Algorithmic Fairness of QPrediction Cardiometabolic Risk Prediction Models

Inchuen Huynh[1], Denise Utochkin[1], Alexandros Katsiferis[2], Tri-Long Nguyen[3,1], and Tibor V Varga[1]

[1]Copenhagen Health Complexity Center, Department of Public Health, University of Copenhagen, Copenhagen, Denmark.
[2]Section for Health Data Science and AI, Department of Public Health, University of Copenhagen, Copenhagen, Denmark.
[3]Section of Epidemiology, Department of Public Health, University of Copenhagen, Copenhagen, Denmark.

Corresponding author:
Tibor V. Varga
ORCID: `https://orcid.org/0000-0002-2383-699X`
Postal address: Øster Farimagsgade 5, DK-1353 Copenhagen K
Copenhagen Health Complexity Center, Department of Public Health, University of Copenhagen, Copenhagen, Denmark.

Abstract

Background There is a lack of fairness evaluation of clinical prediction models used in routine practice. This study aimed to evaluate the fairness of three established risk prediction models belonging to the QPrediction family: QRISK3, QDiabetes, and QStroke, which estimate 10-year risks for cardiovascular disease, type 2 diabetes, and stroke, respectively.

Methods and findings We used data from the UK Biobank, a large-scale prospective cohort study comprising 502,131 participants. To assess calibration, the predicted risks were compared to the observed cumulative incidence rates, and survival Brier scores were calculated as a summary metric of model calibration. To assess discrimination, we calculated overall and group-specific true positive, true negative, false positive, and false negative rates (TPR, TNR, FPR, FNR). We compared the rates between subgroups of demographics and considered prediction parity when values were close to each other to a prespecified degree. Overall, all models showed systematic overprediction of risks for the UK Biobank validation cohorts, with the degree of overprediction varying between the models. Prediction parity was not reached in FNR and TNR in terms of ethnicity, education level, average household income before taxation, and Townsend deprivation score. For immigration status, there were no major differences in the discrimination metrics, and prediction parity was achieved. For QDiabetes, the trends were similar, apart from ethnicity, where the opposite was observed as for QRISK3.

Conclusions Inclusion of ethnicity and deprivation in the models may have contributed to more consistent calibration across subgroups. Variables were measured only at baseline, while many social and economic conditions can shift over time. Although including social determinants of health can improve predictive performance, the models are typically deployed in a clinical setting to guide individual-level decision making, which can inadvertently shift responsibility for health outcomes onto patients.

1 Introduction

Prognostic predictive models enable the identification of individuals at risk of a disease, providing the opportunity for early preventive interventions. Although hundreds of cardiometabolic clinical prediction models have been published to date [1][2], only a handful of these models have been incorporated into national guidelines [3] or clinical practice [4][5].

In the process of validating prognostic prediction models, the performance of models is most often assessed based on their ability to provide overall accurate predictions, while the fairness of the provided predictions has received less attention. In recent years, a number of alarming reports emerged showcasing algorithms and healthcare processes that benefit privileged population subgroups, while underperforming among those vulnerable [6][7][8][9]. The reasons for such results are manifold but often include developing models using imbalanced datasets, using inappropriate feature or outcome proxies, omitting key socioeconomic determinants of health (SDOH) in model development, and failing to co-create models with vulnerable communities to avoid knowledge gaps in model creation and to assess the needs of the patients [10][11][12].

Algorithmic fairness has become a rapidly developing field of science with the overall goal to conceptualize, quantify, and mitigate biases arising from prediction models. Fairness aspects of predictive models have been gaining increasing attention in recent years, with TRIPOD+AI [13], the prevailing consensus guideline established to improve the reporting of prediction models, calling for all papers reporting on diagnostic or prognostic prediction models to address aspects of fairness. *Fairness* is a subjective term, making the assessment of prediction models from a fairness perspective a multifaceted and complex, if not impossible, task. In this project, we define fairness as a property of predictive models to provide an equal average prediction for prespecified groups in populations based on various attributes that might make a group more vulnerable.

In this article, we evaluate the fairness of three established risk prediction models belonging to the QPrediction family: QRISK3 [14], QDiabetes [15], and QStroke [16], predicting future risk of cardiovascular disease (CVD), type 2 diabetes, and stroke, respectively. All three tools are publicly accessible, with QRISK3, QDiabetes, and QStroke being integrated into primary clinical practice in the United Kingdom (UK), used by over 55% of general practices in England [17] and 46% of general practices in Wales [18], and being the most widely used primary care clinical software in Scotland [19]. QRISK3 is recommended for clinical use in the UK by the National Institute for Health and Care Excellence (NICE) [3]. For risk identification of type 2 diabetes, NICE recommends using computer-based risk assessment tools [20] and up until 2019, QDiabetes was included in the list of tools recommended in the NHS Health Check best practice guideline [21].

Although the performance of QPrediction models has been extensively validated in previous articles [22][23][24][25], fairness aspects have not been featured in these studies. In this article, we use data from the UK Biobank for validation, a longitudinal prospective cohort study with over half a million participants from the UK with extensive information about their health status, sociodemographic information, lifestyle, and other factors.

2 Methods

2.1 Design and Study Population

The UK Biobank is a longitudinal prospective cohort study encompassing over half a million participants from the UK. Extensive individual-level data is available on participants' lifestyles, SDOH parameters, biomarkers, and morbidities. Participants were recruited during 2006-2010 and were between ages 40-69 at the time of recruitment. Four repeat assessments were carried out during

2012-2019. At baseline, questionnaire and interview data, physical measurements, and samples of blood, urine, and saliva were collected as described previously [26]. Follow-up on the participants is conducted through linkages to routinely available national datasets, including death registrations, cancer registration, hospital inpatient episodes, hospital outpatient episodes, and primary care data. The hospital inpatient data used to identify disease incidence for this article covers individual patient journeys up until 2022. This study involves human participants and was approved by the National Health Service's National Research Ethics Service North West (21/NW/0157). Participants gave informed consent to participate in the study before taking part.

We applied different exclusion criteria to assess each of the QPrediction models, as described in Figure 1. For all of the models, participants with missing Townsend deprivation scores were excluded as well as those with preexisting diagnoses of the disease outcome of interest. Additionally, participants using statins at study entry were excluded from the study population for QRISK3 and those using anticoagulants were excluded from the study population for QStroke. The diagnostic codes used to define the exclusion criteria related to pre-existing diseases correspond to those defined in the original articles for the development of the models.

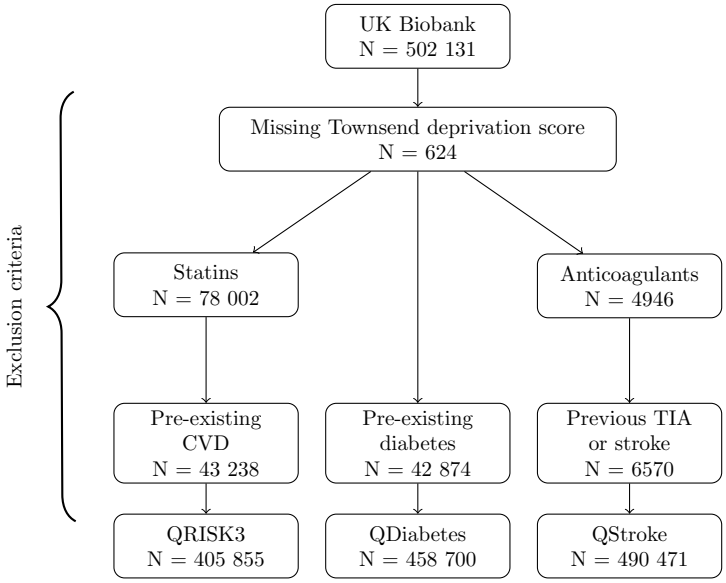

Figure 1: Flowchart of exclusion criteria for each QPrediction model. Note that there are overlaps between groups with various exclusion criteria, therefore the total number of excluded individuals is less than the sum of the individuals under each exclusion criterion.

