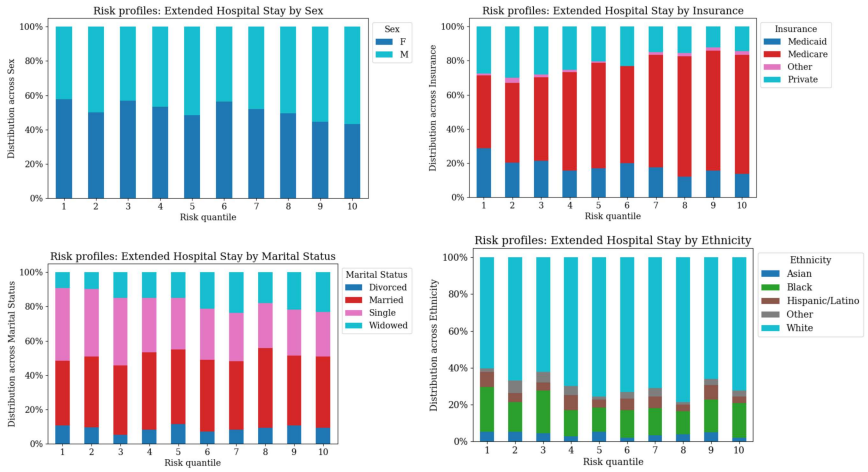

Figure A.1: Risk stratification results in extended hospital stay prediction using the unimodal time-series model (**TS-LSTM**). The stacked barchart highlights the patient profiles per risk group, grouped by sensitive attribute.

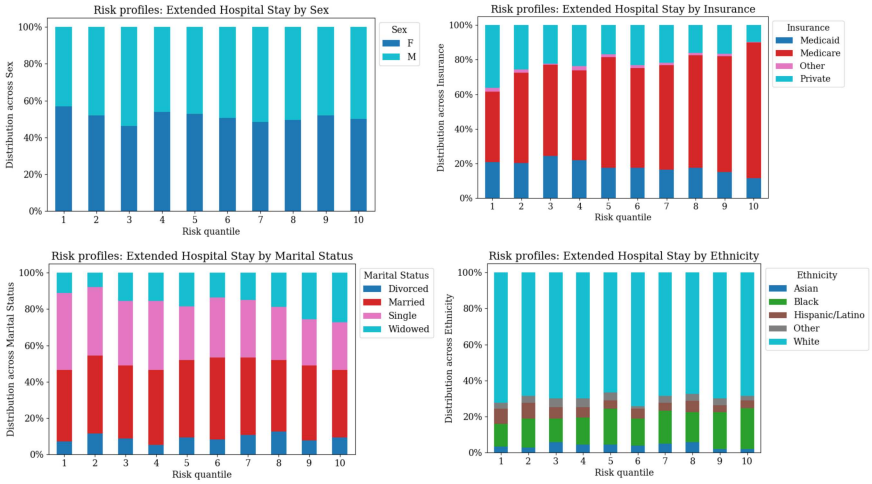

Figure A.2: Risk stratification results in extended hospital stay prediction using the unimodal free-text model (**NT-TF-E**). The stacked barchart highlights the patient profiles per risk group, grouped by sensitive attribute.

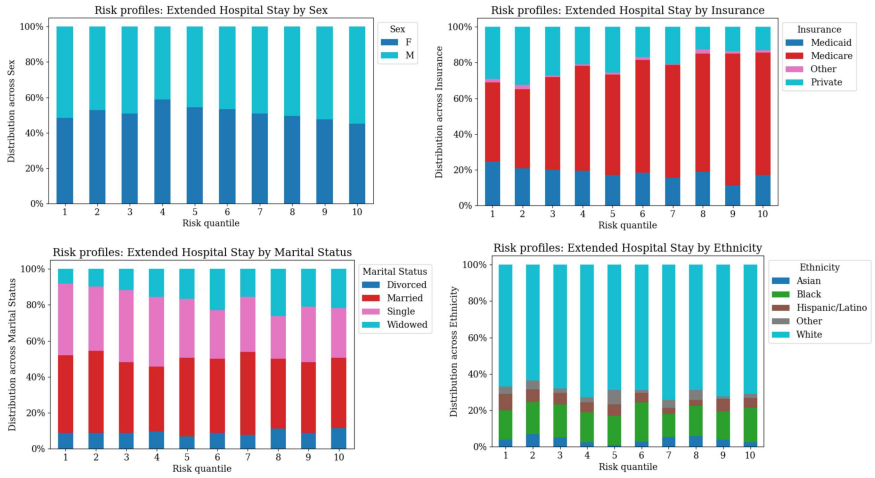

Figure A.3: Risk stratification results in extended hospital stay prediction using the fused static and time-series model (**IF-EHR+TS**). The stacked barchart highlights the patient profiles per risk group, grouped by sensitive attribute.

