ARTICLE   **OPEN**

Representational ethical model calibration

Robert Carruthers[1✉], Isabel Straw[2], James K. Ruffle[2], Daniel Herron[3], Amy Nelson[2], Danilo Bzdok[4], Delmiro Fernandez-Reyes[1], Geraint Rees[2] and Parashkev Nachev[2✉]

Equity is widely held to be fundamental to the ethics of healthcare. In the context of clinical decision-making, it rests on the comparative fidelity of the intelligence – evidence-based or intuitive – guiding the management of each individual patient. Though brought to recent attention by the individuating power of contemporary machine learning, such epistemic equity arises in the context of any decision guidance, whether traditional or innovative. Yet no general framework for its quantification, let alone assurance, currently exists. Here we formulate epistemic equity in terms of model fidelity evaluated over learnt multidimensional representations of identity crafted to maximise the captured diversity of the population, introducing a comprehensive framework for *Representational Ethical Model Calibration*. We demonstrate the use of the framework on large-scale multimodal data from UK Biobank to derive diverse representations of the population, quantify model performance, and institute responsive remediation. We offer our approach as a principled solution to quantifying and assuring epistemic equity in healthcare, with applications across the research, clinical, and regulatory domains.

INTRODUCTION

Medicine has always been personal, concerned with the individual patient whose specific complaint the physician is asked to address. Under pressure to render the underlying intelligence explicit, objective, replicable, and cumulative, evidence-based medicine has shifted the focus to large populations, guiding clinical management by the parameters of simple statistical models that discard individual variation as noise[1,2]. The resultant gain in population-level fidelity may or may not be associated with a loss at the individual level: we do not know, because the studies that generate care policies and those that evaluate them adopt the same inferential approach.

This blind spot extends to *equity* of care: the universal obligation, rooted in Aristotle's notion of *epieikeia*[3], to seek the best achievable outcome for each individual patient. The care recommended by a model is inequitable as far as it fails to fulfil a patient's individual potential for recovery and health[4]. In neglecting individuality, conventional evidence-based medicine conceals our success or failure in maintaining such equity. This defect is all the more important for being *epistemic*, of the upstream knowledge from which all downstream clinical action derives.

Until the advent of machine learning, this major ethical problem had no obvious remedy. But it is now clear that richly expressive models of high-dimensional data can characterize populations with greater fidelity to the individual[5–8]. Whether implicitly or explicitly, such models describe patients in terms of more closely individuating *subpopulations* identified by multiple interacting characteristics, whose distinct structure may be directly material to clinical care, interfere with our ability to determine its optimal form, or both. By revealing differences between subpopulations, machine learning casts a brighter light on epistemic equity than crude population descriptions could provide, and enables us to pursue our deep moral obligation to assure it.

Though a matter of intense study in other domains, there is no accepted framework for defining, diagnosing, or quantifying epistemic equity arising in the guidance of clinical care by algorithmic models, whether simple or complex, traditional or novel, and neither regulatory nor professional bodies currently provide for it. Here we propose such a framework, termed Representational Ethical Model Calibration.

Any quantitative framework here must operationalise the notions of the epistemic equity of a model and the descriptive identity of a patient. We define the former as equal maximisation of model fidelity across the population: where the available knowledge is plausibly invariant, equity means equality; where it varies under some external constraint outside our power to address, equity means equal departure from the attainable maximum. We define the latter as any set of replicable distinguishing characteristics material to the specific healthcare context. For example, the equity of a classifier for detecting ischaemic injury on a brain scan might be measured by balanced accuracy evaluated as a function of age. Observing systematic variation of accuracy with age raises the possibility of inequitable performance across patients so identified.

Many quantitative indices of model fidelity and its variation exist: the optimal choice will vary with the specific application. The appropriate criteria of identity, however, are not so easy to determine. The use of simple descriptors such as age, sex, and ethnicity, taken in isolation, presupposes that they are sufficiently individuating. But, as we have seen, a patient will typically belong to a distinct, replicable – and therefore learnable – subpopulation defined by the interaction of multiple characteristics. The underperformance of a model in such a subpopulation may not be evident from examination of single characteristics alone. Any principled notion of equity obviously cannot exclude groups whose defining identity eludes simple description. Indeed, there is increasing evidence that it is precisely those falling in the intersectional *faultlines* between traditionally recognized groups that may be most vulnerable[9–13]. Moreover, neither the total number nor the nature of the relevant identifying characteristics

[1]Department of Computer Science, University College London, London, UK. [2]UCL Queen Square Institute of Neurology, University College London, London, UK. [3]Research and Development, NIHR University College London Hospitals Biomedical Research Centre, London, UK. [4]Department of Biomedical Engineering, Faculty of Medicine, McGill University, Montreal, Canada. ✉email: robert.carruthers.20@ucl.ac.uk; p.nachev@ucl.ac.uk

may be limited a priori. If a social, environmental, demographic, physiological, pathological or any other replicable distinguishing characteristic – whether self-assigned or externally measured – has a systematic impact on clinical outcomes, we have a moral duty to examine it. Naturally, the wide descriptive space so defined may not be easily navigable. But we can employ *representation learning*[14] to derive rich yet succinct descriptions of the population that render its diversity surveyable.

Our proposed framework therefore combines the evaluation of model performance against identifying descriptors – ethical model calibration – with the derivation of descriptors through representation learning that optimally capture the diversity of the population. In relating observed to ideal performance, it is kin with statistical model calibration[15]. It enables the epistemic equity of a model to be judged against identities defined as richly and comprehensively as available data allow.

Application of the framework is illustrated in Fig. 1. In brief, the fidelity of a given model, quantified by the metric most suited to its task, is evaluated against a succinct description of the population derived from learnt representations of the same (primary) data or other (secondary) data drawn from the same domain. Systematic differences in performance across the population identify potential inequities, and trigger reme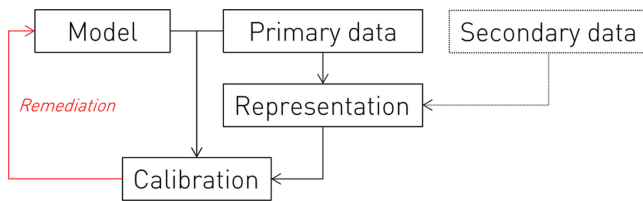diation – action to correct the disparity or limit its downstream impact – by any applicable mechanism, such as acquiring more data, modifying the model, or limiting its application. The cycle of calibration and remediation may be repeated until a result satisfactory on some agreed criterion is obtained. Extending the foregoing example, an ischaemic stroke classifier may be found to perform poorly within a distinct subpopulation with a characteristic age interval and spectrum of co-morbidities. Identifying this subpopulation directs action on the data, the model, or the scope of application, until calibration shows equity has been achieved. Where the representations are based on a generative model of the primary data, it enables immediate remediation by augmenting model retraining with synthetic data from the under-performing subpopulation.