2.2 Variables

We considered variables measured at the baseline assessment to allow for the longest possible follow-up times. The variables were matched to the risk factors included in the QPrediction models (Table S2.1.1-S2.1.3) using data from the baseline assessment as well as from data linkages. For the QRISK3-model, the variables matched were extracted from a previous validation study using UK Biobank data [27], whereas, for the QDiabetes and QStroke models, variables were matched as described in Table S2.2.1-S2.3.3. When there were no direct matches in available UK Biobank variables for a

Model	Outcome	Variables
QRISK3	10-year risk of cardiovascular diseases	age; atrial fibrillation; atypical antipsychotics; body mass index; chronic kidney disease (stage 3-5); erectile dysfunction; ethnicity; diabetes (type 1 or 2); family history of coronary heart disease in first degree relative aged <60 years; migraine; rheumatoid arthritis; severe mental illness; sex; smoking status; systemic lupus erythematosus; systolic blood pressure; systolic blood pressure variability; total cholesterol; Townsend deprivation; steroids; treated hypertension
QDiabetes	10-year risk of type 2 diabetes	age; atypical antipsychotics; body mass index; ethnicity; family history of diabetes in a first degree relative; fasting blood glucose; gestational diabetes; HbA1c; learning disability; manic depression or schizophrenia; polycystic ovary syndrome; previous heart attack, angina, stroke or transient ischemic attack; sex; smoking status; statins; steroids; Townsend deprivation; treated hypertension
QStroke	10-year risk of ischaemic stroke or transient ischaemic attack	age; atrial fibrillation; body mass index; chronic kidney disease (stage 4-5); congestive cardiac failure; coronary heart disease; diabetes (type 1 or 2); ethnicity; family history of coronary heart disease in first degree relative aged <60 years; heart attack or angina; rheumatoid arthritis; sex; smoking status; systolic blood pressure; total cholesterol (cholesterol/HDL ratio); Townsend deprivation; treated hypertension; valvular heart disease

Table 1: Outcomes and variables included in QRISK3, QDiabetes, and QStroke models.

variable in the QPrediction model, the closest possible match was chosen. The QDiabetes model had three variations and the model chosen to calculate the risk score was based on whether the values for HbA1c or fasting glucose were available, or neither. Since the developers found the model including HbA1c without fasting blood glucose to be the best performing [15], we chose to validate this model.

The variables used to define the subgroups for fairness assessments were sex (men/women), age group (35-45/46-55/56-65/66-75), self-reported ethnicity (White/Asian or Asian British/Black or Black British/Chinese/Mixed/Other ethnic group/Prefer not to answer/Do not know), immigration status (no/yes), Townsend deprivation score (binned to quantiles: [-7,-4]/(-4,-2]/(-2,1]/(1,11]), index of multiple deprivation ([0,20]/(20,40]/(40,60]/(60,80]/(80,100]), average household income (Greater than 100,000/52,000 to 100,000/31,000 to 51,999/18,000 to 30,999/Less than 18,000/Prefer not to answer/Do not know), highest education level (College or University degree/National Vocational Qualification (NVQ) or Higher National Diploma (HND) or Higher National Certificate (HNC) or equivalent/Other prof. qual. eg: nursing, teaching/A levels or AS levels or equivalent/O levels or General Certificate of Secondary Education (GCSEs) or equivalent/Certificate of Secondary Education (CSEs) or equivalent/None of the above/Prefer not to answer), with the assumed privileged groups presented first in the parentheses above.

2.3 Outcomes

Outcomes of interest for the QPrediction models were defined using the International Classification of Diseases (ICD). For the validation data, we matched the codings to the ICD10 and ICD9 (ICD

version 9 and 10) equivalents as well as the Office of Population Censuses and Surveys Classification of Interventions and Procedures version 4 (OPCS-4) codes from hospital episode statistics and death registration data. The matching process is described in detail in Table S3. Follow-up time was measured in years defined as time from the date of baseline assessment until the date of an event of cardiovascular disease, type 2 diabetes, or stroke, death from other causes than the event of interest, loss to follow-up, or the administrative censoring date of the hospital inpatient data in the UK Biobank, which differed by country (England: 31 October 2022; Scotland: 31 August 2022; Wales: 31 May 2022) [28].

2.4 Missing Data Strategy

Missing data was imputed using multiple imputation by chained equations (mice) using the *mice* R package [29]. We assumed a missing-at-random missingness pattern. A total of five imputed copies were generated based on the percentage of missingness in the variables (2.8%) [30]. Convergence was checked visually for all the datasets, and the baseline characteristics were checked before and after the imputations. Results were pooled using Rubin's rules [31].

2.5 Algorithmic Fairness in Calibration

Numerous *algorithmic group fairness* metrics have been developed to assess equal predictive performance; therefore, in general, it is paramount for researchers to accurately outline the purpose of the utilized metrics in relation to the goals of research projects. In our work, we measured algorithmic group fairness through two distinct routes: by quantifying differences in the calibration of models across groups and in prediction error to gain a holistic picture of performance parity.

The fairness of the QPrediction models was evaluated using metrics of calibration and discrimination. The purpose of calibration is to quantify the agreement between observed risks and predictions obtained from the models. Here, we compared the predicted probabilities of the QPrediction models to the observed cumulative incidences derived from weighted Kaplan-Meier estimates. The Kaplan-Meier estimates were weighted using inverse probability censoring weighting (IPCW) to account for competing risk, which occurred when a participant experienced a competing event that prevented them from experiencing the event of interest. Competing risk can be thought of as a special case of informative censoring, where the censoring mechanism is associated with the event of interest, for example, by sharing similar risk factors. Not accounting for competing risk and informative censoring could lead to an overestimation of the risk since the population is not representative of the true population at risk of experiencing the event of interest [32]. With IPCW, the uncensored population is up-weighted for each participant lost to follow-up due to informative censoring [33][34]. The derived weighted Kaplan-Meier risk estimates were validated against the correct Aalen-Johansen risk estimates of the event of interest. In case the weighted Kaplan-Meier estimates matched the ones derived using the Aalen-Johanssen estimator, we deemed the IPCW method as properly accounting for competing risk. The weights obtained using IPCW were used to calculate survival Brier scores, as the time-dependent mean squared difference of the predicted and observed values, as a summary metric of model calibration. The Brier scores have a theoretical range of 0-1, 0 representing perfect calibration, and 1 representing the poorest calibration. For a visual check of calibration, the predicted risks (obtained from the QPrediction models) and observed risks (obtained from weighted Kaplan-Meier estimates) were binned into deciles based on the predicted risk scores and plotted against each other. In addition, differences between the predicted and observed risks were plotted to compare the model performances for different categories of each subgroup. The Brier scores and calibration curves were compared across population subgroups. The survival Brier

scores were not calculated for the QStroke model, as the IPCW weighted estimates did not match the Aalen-Johanssen estimates.

2.6 Algorithmic Fairness in Discrimination

For discrimination comparisons, we calculated key prediction indicators using the four main metrics of the confusion matrix for binary predictions: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Specifically, we quantified overall and group-specific true positive rates (TPR or sensitivity, TP/TP+FN), true negative rates (TNR, specificity, TN/FP+TN), false positive rates (FPR, 1-specificity), and false negative rates (FNR, 1-sensitivity). We considered prediction parity with regard to each metric when values were close to each other to a prespecified degree (80% of the values in the assumed privileged group). The goal of equalizing errors is to further prevent minoritized groups from disproportionately bearing the negative impacts caused by errors and thus perpetuating cycles of inequality [35]. We chose these metrics of discrimination as they are commonly used and widely understood in clinical settings for binary prediction problems. All values for the confusion matrices are weighted using the weights derived using IPCW.