A.3 Additional error analysis

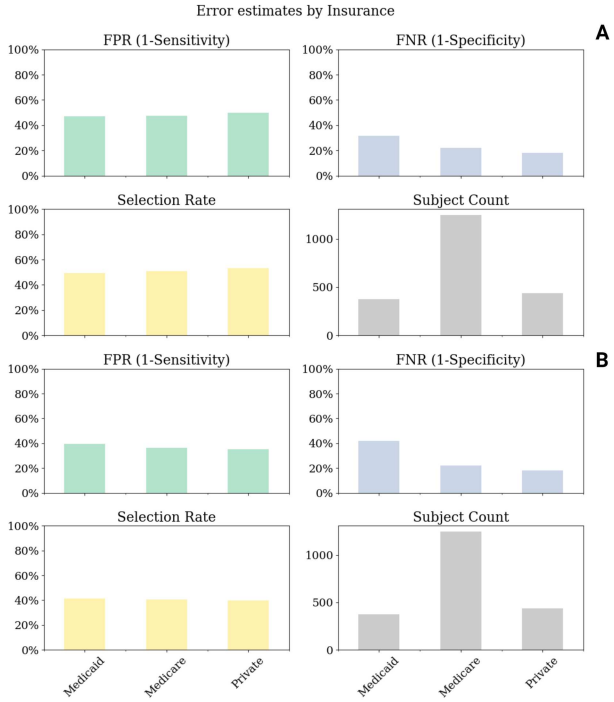

Figure A.4: Error analysis by insurance type, comparing **FPR**, **TPR** and selection rate between the unimodal tabular model (**EHR-MLP**: A) and the fully-fused model (**IF-EHR+TS+NT**: B).

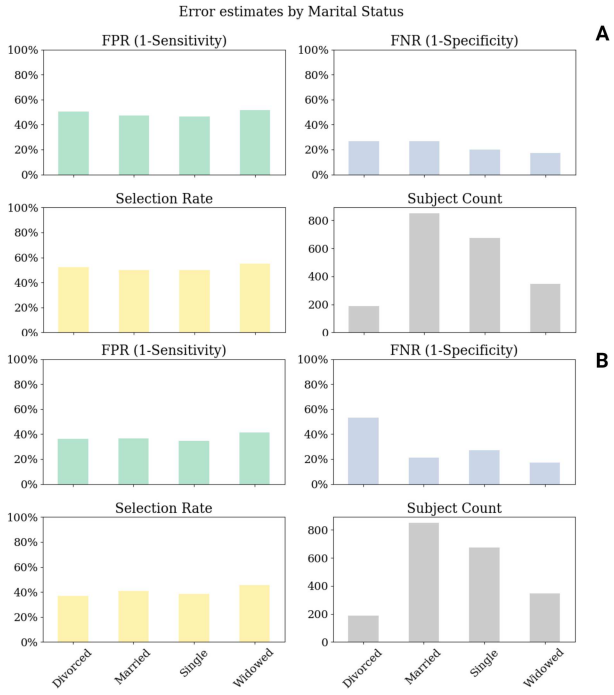

Figure A.5: Error analysis by marital status, comparing **FPR**, **TPR** and selection rate between the unimodal tabular model (**EHR-MLP**: A) and the fully-fused model (**IF-EHR+TS+NT**: B).

