Learning Normal Representations for Blood Biomarkers

Blood-based biomarkers underpin clinical diagnosis and monitoring, yet their interpretation relies largely on fixed population reference intervals that ignore stable, intra-patient variability. As such, population-based interpretation can mask meaningful deviation from an individual's baseline, risking delayed disease detection. To remedy this, some have proposed personalized interpretation using individual histories, but these methods overfit to sparse data, inflating false-positive rates and risking unnecessary follow-up testing and procedures. Here, using nearly two billion longitudinal laboratory measurements from 1.5 million adults in Clalit Health Services and the eICU Collaborative Research Database, we show that while laboratory values are highly individual,, purely personalized intervals overcorrect, flagging up to 68% measurements as abnormal that are often not associated with adverse clinical outcomes. We introduce NORMA, a transformer trained on 3.4 million longitudinal sequences that generates reference intervals by conditioning on both a patient's history and population-level expectations for health. NORMA-derived intervals detect abnormalities a median of 9.3 months before population intervals in outpatient care and achieve higher precision for predicting mortality, acute kidney injury, and chronic disease across both cohorts. These findings caution against unconstrained personalization in laboratory medicine and demonstrate that anchoring individual trajectories to population-level priors preserves the clinical signal that may be lost with purely personalized approaches. To promote transparency, we publicly release the model, code, and an interactive user interface for accessible, individualized laboratory interpretation.

Disentangling Proxies of Demographic Adjustments in Clinical Equations

The use of coarse demographic adjustments in clinical equations is increasingly scrutinized. Race adjustments have sparked debate, with medical societies recently recommending race-neutral equations. However, removing race by averaging race-specific equations or refitting without race does not address the underlying causes of observed differences. We present ARC (Approach for identifying pRoxies of demographic Correction), a framework to identify factors underlying group-level differences and inform more precise clinical equations. We apply ARC to spirometry, ubiquitous measures of pulmonary function traditionally race-stratified, across National Health and Nutrition Examination Survey (NHANES) and UK Biobank, comprising 147,552 participants. Sociodemographic or exposure measures did not explain reference lung function differences across race groups beyond those explained by age, sex, and height. Waist circumference and sitting height accounted for up to 13% of the remaining lung-volume differences between healthy Black and White adults. We introduce ARC_{PFT}, equations that incorporating such individual-level factors, which outperform the race-neutral GLI-Global equation, recommended by pulmonary societies, in both cohorts. Compared to the GLI-Global, inclusion of sitting height and waist circumference in ARC_{PFT} decreased the mean absolute error (MAE) by 15% among Black participants in the UK Biobank and by 22% in NHANES. ARC_{PFT} also reduced vulnerability to domain shift, with MAE 12.5% and 25.5% lower than race-stratified models in Asian and Hispanic populations,

respectively. This approach helps identify proxies of imprecise demographic adjustments and supports the development of personalized equations across clinical contexts.

Directing Generalist Vision-Language Models to Interpret Medical Images Across Populations

Abstract

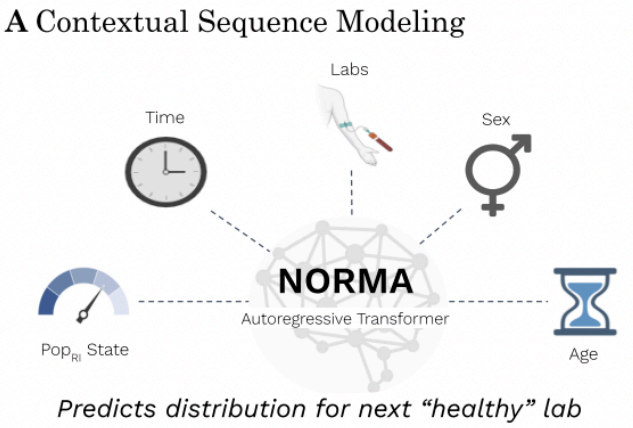
With the proliferation of large multimodal foundation models and increasing use by physicians and patients alike, it is crucial to evaluate the performance and safety of these models across populations and data modalities, including both text and images that can be interpreted by emerging generalist vision-language models. It is equally crucial to use nuanced performance measures that illuminate how users may explicitly or implicitly alter model behavior via prompt formulations in typical use. Here, using two leading models, Gemini Pro Vision (Google) and GPT-4 with Vision (OpenAI), we systematically evaluate how steerable vision-language models are in three common medical imaging tasks: the detection of malignancies in dermatological lesions, detection of abnormalities in chest X-ray radiographs, detection of tumors and stroma in histological samples. These models have guardrails in place to prevent the interpretation of medical images; however, we show that guardrails can be circumvented using relatively simple prompting techniques. We further measure performance and bias across demographic groups and how performance metrics change across prompting strategies, finding profound differences in the way vision-models models exchange sensitivity and specificity by image type, prompting strategy, and by population group. We found that Gemini consistently outperformed GPT-4V in diagnostic accuracy, with maximal balanced accuracies of 0.67 (+/- 0.04) in dermatology, approaching the performance of human dermatologists, 0.75 (+/- 0.04) in radiology images, and 0.81 (+/- 0.02) in histology samples. However, the performance was lowest in skin lesions on darker skin tones, X-rays of older patients, and varied across image pixel intensities. While prompt engineering improved model performance, large foundation models are still far from being reliable for medical image characterization and prone to bias. These findings underscore the necessity for models to be trained on diverse datasets to enhance performance across populations and the importance of placing vision-language model evaluations in context.