This approach is applicable to any model, whether conventional or machine learning-based, any metric of performance and its disparity, and any method of representation learning[16–19]. It leaves the nature of the remediation open, to be chosen as specific circumstances dictate, and distinguishes remediation from the calibration used to guide it. Here we demonstrate its use, end-to-end, in the context of predicting glycaemic control – as indexed by glycated haemoglobin (HbA1c) concentrations[20] – from large-scale, high-dimensional data in UK Biobank[21]. We choose glycaemic control owing to the importance of glucose intolerance and complex patterns of its susceptibility. We employ deep representation learning based on autoencoders owing to their architectural simplicity and expressivity, and established applications in healthcare[5].

We show how the framework can be used to detect the systematic epistemic inequity of a model with respect to subpopulations concealed by the richness of their identity, and guide remediation in pursuit of more equitable model performance. Although model epistemic equity is only one aspect of equity, itself only one aspect of medical ethics, the position of models at the apex of evidence-based clinical decision-making lends the highest ethical significance to the equity of their performance. Our results are relevant to the domains of quantitative ethics, multidimensional fairness, and the regulation of mathematical models in healthcare[22].
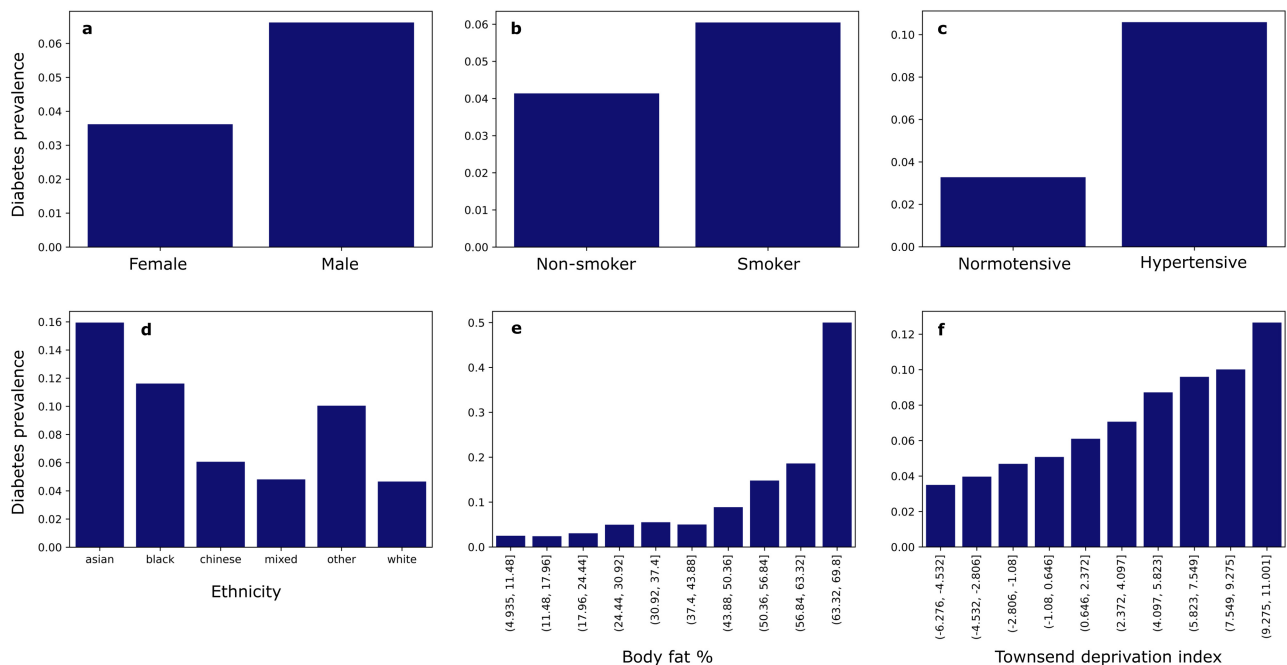

Fig. 1 The Representational Ethical Model Calibration Framework. The fidelity of a candidate model with respect to subpopulations identified by representation learning (performed on either primary or secondary data) is quantified in an ethical calibration step that informs appropriate remedial action, within an iterative process repeated until an agreed criterion of model equity is reached.

Fig. 2 Diabetes prevalence by variable. Higher prevalence was seen in males (**a**), smokers (**b**), those with high blood pressure (**c**), certain ethnicities (**d**), those with higher body fat % (**e**), and the more deprived (**f**).
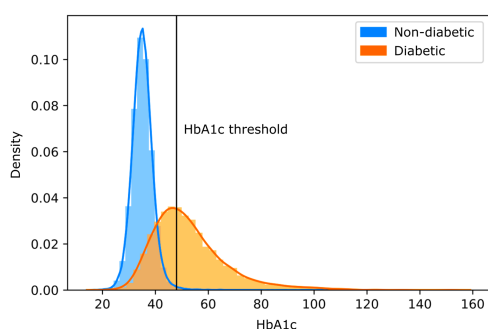
Fig. 3 Relationship of diabetes and glycated haemoglobin (HbA1c). Those without diabetes tended to have HbA1c below the diagnosis threshold of 48, while those with diabetes had a wide range of HbA1c both above and below the threshold.

RESULTS

Associations of impaired glycaemic control

We evaluated a random selection of UK Biobank records split into 150,000 training and 50,000 validation sets, including a range of demographic, social, lifestyle, physiological, and morbidity features potentially relevant to glycaemic control (see Methods). The commonest cause of impaired glycaemic control – diabetes – showed variation with sex, smoking, hypertension, ethnicity, body fat composition, and social deprivation consistent with previous data from populations with a similar age distribution (Fig. 2). A diagnosis of diabetes was associated with higher and more widely dispersed HbA1c, reflecting variable success in the clinical management of the underlying disorder (Fig. 3). These observations suggest predictive models of HbA1c based on this data can be considered representative of a plausible real-world modelling scenario.