While for calibration we considered time-to-event statistics, for discrimination, we considered binary outcomes. The thresholds used to dichotomize individual risks were chosen from QPrediction guidelines for clinical use. These thresholds hold clinical significance, as individuals surpassing these are likely to receive additional healthcare resources or be prioritized for preventive action. For QRISK3, NICE recommends offering statins for people with a score of 10% or more for the prevention of cardiovascular disease [3]. For QDiabetes, the NHS Health Check Best Practice Guideline 2019 considers people with a QDiabetes score of 5.6% or more to be at high risk of diabetes [21]. As QStroke has not been recommended by any national guidelines, we omitted QStroke from the discrimination analysis as the dichotomization of the risk scores could not be clinically reasoned, and the impact of the results would not be interpretable in a clinical setting.

For assessing the cost of misclassification, we calculated net benefits for QRISK3 and QDiabetes for each demographic subgroup. Net benefit (NB) is the expected number of true-positive treatments with a subtracted weighted penalty for unnecessary treatment per patient [36]. The weights for the unnecessary treatments are calculated using the thresholds from QPrediction guidelines for clinical use mentioned earlier. The calculated net benefits are compared to scenarios where every individual is treated (TrA) and where no one is treated (TrN) to determine the performance of the model. In addition, we calculated the net interventions avoided (NI avoided) to assess whether using the QPrediction models helps reduce unnecessary treatments.

This study was undertaken following the TRIPOD+AI guidelines [13] (see checklist in Table S10).

2.7 Software and Code

R was used for data cleaning and analysis, and version control was done using Git. The codes are available at GitHub: `https://github.com/ihuynhi/qpred_uk_biobank`. No study protocol was prepared beforehand.

3 Results

In total, 502,131 UK Biobank participants were considered for inclusion in this study. After applying model-specific exclusion criteria, we analyzed data from 405,855 participants for QRISK3, 458,700 participants for QDiabetes, and 490,471 participants for QStroke. The median follow-up times

were 13.60 years for QRISK3, 13.62 years for QDiabetes, and 13.63 years for QStroke. Baseline characteristics of the model-specific cohorts can be found in the supplemental tables (Table S4.1-S4.3). The baseline cohort (QRISK3) had a mean age of 55.5 (SD = 8.1), and was 57.9% female and 94.2% self-reported their ethnicity as White.

Variable	UK Biobank, female n = 234 808	QRISK3 derivation cohort, female n = 4 019 956	UK Biobank, male n = 171 047	QRISK3 derivation cohort, male n = 3 869 847
Age (years), mean (SD	55.6 (8.0)	43.3 (15.3)	55.4 (8.2)	42.6 (14.0)
Systolic blood pressure (mm Hg), mean (SD)	136.4 (20.2)	123.2 (18.2)	142.2 (18.4)	129.2 (16.3)
Measure of systolic blood pressure variability (SD of repeated measures), mean (SD)	5.4 (4.5)	9.3 (6.2)	5.2 (4.3	9.9 (6.8)
Total cholesterol: high-density lipoprotein cholesterol ratio (mmol/L), mean (SD)	3.9 (1.0)	3.7 (1.2)	4.6 (1.1)	4.4 (1.4)
Family history of coronary heart disease in first degree relative ¡60 years	100846 (42.9)	481 628 (12.0)	61886 (36.2)	357 987 (9.3)
Self-reported ethnicity				
White or not stated	221511 (94.3)	3 564 651 (88.7)	160779 (94.0)	3 435 408 (88.8)
Indian	2330 (1.0)	77 683 (1.9)	1977 (1.2)	81 805 (2.1)
Pakistani	555 (0.2)	39 541 (1.0)	738 (0.4)	46 948 (1.2)
Bangladeshi	50 (0.02)	31 930 (0.8)	106 (0.1)	42 111 (1.1)
Other Asian	690 (0.3)	53 559 (1.3)	673 (0.4)	45 753 (1.2)
Black Caribbean	2383 (1.0)	37 781 (0.9)	1341 (0.8)	30 610 (0.8)
Black African	1421 (0.6)	77 813 (1.9)	1388 (0.8)	71 245 (1.8)
Chinese	888 (0.4)	33 767 (0.8)	497 (0.3)	23 730 (0.6)
Other ethnic group	4588 (2.0)	103 231 (2.6)	3201 (1.9)	92 237 (2.4)
Townsend deprivation score, mean (SD)	-1.4 (3.0)	0.4 (3.2)	-1.3 (3.1)	0.5 (3.3)
Body mass index (kg/m2), mean (SD)	26.7 (5.0)	25.4 (5.1)	27.4 (4.1)	25.9 (4.2)
Smoking status				
Non-smoker	141372 (60.2)	2 051 803 (51.0)	88892 (52.0)	1 463 941 (37.8)
Ex-smoker	71834 (30.6)	589 521 (14.7)	59840 (35.0)	594 265 (15.4)
Light smoker (¡10 a day)	8791 (3.7)	434 954 (10.8)	8415 (4.9)	507 523 (13.1)
Moderate smoker (10–19 a day)	6939 (3.0)	226 128 (5.6)	5207 (3.0)	251 170 (6.5)
Heavy smoker (20 or over a day)	4428 (1.9)	115 890 (2.9)	5693 (3.3)	188 857 (4.9)
Atrial fibrillation	1016 (0.4)	15 177 (0.4)	1822 (1.1)	20 098 (0.5)
Erectile dysfunction	NA	NA	351 (0.2)	90 753 (2.3)
Migraine	11398 (4.9)	257 825 (6.4)	3099 (1.8)	103 995 (2.7)
Rheumatoid arthritis	3313 (1.4)	45 700 (1.1)	1225 (0.7)	20 997 (0.5)
Chronic kidney disease (stage 3-5)	189 (0.1)	19 396 (0.5)	129 (0.1)	12 254 (0.3)
Severe mental illness	846 (0.4)	274 069 (6.8)	731 (0.4)	167 115 (4.3)
Systemic lupus erythematosus	457 (0.2)	4010 (0.1)	47 (0.03)	365 (0.0)
Diabetes type 1	2618 (1.1)	10 060 (0.3)	3541 (2.1)	11 617 (0.3)
Diabetes type 2	2922 (1.2)	48 022 (1.2)	3746 (2.2)	58 395 (1.5)
Atypical antipsychotic	487 (0.2)	19 140 (0.5)	467 (0.3)	20 123 (0.5)
Corticosteroid	1579 (0.7)	96 955 (2.4)	1024 (0.6)	56 533 (1.5)
Treated hypertension	30543 (13.0)	223 494 (5.6)	23152 (13.5)	164 255 (4.2)

Table 2: Baseline charasteristics of UK Biobank cohort used for validating QRISK3 as well as the QResearch cohort used to derive the models. The tables of baseline charasteristics for QDiabetes and QStroke can be found from the supplements.

3.1 Calibration

All models showed systematic overprediction of risks for the UK Biobank validation cohorts based on the calibration plots (Figure S8.1) and survival Brier scores (Table S9.1, S9.2), with only the degree of overprediction varying between the models. As the survival Brier scoe

The survival Brier scores ranged between 0.59, the best calibration across our analysis demonstrated for the oldest age group (65-75 year-olds), to 0.93, the poorest calibration across our analysis demonstrated for the youngest age group (35-45 year-olds). When comparing the differences between observed and predicted risks, only minor differences could be seen across subgroups. For example, Figure 2 shows the risk differences based on the highest education level, and the overprediction was the highest in QRISK3 for participants with a college or university degree by a small margin. For QDiabetes and QStroke, the differences appeared negligible. Similar patterns could be observed for other subgroups.