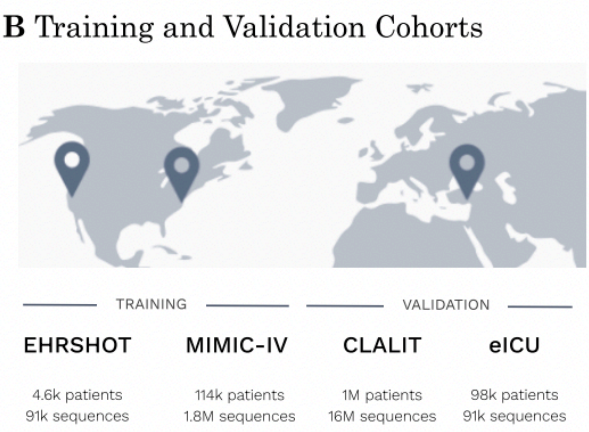
TRACE: A data-driven framework for explaining sensitive attribute detection from medical imaging

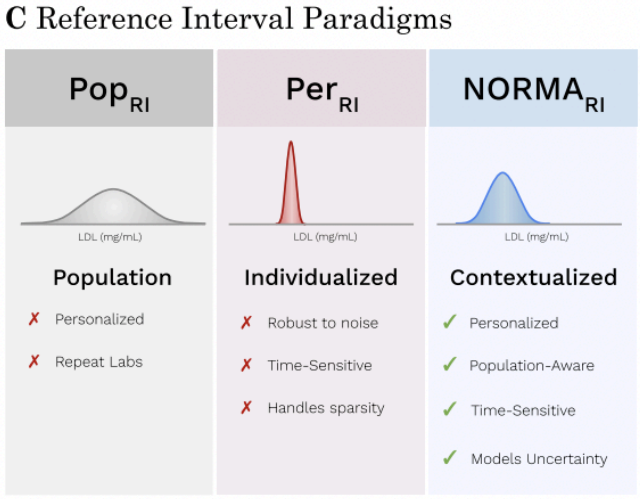
Abstract

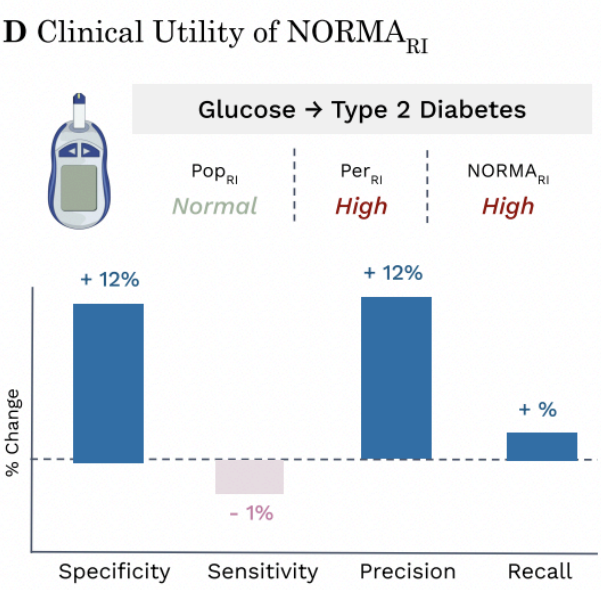
Deep learning models have demonstrated surprisingly strong performance in predicting demographic attributes, including sex, age, and race, from medical images. While clinical prediction models can use such features to improve performance, they may inadvertently rely on spurious correlations between the outcome and demographic features. To prevent potential biases and improve generalizability, it is crucial to understand the underlying reasons why these models can predict demographics from medical images. Here, we introduce TRACE, a framework designed to identify clinical characteristics and imaging artifacts that influence the prediction of sensitive attributes (age, sex, and race) from medical images, and apply it to subgroup detection in chest X-rays using three complementary approaches: (1) distorting images to evaluate the contribution of anatomical morphology and size; (2) utilizing tabular-based model consisting of features such as (e.g. BMI, left lung size, support devices); and (3) examining how subgroup-related features are encoded within learned subgroup embeddings via transfer learning. Our findings reveal that subgroup prediction is influenced by correlated factors such as BMI and imaging artifacts, such as acquisition protocol (AP/PA view); and the pathology label, “No Finding.” Additionally, we show that models rely in part on the morphology of anatomical features, with AUCs of (X-Y) when using anatomical outlines. Our results suggest that race prediction can be further explained by measurable features through off-the-shelf interpretability techniques, paving the way for new methods and techniques to be designed to improve the reliability of medical imaging models.

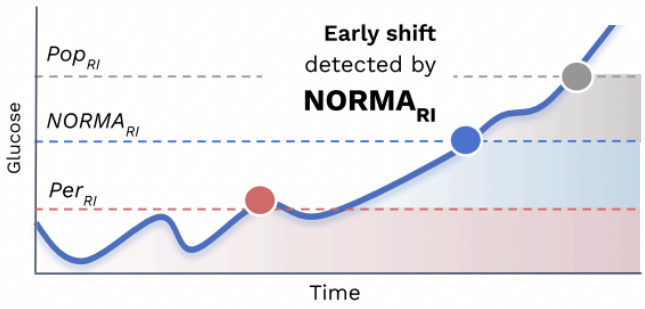
NORMA:

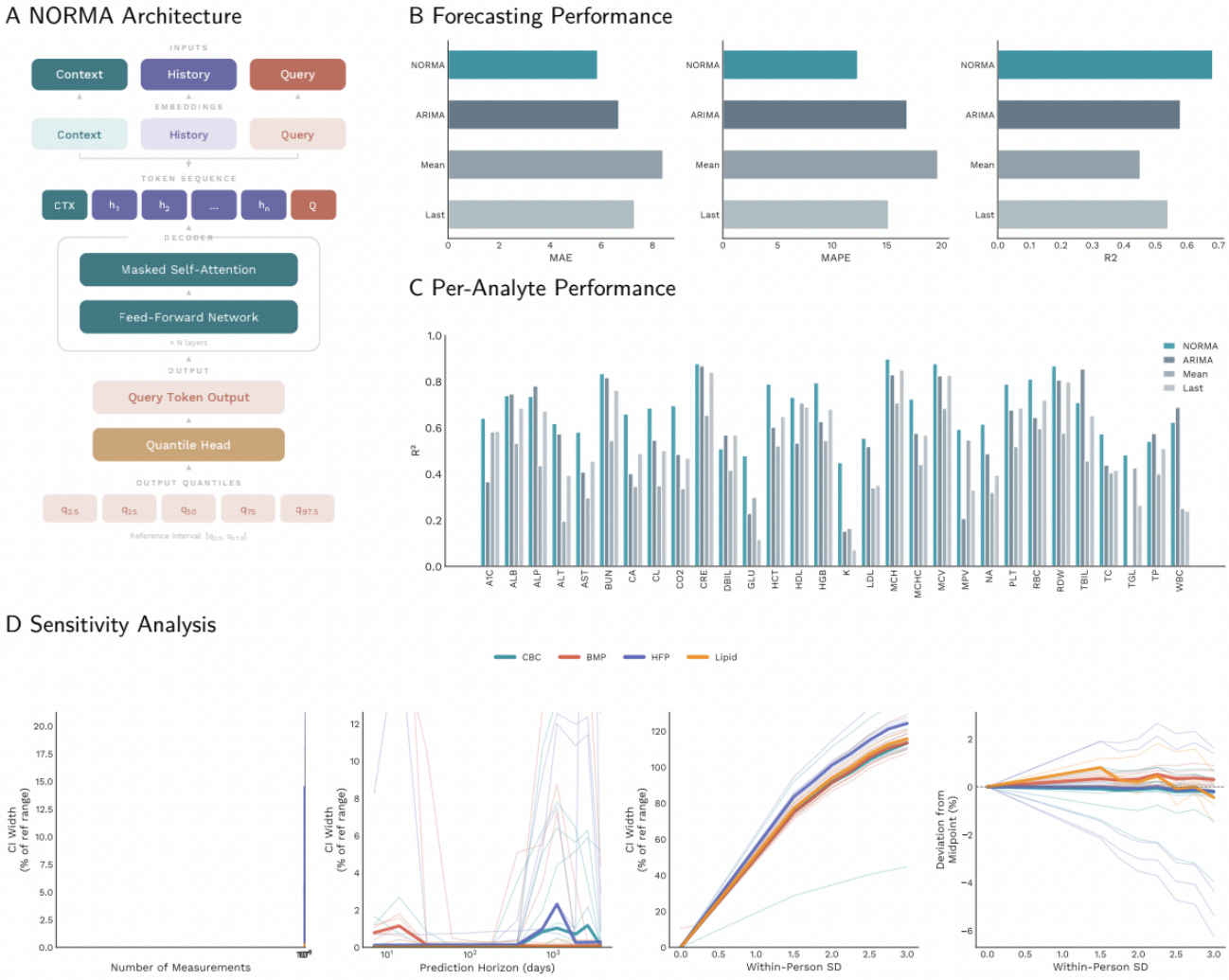

A Contextual Sequence Modeling
Labs
Time
Sex
NORMA
Autoregressive Transformer
Pop$_{RI}$ State
Age
Predicts distribution for next "healthy" lab


B Training and Validation Cohorts
TRAINING
VALIDATION
EHRSHOT
MIMIC-IV
CLALIT
eICU
4.6k patients
91k sequences
114k patients
1.8M sequences
1M patients
16M sequences
98k patients
91k sequences
20 years | 4M sequences | 30 labs
Complete Blood Count
Hepatic Function
Metabolic Function
Lipid Panel