Basic ethical model calibration

A regression model based on a conventional fully-connected feed-forward net with three hidden layers, an architecture chosen for its controllable flexibility (see Methods), was evaluated across the population as a whole. Near-identical root mean squared errors were observed on the training and validation sets – 6.099 and 6.097, respectively – corresponding to normalised root mean squared errors (NRMSE) of 0.169 for both. Nonetheless, examination of performance independently stratified by sex, smoking, and deprivation revealed substantial disparities in model fidelity (Table 1), with evident underperformance for men, smokers, and the socially more deprived.

Representational ethical model calibration

To permit the identification of underperformance localised to more complex subpopulations defined by the interactions of multiple factors, we used an autoencoder to embed participants in a two-dimensional latent representational space that compactly described their high-dimensional similarities and differences (see Methods). Labelling the embedding by individual regression errors (Fig. 4b) revealed potentially structured variation in fidelity; labelling it by key descriptive features revealed its organisation (Fig. 4d–i). To facilitate the identification of a tractable number of characteristic subpopulations, the latent space was segmented into fifty groups using a Gaussian Mixture Model (GMM). Examining the five largest groups in the bottom 25th regression performance centile revealed a diversity of patterns of individual features, most of them shown on permutation testing to be significantly different from the rest of the population (Table 2). Crucially, model fidelity varied more widely than across basic features of ethical concern, and reached

Table 1. Stratified model performance to demonstrate inequity. NRMSE is shown across three example variables.

	Sex		Smoking		Deprivation	
	F	M	No	Yes	Low	High
Training	0.152	0.186	0.162	0.176	0.156	0.187
Validation	0.151	0.187	0.161	0.177	0.158	0.185

The deprivation index was binarized into high and low by mean split. The model's overall training NRMSE was 0.169. Consistent performance differences are shown, with the model performing better (lower NRMSE) for women, non-smokers and the less deprived.

lower values, illustrating the need for evaluating feature interactions in the quantification of equity.

Remediation

Revealing the pattern of inequitable performance allows us to target our efforts at remediation. The optimal approach to remedying inequitable models – in healthcare and elsewhere – is the subject of intense study, and will vary circumstantially in feasibility and effectiveness. Here we illustrate only one approach to remediation, focused on model training. The subpopulations exhibiting higher than median NRMSE were designated as under-served. A simple strategy of remediation was then applied, oversampling these subpopulations in model training, generating a very different pattern of performance with less pronounced disparities across the population, but at the cost of reduced fidelity overall (Figs. 5, 6). These observations can be formalised in terms of NRMSE, and standard indices of distributional equality such as the Gini coefficient (Tables 3, 4). Varying the degree of oversampling moved NRMSE and Gini coefficient scores in the expected directions, as shown in Fig. 7. Performance disparities persisted across groups, such as men and women, defined by dimensions other than those selected for remediation. Note our objective here is not to devise or implement an optimal approach to remediation, but to show how the calibration and remediation processes relate. An optimal approach would improve equity without deleterious impact on other groups or overall performance.

DISCUSSION

We have formulated a framework, Representational Ethical Model Calibration, for detecting and quantifying inequity in model performance distributed across subpopulations defined by multiple interacting characteristics. Central to the framework is the use of representation learning to compress and render navigable the high-dimensional space of patient diversity over which equitable performance must be evaluated. We have demonstrated ethical model calibration on large-scale UK-biobank data in the context of a common morbidity – impaired glycaemic control – and a simple, purely illustrative approach to remediation.

Here we examine ten aspects of the conceptualisation, implementation, and application of our approach.

First, if epistemic equity cannot be assumed where, as here, the modelling task is comparatively simple and the data are balanced and abundant, the case for evaluating it explicitly across all models employed in healthcare ought to be hard to resist. The narrow recruitment mechanism makes a test based on UK Biobank conservative, for the underlying heterogeneity is likely to be less pronounced than that observed in clinical reality. Equally, our analysis shows that calibration across familiar, observed, unitary features such as demographics cannot reveal inequity across unfamiliar, latent, composite features accessible only through
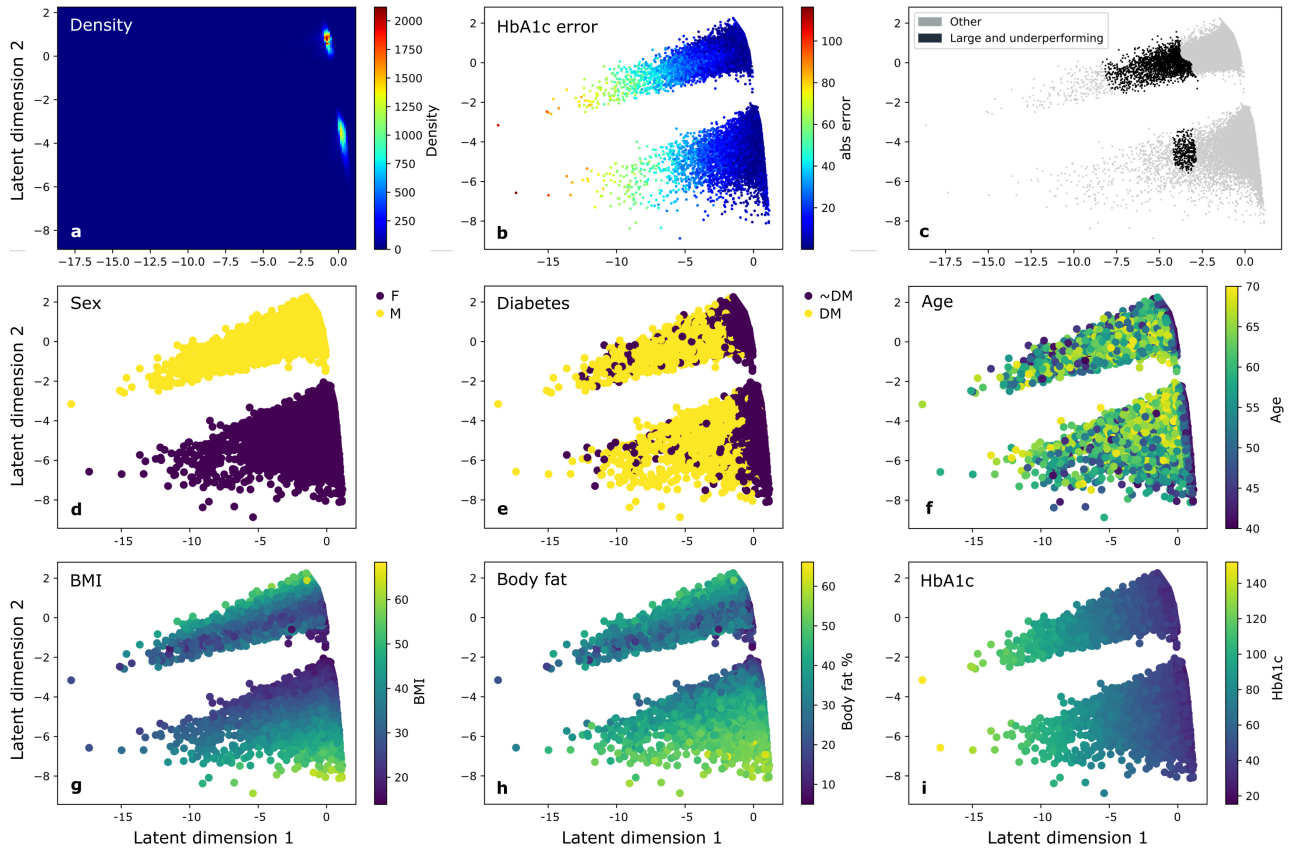
Fig. 4 Two-dimensional latent space. The space is coloured by data density (**a**), model error (**b, c**), and the values of selected variables (**d–i**). The space appears to be dominantly clustered by sex (**d**). The largest groups in the worst 25th performance percentile are shown in (**c**), associated with higher levels of HbA1c (**i**).