3.2 Discrimination

For ethnicity, the TNR and FNR values were lower for participants who self-reported as White in QRISK3 compared with any other ethnicity, apart from Asian and Asian British, although the sample size of the latter group was small. That is, the participants who self-reported as White were less likely to remain undiagnosed, both in case of them truly having no CVD (TNR) and in case

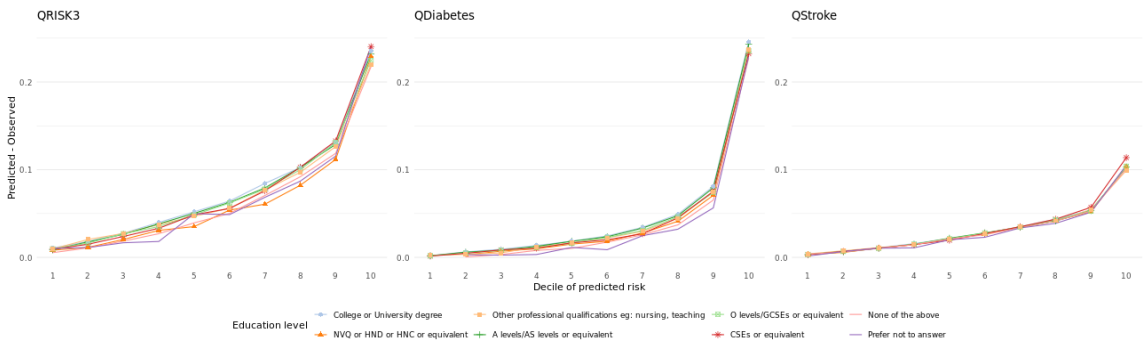

Figure 2: Difference of predicted and observed values based on education level for QRISK3, QDiabetes, and QStroke.

they truly having CVD (FNR). In other words, White participants were in overall more likely to be diagnosed with CVD: they were subjects to both overestimation (FPR) and accurate diagnosis of CVD (TPR). Assuming White to be the privileged group and considering sample size, prediction parity was not reached in any of the subgroups in terms of TPR and FPR. In terms of education level, average household income before taxation, Townsend deprivation score, and index of multiple deprivation, the TNR and FNR seemed generally lower in the lower levels of education, income, and deprivation score. Here, prediction parity was not achieved in FNR and TNR, but was achieved in FPR and TPR for all of the variables mentioned apart from income. For immigration status, there were no major differences in the discrimination metrics, and prediction parity was achieved. For QDiabetes, the trends were similar, apart from ethnicity, where we could observe the opposite pattern as shown for QRISK3.

Apart from the age group 35-45, all the net interventions avoided were negative (Table S9.3) which translates to over-treatment and increase of unnecessary interventions, corollary to the overprediction observed in calibration and FPRs. Comparing NB and TrN for different ethnicities in QRISK3, the NBs were slightly higher than the TrN values for participants who self-reported as Black or Black British, Asian or Asian British, Do not know, and Prefer not to answer. For all the ethnicities, the TrA values were higher than the NB values. That is, deciding to treat everyone seems to yield less unnecessary interventions compared to assigning interventions based on the QRISK3.

4 Discussion

In our analysis, all models demonstrated systematic overprediction overall and across all evaluated subgroups defined by protected attributes, such as sex, age, ethnicity, immigrant status, income, education, and area-level deprivation. When comparing predictive performance across subgroups, we observed substantial variation in discrimination metrics across certain groups. However, calibration was relatively consistent, with no large differences observed across subgroups except for a few comparisons.

Our analysis allowed us to quantify average predictive metrics across population subgroups in the UK; by comparing these metrics, we were able to evaluate aspects of algorithmic fairness. While a wide range of algorithmic group fairness metrics have been proposed to assess parity in predictive performance, their interpretation depends heavily on the context and objectives of the study. Here, we assessed group fairness along two distinct routes: differences in prediction error and differences

	QRISK3									
	Sample size	Survival brier score	TPR	FPR	TNR	FNR	TPR parity	FPR parity	TNR parity	FNR parity
Overall	405855	0.78	0.63	0.38	0.62	0.37	1.00	1.00	1.00	1.00
Sex										
Male	234808.0	0.77	0.78	0.46	0.54	0.22	Ref	Ref	Ref	Ref
Female	171047.0	0.80	0.53	0.27	0.73	0.47	0.68	0.58	1.35	2.13
Age group										
[35,45]	60888.0	0.93	0.08	0.02	0.98	0.92	Ref	Ref	Ref	Ref
(45,55]	131177.0	0.86	0.25	0.11	0.89	0.75	3.04	6.23	0.91	0.82
(55,65]	166011.0	0.72	0.68	0.58	0.42	0.32	8.33	32.81	0.42	0.35
(65,75]	47779.0	0.59	0.96	0.95	0.05	0.04	11.80	53.56	0.05	0.04
Ethnicity										
White	383023.8	0.78	0.64	0.39	0.61	0.36	Ref	Ref	Ref	Ref
Mixed	2533.6	0.85	0.38	0.17	0.83	0.62	0.60	0.44	1.35	1.71
Other ethnic group	3675.8	0.84	0.44	0.19	0.81	0.56	0.69	0.49	1.33	1.54
Prefer not to answer	1300.2	0.79	0.47	0.29	0.71	0.53	0.73	0.74	1.17	1.47
Chinese	1385.0	0.89	0.41	0.10	0.90	0.59	0.64	0.26	1.47	1.64
Black or Black British	6649.2	0.86	0.46	0.15	0.85	0.54	0.72	0.38	1.39	1.49
Asian or Asian British	7119.4	0.75	0.65	0.36	0.64	0.35	1.02	0.92	1.05	0.97
Do not know	168.0	0.79	0.33	0.22	0.78	0.67	0.52	0.58	1.27	1.85
Immigration status										
No	369896.0	0.78	0.63	0.39	0.61	0.37	Ref	Ref	Ref	Ref
Yes	35959.0	0.80	0.62	0.32	0.68	0.38	0.97	0.83	1.11	1.05
Average household income										
Greater than 100,000	20479.6	0.86	0.35	0.15	0.85	0.65	Ref	Ref	Ref	Ref
52,000 to 100,000	76868.6	0.85	0.38	0.19	0.81	0.62	1.09	1.25	0.96	0.95
31,000 to 51,999	95015.6	0.81	0.52	0.29	0.71	0.48	1.51	1.95	0.83	0.73
18,000 to 30,999	86649.6	0.76	0.69	0.46	0.54	0.31	1.98	3.12	0.63	0.48
Less than 18,000	70776.8	0.71	0.77	0.58	0.42	0.23	2.24	3.92	0.49	0.35
Prefer not to answer	39660.6	0.75	0.70	0.49	0.51	0.30	2.03	3.31	0.60	0.46
Do not know	16404.2	0.72	0.77	0.58	0.42	0.23	2.24	3.90	0.49	0.35
Highest education level										
College or University degree	139399.4	0.82	0.54	0.29	0.71	0.46	Ref	Ref	Ref	Ref
NVQ or HND or HNC or equivalent	25162.8	0.76	0.63	0.39	0.61	0.37	1.16	1.35	0.86	0.81
Other professional qualifications eg: nursing, teaching	20154.0	0.74	0.73	0.52	0.48	0.27	1.34	1.80	0.67	0.60
A levels/AS levels or equivalent	47187.4	0.81	0.56	0.32	0.68	0.44	1.04	1.10	0.96	0.95
O levels/GCSEs or equivalent	87101.8	0.78	0.62	0.39	0.61	0.38	1.15	1.33	0.86	0.82
CSEs or equivalent	23220.4	0.83	0.41	0.22	0.78	0.59	0.75	0.77	1.10	1.29
None of the above	59587.2	0.69	0.78	0.65	0.35	0.22	1.44	2.23	0.50	0.48
Prefer not to answer	4042.0	0.73	0.71	0.50	0.50	0.29	1.30	1.71	0.71	0.64
Townsend deprivation score (quantiles)										
[-7,-4]	78670	0.80	0.58	0.34	0.66	0.42	Ref	Ref	Ref	Ref
(-4,-2]	135751	0.79	0.61	0.38	0.62	0.39	1.05	1.09	0.95	0.93
(-2,1]	105950	0.78	0.64	0.38	0.62	0.36	1.11	1.12	0.94	0.85
(1,11]	85484	0.76	0.70	0.42	0.58	0.30	1.22	1.22	0.89	0.70
Index of multiple deprivation										
[0,20]	285207.8	0.79	0.61	0.37	0.63	0.39	Ref	Ref	Ref	Ref
(20,40]	88085.6	0.77	0.66	0.40	0.60	0.34	1.08	1.07	0.96	0.88
(40,60]	27027	0.75	0.68	0.43	0.57	0.32	1.10	1.16	0.91	0.84
(60,80]	5410.2	0.73	0.75	0.46	0.54	0.25	1.22	1.25	0.85	0.65
(80,100]	124.4	0.74	0.97	0.46	0.54	0.03	1.57	1.25	0.85	0.09