A.4 Additional MM-SHAP attribution plots: low-risk case

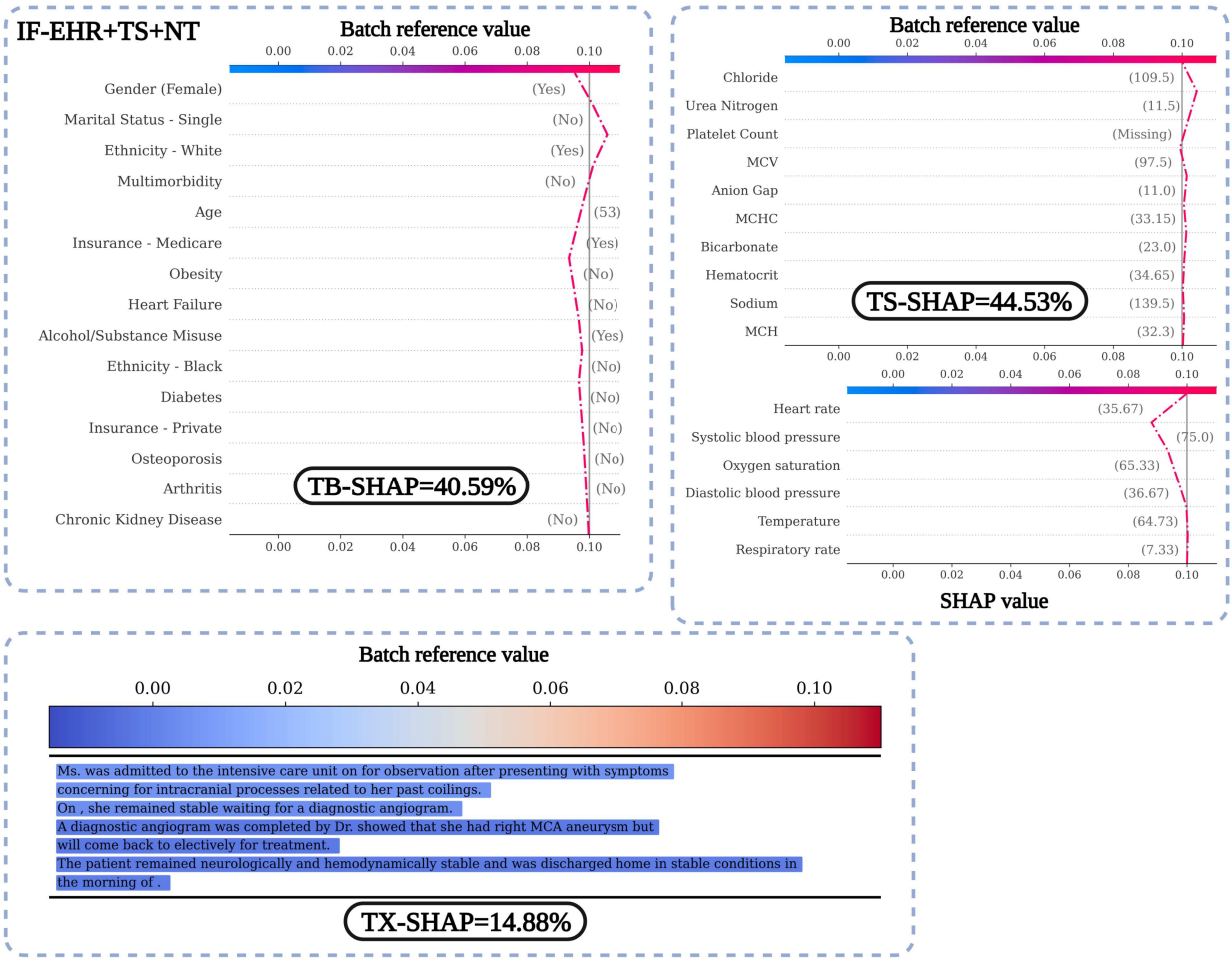

Figure A.6: MM-SHAP feature importance summary for a randomly sampled low-risk individual (risk decile 1) on the test set, using the **IF-EHR+TS+NT** model for prediction of extended stay. Target characteristics: Woman, White ethnicity, Medicare insurance, Married. Top-left: **SHAP** decision plot for the tabular modality and the *TB-SHAP* percentage of dependence; Top-right: **SHAP** decision plots for the time-series modality (vitals and lab tests) and the *TS-SHAP* percentage of dependence; Bottom-left and right: First and last segments from the **SHAP** text plots for the free-text modality (discharge note) and the *TX-SHAP* percentage of dependence;