Table 2. Permutation test results.

	Group				
	22	2	41	33	23
High Blood Pressure	0.0010	0.0010	0.0010	0.0010	0.0010
Smoking	0.0010	0.0010	0.0010	0.1029	0.0010
BMI	0.0010	0.0010	0.0010	0.0010	0.0010
HbA1c	0.0010	0.0010	0.0010	0.0010	0.0010
Age	0.0010	0.0010	0.0010	0.0010	0.0010
Ethnicity: white	0.0579	0.0010	0.0010	0.0010	0.0010
Townsend Deprivation Index	0.1349	0.0010	0.0010	0.0040	0.0010

P-values below a given significance level indicate that the group's mean was significantly different from the remainder of the data for that variable. Applying the Benjamini-Hochberg procedure to control the False Discovery Rate across multiple tests resulted in 85 significant results out of 110 tests at $\alpha = 0.05$.
Selected permutation test results (p-values) are shown for the five largest groups in the worst 25th performance percentile for prediction.

representation learning. A commitment to ethical model calibration implies a commitment to the representational kind.

Second, the epistemic equity of the models used to guide clinical care – our concern here – is obviously not the only kind of equity healthcare must consider. Unwarranted variation in clinical outcomes may arise from a wide diversity of procedural, cultural, social, economic, political, and regulatory factors that operate outside the realm of evidence-guided practice and need to be addressed independently from it[5,16,22–32]. In focusing on epistemic equity we are not denying the importance of other kinds. But since action proceeds from belief, and belief in medicine strives to be objectively evidential, detecting and quantifying epistemic equity will always be a fundamental concern. Equally, in using the qualifier "ethical", we do not mean to imply that detecting equity, still less epistemic equity, exhausts the ethical realm, only to indicate the *purpose* of the calibration our framework enables, applied to an object – a mathematical model – whose ethical expression is limited to fidelity.

Third, to draw attention to subpopulations defined by the interactions of many or unfamiliar identifying characteristics is not to neglect the importance of familiar single characteristics taken in isolation, still less those already identified as posing a risk to equity. Indeed, equity with respect to a single characteristic may be concealed by disordinal interactions. For example, if compared with men of any age a model both over-performs in young women and under-performs in elderly women, calibration against sex alone will not reveal any difference between men and women at all. Even those interested in a particular characteristic cannot afford to ignore its interactions with others.

Fourth, since the objective is to evaluate the epistemic equity of a model with respect to a description of the population that best exposes it to scrutiny, the choice of modelled characteristics will typically be constrained only by feasibility. Prejudice for or against any given characteristic would undermine the very notion of equity we are seeking to promote. Equally, inequities embedded in the source characteristics themselves may propagate to their representation: a problem mitigated by calibrating with multiple different representations, either constructed from different sets of characteristics, or drawn from different levels of a hierarchical
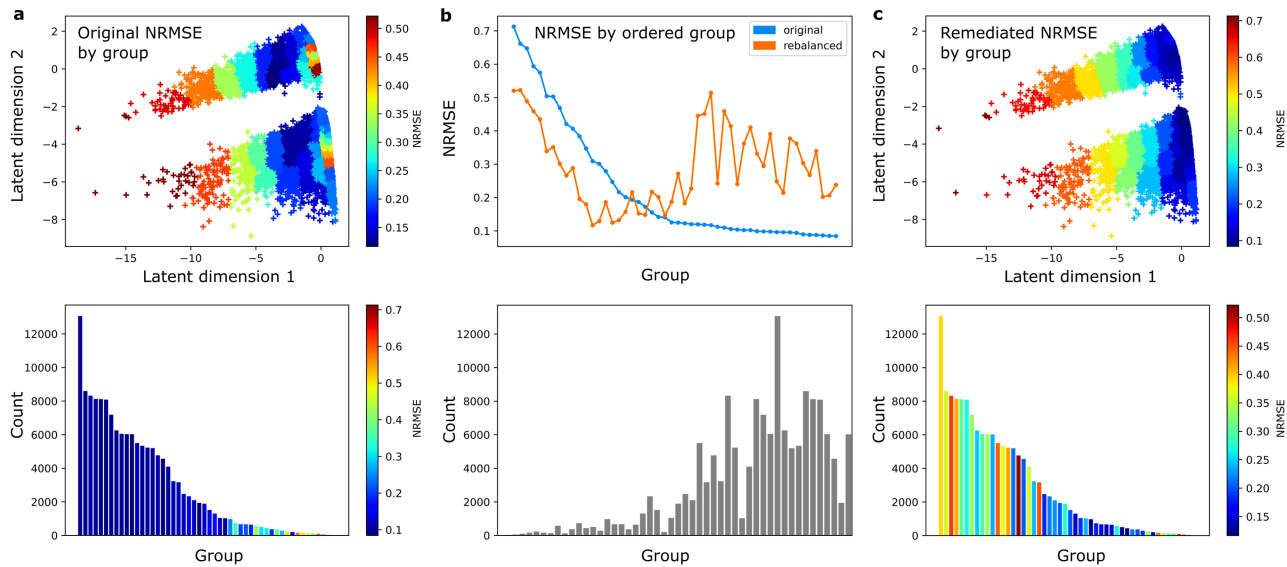
Fig. 5 Model performance by GMM group in the latent space. Performance is shown before (**a**) and after (**c**) remediation. The top panel of (**a**) shows model performance by group, while the bottom panel shows the group counts. The model showed mostly even performance across groups. The top panel of (**b**) shows the effect of remediation. Lower-performing (higher NRMSE) groups show improvements, but the better-performing groups got significantly worse. The bottom panel of (**b**) shows group counts in descending order of the original NRMSE. It can be seen that performance decreases occurred in high-volume groups. The performance distribution worsened overall, shown in (**c**), and this would likely offset any gain in equity.