Table 3: Table of sample sizes, survival Brier scores, confusion matrices, and rate parities for demographic subgroups for QRISK3.

	QDiabetes									
	Sample size	Survival brier score	TPR	FPR	TNR	FNR	TPR parity	FPR parity	TNR parity	FNR parity
Overall	458700	0.90	0.59	0.24	0.76	0.41	1.00	1.00	1.00	1.00
Sex										
Male	254758.0	0.90	0.65	0.26	0.74	0.35	Ref	Ref	Ref	Ref
Female	203942.0	0.91	0.54	0.22	0.78	0.46	0.84	0.86	1.05	1.30
Age group										
[35,45]	60960.0	0.95	0.48	0.11	0.89	0.52	Ref	Ref	Ref	Ref
(45,55]	136384.0	0.92	0.58	0.19	0.81	0.42	1.21	1.72	0.91	0.81
(55,65]	196067.0	0.89	0.61	0.29	0.71	0.39	1.29	2.56	0.80	0.74
(65,75]	65289.0	0.87	0.60	0.33	0.67	0.40	1.27	2.93	0.75	0.76
Ethnicity										
White	434852.4	0.91	0.58	0.23	0.77	0.42	Ref	Ref	Ref	Ref
Mixed	2671.6	0.85	0.65	0.39	0.61	0.35	1.12	1.67	0.80	0.84
Other ethnic group	3849.6	0.83	0.78	0.44	0.56	0.22	1.34	1.89	0.73	0.53
Prefer not to answer	1455.2	0.82	0.81	0.47	0.53	0.19	1.40	2.01	0.69	0.45
Chinese	1406.0	0.89	0.54	0.30	0.70	0.46	0.93	1.30	0.91	1.10
Black or Black British	6748.6	0.79	0.77	0.52	0.48	0.23	1.33	2.24	0.62	0.54
Asian or Asian British	7532.6	0.85	0.61	0.36	0.64	0.39	1.06	1.53	0.84	0.92
Do not know	184.0	0.83	0.70	0.46	0.54	0.30	1.21	1.97	0.71	0.72
Immigration status										
No	419853.0	0.90	0.58	0.24	0.76	0.42	Ref	Ref	Ref	Ref
Yes	38847.0	0.88	0.65	0.31	0.69	0.35	1.11	1.29	0.91	0.85
Average household income										
Greater than 100,000	21976.0	0.93	0.51	0.16	0.84	0.49	Ref	Ref	Ref	Ref
52,000 to 100,000	81948.0	0.93	0.55	0.17	0.83	0.45	1.08	1.08	0.98	0.92
31,000 to 51,999	103959.8	0.91	0.53	0.21	0.79	0.47	1.05	1.32	0.94	0.95
18,000 to 30,999	99476.4	0.90	0.58	0.26	0.74	0.42	1.14	1.65	0.88	0.85
Less than 18,000	86526.4	0.88	0.62	0.32	0.68	0.38	1.23	2.05	0.80	0.76
Prefer not to answer	45778.0	0.89	0.61	0.28	0.72	0.39	1.20	1.75	0.86	0.79
Do not know	19035.4	0.87	0.65	0.33	0.67	0.35	1.28	2.07	0.80	0.72
Highest education level										
College or University degree	151750.8	0.92	0.56	0.19	0.81	0.44	Ref	Ref	Ref	Ref
NVQ or HND or HNC or equivalent	29505.2	0.89	0.59	0.27	0.73	0.41	1.07	1.42	0.90	0.91
Other professional qualifications eg: nursing, teaching	23437.4	0.89	0.62	0.28	0.72	0.38	1.11	1.45	0.89	0.86
A levels/AS levels or equivalent	51475.4	0.91	0.53	0.21	0.79	0.47	0.95	1.10	0.98	1.06
O levels/GCSEs or equivalent	97356.8	0.90	0.57	0.24	0.76	0.43	1.03	1.25	0.94	0.96
CSEs or equivalent	25010.2	0.91	0.58	0.23	0.77	0.42	1.04	1.19	0.95	0.95
None of the above	75393.0	0.87	0.63	0.34	0.66	0.37	1.14	1.78	0.81	0.82
Prefer not to answer	4771.2	0.87	0.63	0.33	0.67	0.37	1.14	1.71	0.83	0.82
Townsend deprivation score (quantiles)										
[-7,-4]	88168.0	0.92	0.53	0.19	0.81	0.47	Ref	Ref	Ref	Ref
(-4,-2]	153310.0	0.91	0.55	0.22	0.78	0.45	1.04	1.16	0.96	0.96
(-2,1]	119538.0	0.90	0.59	0.25	0.75	0.41	1.11	1.33	0.92	0.88
(1,11]	97684.0	0.88	0.65	0.32	0.68	0.35	1.23	1.69	0.84	0.74
Index of multiple deprivation										
[0,20]	320227.4	0.91	0.56	0.22	0.78	0.44	Ref	Ref	Ref	Ref
(20,40]	100400.4	0.89	0.64	0.29	0.71	0.36	1.14	1.33	0.91	0.82
(40,60]	31428.0	0.87	0.64	0.34	0.66	0.36	1.16	1.59	0.84	0.80
(60,80]	6473.2	0.86	0.62	0.37	0.63	0.38	1.11	1.71	0.80	0.86
(80,100]	171.0	0.84	0.67	0.40	0.60	0.33	1.21	1.85	0.76	0.74

Table 4: Table of sample sizes, survival Brier scores, confusion matrices, and rate parities for demographic subgroups for QDiabetes.

in model calibration across subgroups. Together, these measures provide a more holistic picture of performance disparities.

Models trained on biased datasets or without explicitly addressing data biases are likely to perpetuate and amplify existing inequities in clinical decision-making [6][37][8]. Prior research suggests that incorporating social determinants of health (SDOH) into clinical prognostic prediction models can enhance performance for specific population subgroups [38]. QPrediction models integrate certain SDOH, such as area-level deprivation, measured using the Townsend deprivation score, and self-reported ethnicity. However, the representation of these variables is often overly simplistic and unbalanced. These simplistic categorizations obscure the complexity of social and cultural identities, raising concerns about whether such representations meaningfully enhance fairness, even when explicitly included in risk models. Despite these limitations, we hypothesize that the inclusion of ethnicity and deprivation in the QPrediction models may have contributed to more consistent calibration across subgroups defined by protected attributes, which are themselves often correlated, such as ethnicity and socioeconomic status. This could partially explain why we observed relatively small differences in calibration curves across subgroups. Furthermore, the original QPrediction derivation cohorts included a higher proportion of participants from ethnic minority backgrounds compared to our validation cohorts drawn from the UK Biobank. This greater diversity in the development datasets may have supported more equitable model performance, particularly in terms of calibration, across population subgroups.