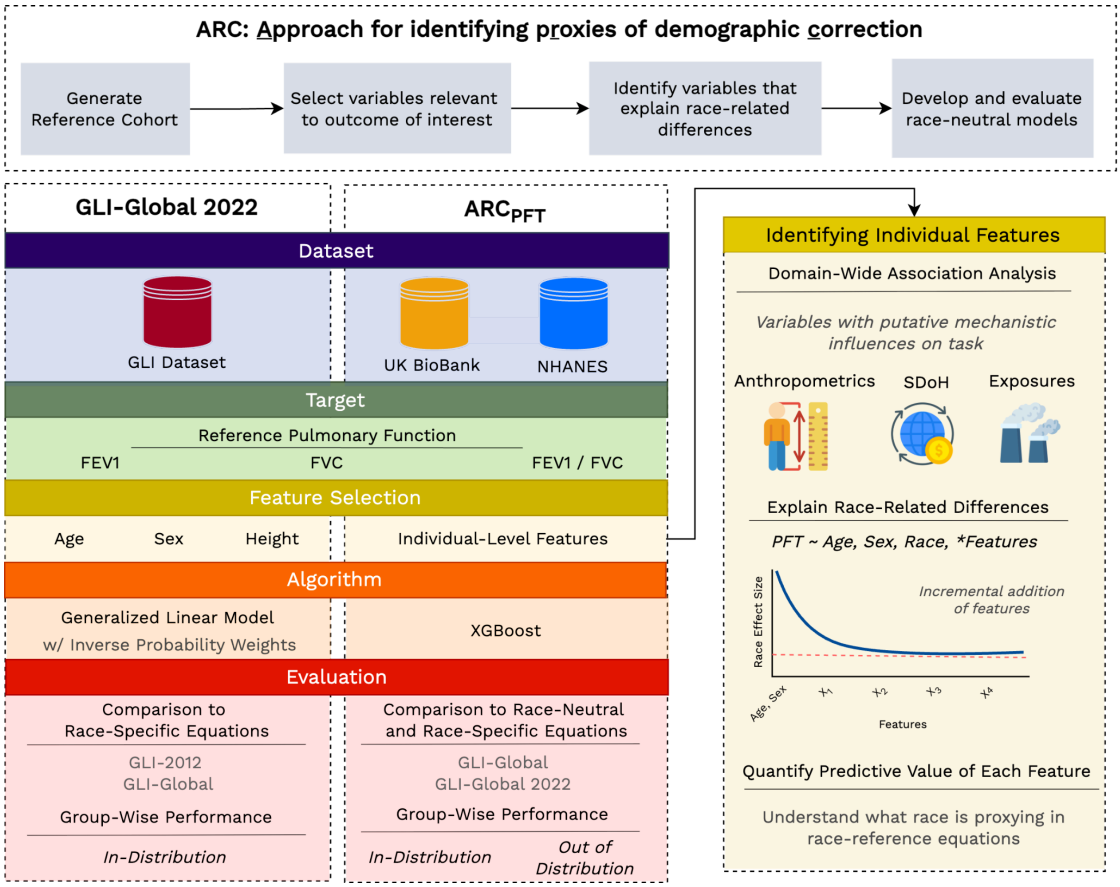

C Reference Interval Paradigms
Pop$_{RI}$
Per$_{RI}$
NORMA$_{RI}$
LDL (mg/mL)
Population
✗ Personalized
✗ Repeat Labs
Individualized
✗ Robust to noise
✗ Time-Sensitive
✗ Handles sparsity
Contextualized
✓ Personalized
✓ Population-Aware
✓ Time-Sensitive
✓ Models Uncertainty


D Clinical Utility of NORMA$_{RI}$
Glucose → Type 2 Diabetes
Pop$_{RI}$
Normal
Per$_{RI}$
High
NORMA$_{RI}$
High
+ 12%
+ 12%
+ %
- 1%
% Change
Specificity
Sensitivity
Precision
Recall


E Early Risk Stratification
Pop$_{RI}$
NORMA$_{RI}$
Per$_{RI}$
Early shift
detected by
NORMA$_{RI}$
Glucose
Time


A NORMA Architecture
INPUTS
Context
History
Query
EMBEDDINGS
Context
History
Query
TOKEN SEQUENCE
CTX
h1
h2
...
hn
Q
DECODER
Masked Self-Attention
Feed-Forward Network
× N layers
OUTPUT
Query Token Output
Quantile Head
OUTPUT QUANTILES
q2.5
q25
q50
q75
q97.5
Reference interval: [q2.5, q97.5]
B Forecasting Performance
NORMA
ARIMA
Mean
Last
MAE
MAPE
R2
C Per-Analyte Performance
R²
NORMA
ARIMA
Mean
Last
AIC
ALB
ALP
ALT
AST
BUN
CA
CL
CO2
CRE
DBIL
GLU
HCT
HDL
HGB
K
LDL
MCH
MCHC
MCV
MPV
NA
PLT
RBC
RDW
TBIL
TC
TGL
TP
WBC
D Sensitivity Analysis
CBC
BMP
HFP
Lipid
CI Width (% of ref range)
Number of Measurements