decomposition of a single set, yielding a multi-scale perspective on equity.

Fifth, although contemporary discourse on model equity is tightly focused on complex models based on machine learning[33–35], the question of epistemic equity arises with any model architecture – simple, complex, transparent or opaque – indeed with intuitive decision-making too. The simple models of traditional evidence-based medicine, crafted in ignorance of population heterogeneity, are not more equitable but merely blind to the violations of equity they may commit. It precisely because complex models seek to ground beliefs about individuals in a local feature space that variations over the population become both surveyable and remediable. But it is also true that greater model flexibility can be associated with greater model fragility, potentially amplifying disparities through subpopulation-specific catastrophic failure[36–40]. In any event, all models – and intuitive decision-making – should be evaluated, for no practice is immune to inequity.

Sixth, the causes of epistemic inequity will vary in their susceptibility to remediation[41–46]. The irreducible, random component of variation may be greater in one subpopulation than another through constitutional differences in the underlying biological processes nothing could possibly equalise, such as those associated with ageing. There may be variations in data representation, class balance, labelling, and noise that are outside the power of healthcare systems to address, imposing practical, circumstantial limits on equity. But that a representation model can identify a distinct subpopulation shows its features are learnable, and that underperformance is therefore potentially addressable, even if it may not be found to be so on subsequent examination. Our framework neither assumes nor accepts differential limits on epistemic quality: its task is to focus attention on where they most need to be examined.

Seventh, none of the components of the framework – the target model, the representation model, the metrics of fidelity and equity, or the approach to remediation – are generally prescribed. It is natural that each should be adapted to the specific task and its circumstances: a strength of the approach is its flexibility. In particular, the expressivity of the representation should be tuned

to the heterogeneity of the data and its learnability under the applicable data and computational regime. Where supported by the regime, it is appropriate to use a finely granular representation; where not, a coarser representation may still capture systematic intersectional effects non-representational calibration would miss. Hierarchically organised representations provide a graceful way of manipulating representational expressivity without the need to retrain the representational model: one simply chooses the most suitable level of the descriptive hierarchy. In general, our innovation is not in the ingredients but in the way in which they are put together to provide a robust, comprehensive solution to the problem of quantifying equity in populations of the heterogeneity likely to obtain in reality.

Eighth, although most naturally derived from the data on which the target model is trained, the representations used in calibration may originate from another source as long as the test data can be mapped onto the same representational space. The task of the representation model – to redescribe the population in a way that makes its heterogeneity legible – does not require the target label, and can therefore be accomplished with larger scale data from elsewhere. This mechanism can even be used in remediation to augment the target data to include patterns of variation derived from another source. For example, one might use learnt patterns of age-related changes in brain morphology to deform an independent set of brain images across a wider range of aged appearances[47].

Ninth, remediation need not be confined to model retraining, but may encompass any action that improves the quality of the decision-making the target model is used to guide, including simply narrowing the applicable scope of a model. Theoretically the most potent action, though perhaps the least discussed in the literature, is acquisition of new data selected by its predicted impact on model equity[48,49]. Just as such active learning may make the decision boundaries of a discriminative model easier to delineate, so it may ensure they are equitably configured. Indeed it would be entirely natural to add an equity constraint to an active learning or sequentially optimised experimental design modelling framework. In general, it should be part of the objective of remediation to attain improvements in equity *without* impact
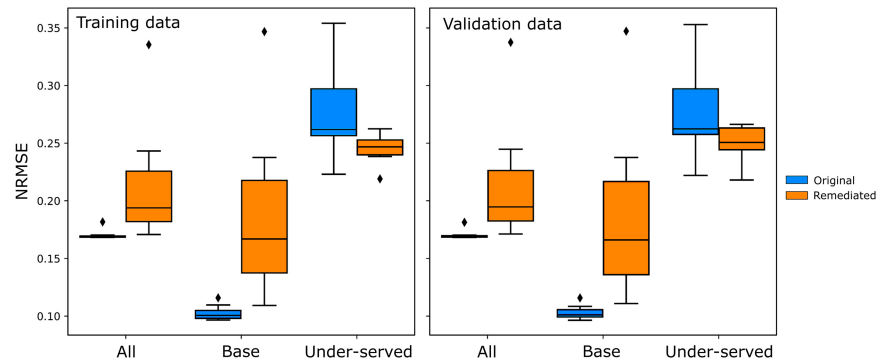
Fig. 6 Effect of remediation. NRMSE is shown for the whole dataset, the base group and the under-served group, before and after remediation, over $n = 10$ trials, on training and validation data. Performance was worse on the under-served group, and this improved after rebalancing. However, there was a high cost in base group performance. See differences in Table 3. The boxplots show the median (centre line), 25th and 75th centiles (box), 1.5 times the interquartile range (whiskers), and outliers (diamonds).

Table 3. Difference in Normalised Root Mean Square Error (NRMSE) before and after rebalancing.

	Mean Difference in NRMSE		Standard Deviation	
	Training	Validation	Training	Validation
All	−0.0411	−0.0420	0.0466	0.0470
Base	−0.0817	−0.0809	0.0666	0.0667
Under-served	0.0315	0.0268	0.0335	0.0300

Mean differences and standard deviations of differences are presented across $n = 10$ trials. Results are presented for the entire dataset, the base group and the under-served group. A positive difference indicates better model performance. It can be seen that base group performance worsened on average while under-served group performance improved. NRMSE values are illustrated in Fig. 6.