While calibration was relatively similar across subgroups, it should be emphasized that we observed overprediction for all subgroups in our analysis. Given that QPrediction models are applied in the clinic, one explanation for the observed overprediction is that some participants may have previously received preventive care as a result of QPrediction risk scores, systematically lowering the actual event rates. Similarly, the prevented adverse outcomes have most likely caused the increase in FPRs as well as inflated the net interventions avoided.

Despite similar calibration across population subgroups, we observed substantial disparities in FPR and FNR. Among women, for example, the FPR for QRISK3 was approximately 40% lower than that for men. False negatives, on the other hand, were twice as high in women compared to men. Similar, though less pronounced, patterns were observed in the QDiabetes model. These results suggest that women were less likely to be incorrectly flagged as high-risk, but more likely to be incorrectly assessed as low-risk. This phenomenon may result in the undertreatment of women, as true cases may go unrecognized.

For QRISK3, we found that older individuals, those with lower income levels, and those with lower educational attainment had higher FPRs and lower FNRs. These trends were largely mirrored in QDiabetes. For income, disparities in FPR larger than 50% were common in the observed data. Patterns for ethnicity were different between QRISK3 and QDiabetes. In QRISK3, individuals who self-identified as non-White experienced higher FNRs and lower FPRs, meaning the model was more likely to miss true cases among minority groups. Conversely, QDiabetes showed the opposite pattern: higher FPRs and lower FNRs among non-White groups, suggesting a tendency to overpredict risk in these populations, similar to another study predicting undiagnosed hypertension [39]. When analyzing immigration status, QRISK3 showed parity, with no major differences in error rates between immigrants and non-immigrants. However, in QDiabetes, immigrants had approximately 30% higher FPR, suggesting a greater likelihood of being incorrectly flagged as high-risk.

For QRISK3, both FPR and FNR were relatively stable across quintiles of the Townsend deprivation score, suggesting that including this variable in the model may have helped mitigate socioeconomic disparities in classification errors. In contrast, QDiabetes showed similar FNRs across deprivation quintiles, but FPRs were higher among individuals with greater deprivation, indicating potential over-prediction in more socioeconomically disadvantaged groups.

The observation that calibration was consistent across subgroups, while FPR and FNR differed substantially, is not surprising and is well-explained by the fairness impossibility theorem [40][41]. This theorem demonstrates that in the presence of unequal base rates (i.e., different outcome prevalences across groups), it is mathematically impossible to simultaneously satisfy calibration and particular discrimination metrics. In other words, ensuring that predicted probabilities are accurate for each group will inherently lead to disparities in error rates. A well-known example of this trade-off can be observed in ProPublica's investigation of the COMPAS risk assessment tool, where the model was found to be similarly calibrated for Black and White defendants but exhibited stark imbalances in FPR and FNR across racial groups [42].

We highlight that while expanding datasets to be more representative of population subgroups and incorporating SDOH into model development can improve predictive performance and potentially advance equity [10], such technical solutions are ultimately limited. Algorithmic bias and unfairness [43] often reflect underlying structural inequities that are embedded and also extend beyond the healthcare system. Even when models like the QPrediction algorithms integrate variables such as area-level deprivation or self-reported ethnicity, the models are typically deployed in a clinical setting to guide individual-level decision making, such as initiating preventive medications. This can inadvertently shift responsibility (and perhaps blame) for health outcomes onto patients, reinforcing individualistic explanations for disease while obscuring the structural and societal forces, such as racism, poverty, and unequal access to care, that shape health risks in the first place [44].

The validity of our results relies on the assumption that the variables and outcomes in the UK Biobank accurately reflect real-world events and characteristics. However, several forms of data bias must be considered; rather than enlisting these as analytical limitations, as is usually done in scientific literature, we present these points as key discussion points here. Outcome ascertainment in the Biobank relies on electronic health records and linked data, which may under-capture certain conditions, particularly in groups with lower healthcare utilization [45]. Moreover, variables were measured only at baseline, while many social and economic conditions, such as employment status, neighborhood deprivation, and even self-identified ethnicity, can shift over time [46]. The long follow-up periods for events mean that important social changes occurring between baseline and event onset are not captured, potentially affecting our analysis. Furthermore, self-reported data is prone to systematic reporting errors, which might further decrease the validity of the findings [47].

Ethnicity, in particular, poses complex challenges [48][49], and while the incorporation of these constructs has been shown to reduce algorithmic bias [50], their use is still controversial for multiple reasons. This study, like many others, relied on self-reported ethnic background. While self-identification is a valid and commonly used method, it is ambiguous and subject to both individual interpretation and broader social dynamics. Research has shown that self-assigned and socially-assigned race or ethnicity can diverge significantly [51], with meaningful implications for health outcomes and discrimination. The categories available in the UK Biobank and the QPrediction models are coarse and rooted in geographic or national labels. For instance, broad racial labels like "White" are categorized alongside specific nationalities such as "Indian" and "Pakistani", and national identities such as "British" and "Irish" are grouped at the same level as "African", which encompasses an entire continent that is the most genetically diverse in the world [52]. While the latter phenomenon is likely due to practical considerations related to sample size, it is undeniable that grouping diverse populations under umbrella labels conceals meaningful heterogeneity in culture, biology, and lived experience. It also remains unclear whether these categories are intended to reflect shared genetics, cultural heritage, or shared experiences of being racialized, and how histories of colonization, migration, or social stratification influence these identities. Due to the crudeness of the ethnic categories, further research is needed to assess whether the inclusion of ethnicity in the models is actually useful for minoritized groups.

In the EU alone, health inequalities account for 20% of total healthcare costs and related welfare losses amount to nearly 1 trillion EUR per year [53]. More importantly, the amount of avoidable health inequalities caused by structural inequities [54] cannot be adequately quantified. Despite the potential for reducing health(care) inequalities through the use of bias mitigation techniques, we argue that it is unlikely that algorithmic bias mitigation strategies, alone, would be adequate in addressing the structural drivers of health disparities rooted in racial capitalism, colonialism, and patriarchy [55][56][57]. Closing the inequity gap is a self-evident moral imperative: as a WHO report from 2008 puts it, currently observed health inequalities linked to social disadvantage "should simply never happen" [54].

5 Acknowledgements

No generative AI tools have been used in the creation of this article.

6 Conflicts of Interest

None declared.

7 Funding

The Copenhagen Health Complexity center is funded by TrygFonden. IH, DU, and TVV are supported by the 'Data Science Investigator - Emerging 2022' grant from Novo Nordisk Foundation (NNF22OC0075284) and the Department of Public Health (University of Copenhagen). AK acknowledges funding from the Novo Nordisk Foundation via The Novo Nordisk Young Investigator Award (NNF20OC0059309).

8 CRediT Author Statement

The contributors are listed alphabetically.
Conceptualization: IH, DU, TVV
Methodology: IH, AK, TLN, DU, TVV
Software: IH
Validation: IH
Formal analysis: IH
Investigation: IH, TVV
Resources: TVV
Data Curation: IH
Writing - Original Draft: IH, DU, TVV
Writing - Review & Editing: IH, AK, TLN, DU, TVV
Visualization: IH
Supervision: TLN, DU, TVV
Project administration: TVV
Funding acquisition: TVV

References

[1] Johanna A. A. G. Damen et al. "Prediction models for cardiovascular disease risk in the general population: systematic review". eng. In: *BMJ (Clinical research ed.)* 353 (May 2016), p. i2416. ISSN: 1756-1833. DOI: 10.1136/bmj.i2416.