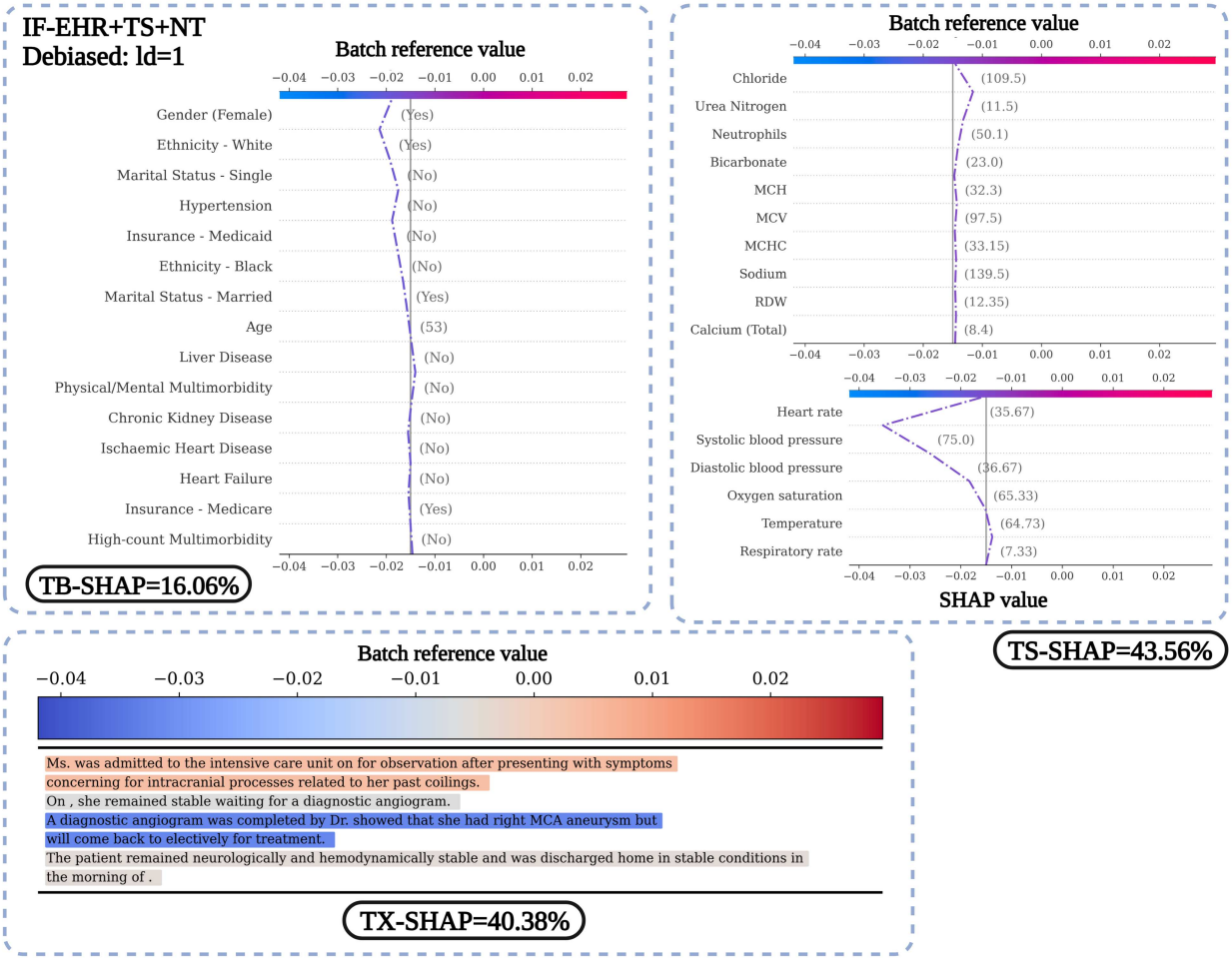

Figure A.7: MM-SHAP feature importance summary for a randomly sampled low-risk individual (risk decile 1) on the test set, using the **IF-EHR+TS+NT** model with debiasing at $\lambda = 1$. Target characteristics: Woman, White ethnicity, Medicare insurance, Married. Top-left: **SHAP** decision plot for the tabular modality and the *TB-SHAP* percentage of dependence; Top-right: **SHAP** decision plots for the time-series modality (vitals and lab tests) and the *TS-SHAP* percentage of dependence; Bottom-left and right: First and last segments from the **SHAP** text plots for the free-text modality (discharge note) and the *TX-SHAP* percentage of dependence;