Table 4. Gini coefficients. Mean Gini coefficients, differences and standard deviations are presented, before and after rebalancing, across $n = 10$ trials.

	Mean Gini Coefficient		Standard Deviation	
	Training	Validation	Training	Validation
Original	0.1594	0.1595	0.0163	0.0170
Rebalanced	0.1466	0.1470	0.0309	0.0299
Difference	0.0128	0.0124	0.0321	0.0324

A Gini coefficient of zero indicates equality and a Gini coefficient of one indicates maximum inequality. A positive difference can be observed, which indicates a decrease in Gini coefficient on average and hence a more equitable distribution of performance across groups. However, the standard deviations indicate that this behaviour was not consistent.

on other subpopulations or on the population as a whole[47]. Only remediation methods that add information, through additional data or more accurate prior beliefs, could plausibly combine joint improvements in equity and overall fidelity. Since knowledge in healthcare is not a fixed quantity, to be divided more or less evenly across the population, redistributive approaches are less appropriate here than in other domains of activity[33–35,41,50]. This is not a zero-sum game.

Finally, we should recognize the deep union between epistemic equity and the individuation of care: neither is possible without the other. Equity implies maximising our knowledge of the optimal care of a patient, identified as richly as the task demands, up to the practically achievable limit; successful individuation implies having attained that maximum. Both require the flexible, highly expressive models contemporary machine learning has only recently supplied but medicine has always needed.

In summary, our proposed framework enables the assurance of the epistemic equity of any model in healthcare – whether simple of complex – under ethically the most general notion of identity: one defined not merely by demographics but by *any* set of characteristics that define a distinct group, alone or in interaction. Our approach places ethical model calibration on a robust conceptual and algorithmic footing, advancing the application of quantitative ethics to medicine, and promoting equitable clinical care at the highest level: the knowledge it rests on.

METHODS

Dataset

The dataset for this study was drawn from the UK Biobank. The UK Biobank was established as a major prospective study with significant involvement from the UK Medical Research Council and the Wellcome Trust[21], and has become an important open-access resource for medical researchers across the UK and worldwide. The subset of data contained 150,000 records in the training set and 50,000 records in the validation set, each record representing a distinct individual. The variables in the dataset were as follows (UK Biobank field number in brackets): Demographic: Sex (31), Age (33), Smoking (20116), Ethnicity (21000), Townsend Deprivation Index (189); Investigatory: Haemoglobin (30020), Glycated Haemoglobin (30750), Body Mass Index (21001), Weight (21002), Body Fat % (23099); Medical Diagnoses: Diabetes (2443), High blood pressure (6150), Heart Attack / Angina / Stroke (6150), Blood Clot / Emphysema / Lung Clot, Asthma, Hayfever/Rhinitis/Eczema (6152), Other Serious Condition (2473).

The condition of interest in this study was Type II diabetes: globally a leading, and increasing, cause of morbidity and mortality[51], predicted to become the most prevalent condition in the UK Biobank cohort[21]. Diabetes was chosen in order to test the framework with a realistic medical problem, and a selection of variables of ethical interest. It was expected that the diagnosis of diabetes, and by association raised levels of HbA1c, would be predictable from the data. This was known to be plausible due to existing work[52] on diabetes prediction using UK Biobank data, which influenced the variable selection. The dataset was used for the regression task of predicting the level of glycated haemoglobin (HbA1c) using the other variables, excluding the presence of a diabetes diagnosis.

To prepare the data, a small number of records with what appeared to be outlier values of HbA1c were removed. Some other variables of interest were dropped including Income and
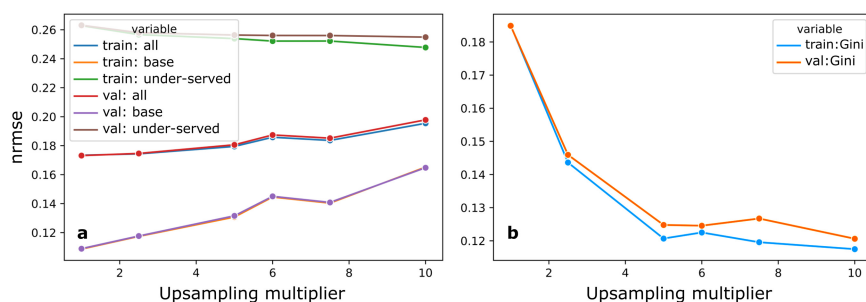
Fig. 7 Effect of upsampling multiplier on performance metrics. Panel (**a**) shows NRMSE for the entire dataset, the base group and the under-served group, for training and validation sets. Panel (**b**) shows the Gini coefficient. In both training and validation sets, increasing the upsampling multiplier improved model performance on the under-served group, while negatively affecting performance on the base group and overall. The Gini coefficient tended to drop as upsampling increased, mostly indicating increased equity in the distribution for higher levels of upsampling.

Forced Expiratory Volume, owing to missing data. The ethnicity codes in the data were also grouped into broader categories for ease of illustration. The diagnoses of heart attack, angina and stroke were combined into one variable due to the format of the original data and low prevalence, as was Blood Clot/Emphysema/Lung Clot and Hayfever/Rhinitis/Eczema for the same reasons. Any records with missing values were removed; the number of records above refers to complete records.

The UK Biobank project was approved by the National Research Ethics Service Committee North West-Haydock (REC reference: 11/NW/0382). An electronic signed consent was obtained from the participants.

Regression model

A standard feed-forward neural network model was used for the regression task. Three hidden layers of neurons were used, with the number of neurons in each layer being (16,16,8), yielding 753 parameters. The ReLU activation function was used at each hidden layer. The mean squared error loss function was used. The model was trained using gradient descent with a batch size of 10 and the Adam optimiser with a learning rate of 0.001, as has been shown to perform well in a variety of contexts[53]. The number of epochs was set following experimentation based on error on the validation set. Variables were normalised before modelling to have a mean of zero and standard deviation of one across participants. The Pytorch library was used.