[2] Yue Cai et al. "Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: a systematic review". In: *BMC Medicine* 22 (Feb. 2024), p. 56. ISSN: 1741-7015. DOI: 10.1186/s12916-024-03273-7.

[3] *Recommendations — Cardiovascular disease: risk assessment and reduction, including lipid modification — Guidance — NICE*. eng. Publisher: NICE. Dec. 2023. URL: https://www.nice.org.uk/guidance/ng238/chapter/Recommendations#statins-for-primary-prevention-of-cardiovascular-disease.

[4] Dansk Selskab for Almen Medicin. *Iskæmisk hjerte-kar-sygdom*. da. 2025. URL: https://www.dsam.dk/vejledninger/hjerte/risikovurdering.

[5] Donna K. Arnett et al. "2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines". In: *Circulation* 140.11 (Sept. 2019). Publisher: American Heart Association, e596–e646. DOI: 10.1161/CIR.0000000000000678.

[6] Michael W. Sjoding et al. "Racial Bias in Pulse Oximetry Measurement". In: *New England Journal of Medicine* 383.25 (Dec. 2020). Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMc2029240, pp. 2477–2478. ISSN: 0028-4793. DOI: 10.1056/NEJMc2029240.

[7] Andrius Vabalas et al. "Deep learning-based prediction of one-year mortality in Finland is an accurate but unfair aging marker". en. In: *Nature Aging* 4.7 (July 2024). Publisher: Nature Publishing Group, pp. 1014–1027. ISSN: 2662-8465. DOI: 10.1038/s43587-024-00657-5.

[8] Ziad Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (Oct. 2019). Publisher: American Association for the Advancement of Science, pp. 447–453. DOI: 10.1126/science.aax2342.

[9] Héléne T. Cronjé et al. "Assessing racial bias in type 2 diabetes risk prediction algorithms". en. In: *PLOS Global Public Health* 3.5 (May 2023). Publisher: Public Library of Science, e0001556. ISSN: 2767-3375. DOI: 10.1371/journal.pgph.0001556.

[10] Tibor V Varga. "Algorithmic fairness in cardiovascular disease risk prediction: overcoming inequalities". en. In: *Open Heart* 10.2 (Nov. 2023), e002395. ISSN: 2053-3624. DOI: 10.1136/openhrt-2023-002395.

[11] Natalia Norori et al. "Addressing bias in big data and AI for health care: A call for open science". In: *Patterns* 2.10 (Oct. 2021), p. 100347. ISSN: 2666-3899. DOI: 10.1016/j.patter.2021.100347.

[12] Marshall H. Chin et al. "Guiding Principles to Address the Impact of Algorithm Bias on Racial and Ethnic Disparities in Health and Health Care". In: *JAMA Network Open* 6.12 (Dec. 2023), e2345050. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2023.45050.

[13] Gary S. Collins et al. "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods". en. In: *BMJ* 385 (Apr. 2024). Publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting, e078378. ISSN: 1756-1833. DOI: 10.1136/bmj-2023-078378.

[14] 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". en. In: *BMJ* 357 (May 2017). Publisher: British Medical Journal Publishing Group Section: Research, j2099. ISSN: 1756-1833. DOI: 10.1136/bmj.j2099.

[15] Julia Hippisley-Cox and Carol Coupland. "Development and validation of QDiabetes-2018 risk prediction algorithm to estimate future risk of type 2 diabetes: cohort study". en. In: *BMJ* 359 (Nov. 2017). Publisher: British Medical Journal Publishing Group Section: Research, j5019. ISSN: 0959-8138, 1756-1833. DOI: 10.1136/bmj.j5019.

[16] Julia Hippisley-Cox, Carol Coupland, and Peter Brindle. "Derivation and validation of QStroke score for predicting risk of ischaemic stroke in primary care and comparison with other risk scores: a prospective open cohort study". In: *The BMJ* 346 (May 2013), f2573. ISSN: 0959-8138. DOI: 10.1136/bmj.f2573.

[17] Evangelos Kontopantelis et al. "Spatial distribution of clinical computer systems in primary care in England in 2016 and implications for primary care electronic medical record databases: a cross-sectional population study". In: *BMJ Open* 8.2 (Feb. 2018), e020738. ISSN: 2044-6055. DOI: 10.1136/bmjopen-2017-020738.

[18] *EMIS - Wales*. en-GB. June 2025. URL: https://www.emishealth.com/news-insights/practices-work-smarter-not-harder-across-wales.

[19] *EMIS - Scotland*. en-GB. June 2025. URL: https://www.emishealth.com/emis-pcs.

[20] *Recommendations — Type 2 diabetes: prevention in people at high risk — Guidance — NICE*. eng. Publisher: NICE. July 2012. URL: https://www.nice.org.uk/guidance/PH38/chapter/Recommendations.

[21] NHS Health Check. *National guidance for NHS Health Check programme*. URL: https://www.healthcheck.nhs.uk/commissioners-and-providers/national-guidance/.

[22] Julia Hippisley-Cox, Carol Coupland, and Peter Brindle. "The performance of seven QPrediction risk scores in an independent external sample of patients from general practice: a validation study". en. In: *BMJ Open* 4.8 (Aug. 2014). Publisher: British Medical Journal Publishing Group Section: Epidemiology, e005809. ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2014-005809.

[23] G. S. Collins and D. G. Altman. "External validation of QDSCORE® for predicting the 10-year risk of developing Type 2 diabetes". en. In: *Diabetic Medicine* 28.5 (2011). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1464-5491.2011.03237.x, pp. 599–607. ISSN: 1464-5491. DOI: 10.1111/j.1464-5491.2011.03237.x.

[24] Bruce Guthrie et al. "The implications of competing risks and direct treatment disutility in cardiovascular disease and osteoporotic fracture: risk prediction and cost effectiveness analysis". en. In: *Health and Social Care Delivery Research* 12.4 (Feb. 2024), pp. 1–275. ISSN: ISSN: 2755-0079. DOI: 10.3310/KLTR7714.

[25] Shona Livingstone et al. "External validation of the QLifetime cardiovascular risk prediction tool: population cohort study". In: *BMC Cardiovascular Disorders* 23.1 (Apr. 2023), p. 194. ISSN: 1471-2261. DOI: 10.1186/s12872-023-03209-8.

[26] Cathie Sudlow et al. "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age". en. In: *PLOS Medicine* 12.3 (Mar. 2015). Publisher: Public Library of Science, e1001779. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1001779.

[27] Ruth E. Parsons et al. "Independent external validation of the QRISK3 cardiovascular disease risk prediction model using UK Biobank". en. In: (Nov. 2023). Publisher: BMJ Publishing Group Ltd and British Cardiovascular Society Section: Cardiac risk factors and prevention. DOI: 10.1136/heartjnl-2022-321231.

[28] UK Biobank. *Data providers and dates*. Mar. 2025. URL: https://biobank.ndph.ox.ac.uk/showcase/exinfo.cgi?src=Data_providers_and_dates.

[29] Stef van Buuren and Karin Groothuis-Oudshoorn. "mice: Multivariate Imputation by Chained Equations in R". en. In: *Journal of Statistical Software* 45 (Dec. 2011), pp. 1–67. ISSN: 1548-7660. DOI: 10.18637/jss.v045.i03.

[30] Ian R. White, Patrick Royston, and Angela M. Wood. "Multiple imputation using chained equations: Issues and guidance for practice". en. In: *Statistics in Medicine* 30.4 (2011). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4067, pp. 377–399. ISSN: 1097-0258. DOI: 10.1002/sim.4067.