Prediction Horizon (days)
Within-Person SD
Deviation from Midpoint (%)

ARC:


ARC: Approach for identifying proxies of demographic correction
Generate Reference Cohort
Select variables relevant to outcome of interest
Identify variables that explain race-related differences
Develop and evaluate race-neutral models
GLI-Global 2022
ARC PFT
Dataset
GLI Dataset
UK BioBank
NHANES
Target
Reference Pulmonary Function
FEV1
FVC
FEV1 / FVC
Feature Selection
Age
Sex
Height
Individual-Level Features
Algorithm
Generalized Linear Model
w/ Inverse Probability Weights
XGBoost
Evaluation
Comparison to Race-Specific Equations
GLI-2012
GLI-Global
Group-Wise Performance
In-Distribution
Comparison to Race-Neutral and Race-Specific Equations
GLI-Global
GLI-Global 2022
Group-Wise Performance
In-Distribution
Out of Distribution
Identifying Individual Features
Domain-Wide Association Analysis
Variables with putative mechanistic influences on task
Anthropometrics
SDoH
Exposures
Explain Race-Related Differences
PFT ~ Age, Sex, Race, *Features
Incremental addition of features
Race Effect Size
Age, Sex
X1
X2
X3
X4
Features
Quantify Predictive Value of Each Feature
Understand what race is proxying in race-reference equations

SHORTCUTS & interpretability