Model performance was measured using Root Mean Square Error (RMSE) and Normalised RMSE (NRMSE), which is RMSE of the model normalised by the average Root Mean Square Error of the subpopulation. This results in a proportional measure between zero and one. The normalisation was done to allow for different variance between subpopulations. For example, patients with diabetes have a higher variance in glycated haemoglobin (HbA1c), and so a higher RMSE could be expected and is not necessarily indicative of decreased model performance. One disadvantage of (N)RMSE is that it can be affected by outliers.

Existing equity metrics such as Statistical Parity Difference or Equal Opportunity Difference were not used. These tend to be classification-oriented, limited in scope and require the definition of favourable vs unfavourable outcomes. For example, judgements around the relative consequences of a disease being under- or over-diagnosed were not addressed, only that model accuracy differs across the dataset.

Unsupervised models

An autoencoder model was used to in an attempt to uncover new subpopulations in the latent space of the data. An autoencoder consists of a two-part neural network, where the first "encoder" network compresses the data into a low-dimensional space, and the second "decoder" network attempts to reconstruct the original data from the compressed representation. The autoencoder architecture consisted of two hidden layers with (8,4) neurons before a "bottleneck" layer of two neurons (the compressed data) and a mirror-image decoder network. The mean squared error loss was used to calculate reconstruction error. For most experiments, the compressed layer was kept at two neurons for visualisation purposes, even though it was unlikely that all the information could be retained in just two variables given the specific nature of the datasets. The other training hyperparameters were set following a similar procedure to the neural network regression model. Non-binary variables were normalised before modelling.

Given a two-dimensional latent space containing key information about relationships in the dataset, a Gaussian Mixture Model (GMM) was used to identify subpopulations within that space. The GMM is a weighted mixture of multivariate Gaussian probability distributions[54]. It can be used to estimate complex densities due to its flexibility and ability to generalise to high dimensions. Here our interest was to segment the latent space into subpopulations, which could be interpreted as probabilistic clustering where each point is assigned to the Gaussian component with the highest probability. Although GMM is effective as a clustering method, there are limitations on the shape of clusters it can find. In addition, the number of components must be specified. To determine the optimal number of components, silhouette scores were calculated. Silhouette scores are based on a comparison of distances between points both within and between clusters[55]. They are close to one for well-separated clusters and close to zero or negative for poorly-separated clusters. For the latent spaces in this study, the mean silhouette scores tended to decrease as the number of Gaussian components in the GMM increased. This indicates that the clusters identified by the GMMs were not well-separated, as there was no clear distinction between them. Nevertheless, the GMM was still used to segment the latent space for analysis, based on the numerical values of the latent space, which were expected to have some degree of local correspondence. A fixed number of fifty components was used to balance between diverse subpopulation identification and having a sufficient number of data points in each subpopulation. Since neural network models such as autoencoders produce distributed or "entangled" representations that are subject to random variation, there is no guarantee of finding meaningful subpopulations with this method, and in practice a greater degree of disentanglement would be necessary to identify consistent and meaningful subpopulations. Note this is just one possible method of segmenting the latent space: it was used here merely to exemplify the general approach to representational ethical model calibration.

Subpopulation performance

The regression model tended to perform differently across subpopulations, deviating from the ideal of equal fidelity for all. To identify the characteristics of those that exhibited particularly poor model performance, a combination of visual analysis and permutation tests was employed. Permutation tests are non-parametric tests that explore all possible random orderings of a variable in a dataset and thus produce a p-value of a variable's mean within a subpopulation[56]. This provides an objective way to determine which variables are important in defining that subpopulation. It does not, however, give specific details and the overall results require domain knowledge to interpret. For computational reasons, it was unfeasible to calculate all possible permutations, and so the standard approximation of 1000 rounds was used. Due to the number of tests (testing each of 25 variables across 5 groups), the Benjamini-Hochberg procedure[57] was used to control the False Discovery Rate at $a = 0.05$.

To quantify the overall level of inequity in a population of model predictions, the Gini coefficient was calculated. This economic metric was originally used to index the dispersion of income differences in a population[58]. Its calculation is based on the relative mean absolute difference in incomes across the population. A Gini coefficient of zero indicates equality of income and a Gini coefficient of one indicates maximum inequality. If the distribution is random, the Gini coefficient has an expected value in the region of 0.33. In our case, "income" becomes model fidelity as captured by NRMSE. The Gini coefficient is a useful summary statistic, particularly when examining the impact of a remediation algorithm on a population of model results. Note other indices of equity may be used here, dependent on the measure of fidelity most appropriate to the specific task: our use of Gini is intended to be illustrative.

Remediation

Once model inequity was identified, the chosen remediation approach was to oversample the underperforming group, before retraining the model on the rebalanced dataset. This is referred to as rebalancing. There is no consensus on the best method for remediation[59]: rebalancing is used here merely to illustrate remediation in the context of the broader calibration framework. The underperforming group was defined to be data points in subpopulations with a performance metric below the median performance for all. This reduced the emphasis on particular subpopulations, which were automatically resampled in proportion to their size. This method was selected after previous experiments showed that rebalancing based on individual groups could push already-marginalised groups to have even worse results. Hence the dataset was split into an "under-served" (underperforming) group and a "base" (remainder) group. The oversampling multiplier was set to optimise model performance following experimentation (Fig. 7). The experiment was repeated 10 times and the variance in results is shown in Fig. 6 and Tables 3, 4.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The UK Biobank resource is available to researchers for health-related research in the public interest. All researchers who wish to access the research resource must register with UK Biobank at https://www.ukbiobank.ac.uk.

CODE AVAILABILITY

The code used in this study is available from the corresponding authors on request by email.