[31] Donald B. Rubin. "Introduction". en. In: *Multiple Imputation for Nonresponse in Surveys*. Section: 1 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470316696.ch1. John Wiley & Sons, Ltd, 1987, pp. 1–26. ISBN: 978-0-470-31669-6. DOI: 10.1002/9780470316696.ch1.

[32] Shona Livingstone et al. "Effect of competing mortality risks on predictive performance of the QRISK3 cardiovascular risk prediction tool in older people and those with comorbidity: external validation population cohort study". English. In: *The Lancet Healthy Longevity* 2.6 (June 2021). Publisher: Elsevier, e352–e361. ISSN: 2666-7568. DOI: 10.1016/S2666-7568(21)00088-X.

[33] Maarten Coemans et al. "Bias by censoring for competing events in survival analysis". en. In: *BMJ* 378 (Sept. 2022). Publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting, e071349. ISSN: 1756-1833. DOI: 10.1136/bmj-2022-071349.

[34] N. van Geloven et al. "Correcting for the dependent competing risk of treatment using inverse probability of censoring weighting and copulas in the estimation of natural conception chances". en. In: *Statistics in Medicine* 33.26 (2014), pp. 4671–4680. ISSN: 1097-0258. DOI: 10.1002/sim.6280.

[35] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. en-US. The MIT Press, Dec. 2023. ISBN: 978-0-262-04861-3. URL: https://mitpress.mit.edu/9780262048613/fairness-and-machine-learning/.

[36] Andrew J. Vickers and Elena B. Elkin. "Decision Curve Analysis: A Novel Method for Evaluating Prediction Models". EN. In: *Medical Decision Making* 26.6 (Nov. 2006). Publisher: SAGE Publications Inc STM, pp. 565–574. ISSN: 0272-989X. DOI: 10.1177/0272989X06295361.

[37] Christina Hunter Chapman et al. "Identifying Equitable Screening Mammography Strategies for Black Women in the United States Using Simulation Modeling". In: *Annals of Internal Medicine* 174.12 (Dec. 2021). Publisher: American College of Physicians, pp. 1637–1646. ISSN: 0003-4819. DOI: 10.7326/M20-6506.

[38] Joshua A. N. van Apeldoorn et al. "Adding ethnicity to cardiovascular risk prediction: External validation and model updating of SCORE2 using data from the HELIUS population cohort". eng. In: *International Journal of Cardiology* 417 (Dec. 2024), p. 132525. ISSN: 1874-1754. DOI: 10.1016/j.ijcard.2024.132525.

[39] Yang Hu et al. "Accounting for racial bias and social determinants of health in a model of hypertension control". eng. In: *BMC medical informatics and decision making* 25.1 (Feb. 2025), p. 53. ISSN: 1472-6947. DOI: 10.1186/s12911-025-02873-4.

[40] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. *Inherent Trade-Offs in the Fair Determination of Risk Scores*. arXiv:1609.05807 [cs]. Nov. 2016. DOI: 10.48550/arXiv.1609.05807.

[41] Alvin Rajkomar et al. "Ensuring Fairness in Machine Learning to Advance Health Equity". In: *Annals of internal medicine* 169.12 (Dec. 2018), pp. 866–872. ISSN: 0003-4819. DOI: 10.7326/M18-1990.

[42] *Machine Bias — ProPublica*. May 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[43] Jessica K. Paulus and David M. Kent. "Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities". en. In: *npj Digital Medicine* 3.1 (July 2020). Publisher: Nature Publishing Group, pp. 1–8. ISSN: 2398-6352. DOI: 10.1038/s41746-020-0304-9.

[44] Paris B Adkins-Jackson et al. "Measuring Structural Racism: A Guide for Epidemiologists and Other Health Researchers". In: *American Journal of Epidemiology* 191.4 (Sept. 2021), pp. 539–547. ISSN: 0002-9262. DOI: 10.1093/aje/kwab239.

[45] Emily Bassett et al. "Inconsistency in UK Biobank Event Definitions From Different Data Sources and Its Impact on Bias and Generalizability: A Case Study of Venous Thromboembolism". In: *American Journal of Epidemiology* 193.5 (May 2024), pp. 787–797. ISSN: 0002-9262. DOI: 10.1093/aje/kwad232.

[46] Mary E. Campbell, Jenifer L. Bratter, and Wendy D. Roth. "Measuring the Diverging Components of Race: An Introduction". EN. In: *American Behavioral Scientist* 60.4 (Apr. 2016). Publisher: SAGE Publications Inc, pp. 381–389. ISSN: 0002-7642. DOI: 10.1177/0002764215613381.

[47] Tabea Schoeler, Jean-Baptiste Pingault, and Zoltán Kutalik. "The impact of self-report inaccuracy in the UK Biobank and its interplay with selective participation". en. In: *Nature Human Behaviour* 9.3 (Mar. 2025). Publisher: Nature Publishing Group, pp. 584–594. ISSN: 2397-3374. DOI: 10.1038/s41562-024-02061-w.

[48] Darshali A. Vyas, Leo G. Eisenstein, and David S. Jones. "Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms". In: *New England Journal of Medicine* 383.9 (Aug. 2020), pp. 874–882. ISSN: 0028-4793. DOI: 10.1056/NEJMms2004740.

[49] Shazia M. Siddique et al. "Critical Appraisal for Racial and Ethnic Equity in Clinical Prediction Models Extension: Development of a Critical Appraisal Tool Extension to Assess Racial and Ethnic Equity-Related Risk of Bias for Clinical Prediction Models". In: *Health Equity* 7.1 (Nov. 2023), pp. 773–781. ISSN: 2473-1242. DOI: 10.1089/heq.2023.0035.

[50] Sara Khor et al. "Racial and Ethnic Bias in Risk Prediction Models for Colorectal Cancer Recurrence When Race and Ethnicity Are Omitted as Predictors". eng. In: *JAMA network open* 6.6 (June 2023), e2318495. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2023.18495.

[51] Kellee White et al. "Socially-assigned race and health: a scoping review with global implications for population health equity". In: *International Journal for Equity in Health* 19 (Feb. 2020), p. 25. ISSN: 1475-9276. DOI: 10.1186/s12939-020-1137-5.

[52] "A global reference for human genetic variation". In: *Nature* 526.7571 (2015), pp. 68–74. ISSN: 0028-0836. DOI: 10.1038/nature15393.

[53] Johan P. Mackenbach, Willem J. Meerding, and Anton E. Kunst. "Economic costs of health inequalities in the European Union". en. In: *Journal of Epidemiology & Community Health* 65.5 (May 2011). Publisher: BMJ Publishing Group Ltd Section: Research report, pp. 412–419. ISSN: 0143-005X, 1470-2738. DOI: 10.1136/jech.2010.112680.

[54] World Health Ogranization. *Closing the gap in a generation: health equity through action on the social determinants of health - Final report of the commission on social determinants of health*. en. URL: https://www.who.int/publications/i/item/WHO-IER-CSDH-08.1.

[55] Vanessa Machado de Oliveira. *Hospicing Modernity: Facing Humanity's Wrongs and the Implications for Social Activism*. en. Google-Books-ID: 4noQEAAAQBAJ. North Atlantic Books, Sept. 2021. ISBN: 978-1-62317-625-9.

[56] Ali Alkhatib. "To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–9. ISBN: 978-1-4503-8096-6. DOI: 10.1145/3411764.3445740.

[57] Bob Jessop. "The Crisis of the National Spatio-Temporal Fix and the Tendential Ecological Dominance of Globalizing Capitalism". en. In: *International Journal of Urban and Regional Research* 24.2 (2000). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-2427.00251, pp. 323–360. ISSN: 1468-2427. DOI: 10.1111/1468-2427.00251.