REFERENCES

1. Sackett, D. L. Evidence-based medicine. In *Seminars in perinatology*, vol. 21, 3–5 (Elsevier, 1997).
2. Greenhalgh, T., Howick, J. & Maskrey, N. Evidence based medicine: a movement in crisis? *Bmj* **348**, g3725 (2014).
3. Crisp, R. *Aristotle: Nicomachean Ethics* (Cambridge University Press, 2014).
4. Health equity. https://www.who.int/health-topics/health-equity. Accessed: 2022-08-13.
5. Xiao, C., Choi, E. & Sun, J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J. Am. Med. Inform. Assoc.* **25**, 1419–1428 (2018).
6. Vasiljevic, Z. et al. Smoking and sex differences in first manifestation of cardiovascular disease. *Atherosclerosis* **330**, 43–51 (2021).
7. Bica, I., Alaa, A. M., Lambert, C. & Van Der Schaar, M. From real-world patient data to individualized treatment effects using machine learning: Current and future methods to address underlying challenges. *Clin. Pharmacol. Ther.* **109**, 87–100 (2021).
8. Bzdok, D. & Meyer-Lindenberg, A. Machine learning for precision psychiatry: opportunities and challenges. *Biol. Psychiatry.: Cogn. Neurosci. Neuroimaging* **3**, 223–230 (2018).
9. Lau, D. C. & Murnighan, J. K. Demographic diversity and faultlines: The compositional dynamics of organizational groups. *Acad. Manag. Rev.* **23**, 325–340 (1998).
10. DiBenigno, J. & Kellogg, K. C. Beyond occupational differences: The importance of cross-cutting demographics and dyadic toolkits for collaboration in a us hospital. *Adm. Sci. Q.* **59**, 375–408 (2014).
11. Li, J. & Hambrick, D. C. Factional groups: A new vantage on demographic faultlines, conflict, and disintegration in work teams. *Acad. Manag. J.* **48**, 794–813 (2005).
12. Thatcher, S. & Patel, P. C. Demographic faultlines: A meta-analysis of the literature. *J. Appl. Psychol.* **96**, 1119 (2011).
13. Bambra, C. Placing intersectional inequalities in health. *Heal. Place* **75**, 102761 (2022).
14. Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).
15. Niculescu-Mizil, A. & Caruana, R. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, 625–632 (2005).
16. Weng, W.-H. & Szolovits, P. Representation learning for electronic health records. *arXiv preprint arXiv:1909.09248* (2019).
17. Girkar, U. M. et al. Predicting blood pressure response to fluid bolus therapy using attention-based neural networks for clinical interpretability. *arXiv preprint arXiv:1812.00699* (2018).
18. Landi, I. et al. Deep representation learning of electronic health records to unlock patient stratification at scale. *NPJ digital Med.* **3**, 1–11 (2020).
19. Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.* **6**, 1–10 (2016).
20. WHO. Use of glycated haemoglobin (hba1c) in diagnosis of diabetes mellitus: abbreviated report of a who consultation. Tech. Rep., World Health Organization (2011).
21. Sudlow, C. et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
22. Challen, R. et al. Artificial intelligence, bias and clinical safety. *BMJ Qual. Saf.* **28**, 231–237 (2019).
23. Vayena, E., Blasimme, A. & Cohen, I. G. Machine learning in medicine: addressing ethical challenges. *PLoS Med.* **15**, e1002689 (2018).
24. Yu, A. C. & Eng, J. One algorithm may not fit all: how selection bias affects machine learning performance. *Radiographics* **40**, 1932–1937 (2020).
25. Singh, R. et al. Deep learning in chest radiography: detection of findings and presence of change. *PloS one* **13**, e0204155 (2018).
26. De Fauw, J. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
27. Ferryman, K. & Pitcan, M. Fairness in precision medicine. *Data & Soc.* 1 (2018).

28. Blank, R., Burau, V. & Kuhlmann, E. *Comparative health policy* (Macmillan International Higher Education, 2017).

29. Starke, G., De Clercq, E. & Elger, B. S. Towards a pragmatist dealing with algorithmic bias in medical machine learning. *Med. Health Care Philos.* **24**, 341–349 (2021).

30. Char, D. S., Shah, N. H. & Magnus, D. Implementing machine learning in health care—addressing ethical challenges. *N. Engl. J. Med.* **378**, 981 (2018).

31. Goldacre, B. *Bad pharma: how drug companies mislead doctors and harm patients* (Macmillan, 2014).

32. McCradden, M. D., Joshi, S., Mazwi, M. & Anderson, J. A. Ethical limitations of algorithmic fairness solutions in health care machine learning. *Lancet Digit. Heal* **2**, e221–e223 (2020).

33. Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A. & Nagar, S. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, **63**, 4–1 (2019).

34. Bird, S. et al. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).

35. Saleiro, P. et al. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577* (2018).

36. Sagawa, S., Koh, P. W., Hashimoto, T. B. & Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations* (2019).

37. Tatman, R. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 53–59 (2017).

38. Schupak, A. Google apologizes for mis-tagging photos of African americans https://www.cbsnews.com/news/google-photos-labeled-pics-of-african-americans-as-gorillas/ (2015).

39. Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-toolthat-showed-bias-against-women-idUSKCN1MK08G (2018).

40. Angwin, J., Larson, J., Mattu, S. & Kirchner, L. Machine bias https://www.propublica.org/article/machine-biasrisk-assessments-in-criminal-sentencing (2016).

41. Hellström, T., Dignum, V. & Bensch, S. Bias in machine learning–what is it good for? *arXiv preprint arXiv:2004.00686* (2020).

42. Porta, M. *A dictionary of epidemiology* (Oxford university press, 2014).

43. Olteanu, A., Castillo, C., Diaz, F. & Kıcıman, E. Social data: Biases, methodological pitfalls, and ethical boundaries. *Front. Big Data* **2**, 13 (2019).

44. Torralba, A. & Efros, A. A. Unbiased look at dataset bias. In *CVPR 2011*, 1521–1528 (IEEE, 2011).

45. Sap, M., Card, D., Gabriel, S., Choi, Y. & Smith, N. A. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 1668–1678 (2019).

46. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An introduction to statistical learning*, vol. 112 (Springer, 2013).

47. Pombo, G. et al. Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3d deep generative models. *arXiv preprint arXiv:2111.14923* (2021).

48. Chaloner, K. & Verdinelli, I. Bayesian experimental design: A review. *Stat. Sci.* **10**, 273–304 (1995).

49. Cohn, D. A., Ghahramani, Z. & Jordan, M. I. Active learning with statistical models. *J. Artif. Intell. Res.* **4**, 129–145 (1996).

50. Berk, R., Heidari, H., Jabbari, S., Kearns, M. & Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociol. Methods Res* **50**, 3–44 (2021).

51. Vos, T. et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *Lancet* **396**, 1204–1222 (2020).

52. Dolezalova, N. et al. Development of a dynamic type 2 diabetes risk prediction tool: a uk biobank study. *arXiv preprint arXiv:2104.10108* (2021).

53. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

54. Friedman, J. et al. *The elements of statistical learning*. 10 (Springer series in statistics New York, 2001).

55. Ogbuabor, G. & Ugwoke, F. Clustering algorithm for a healthcare dataset using silhouette score value. *Int. J. Comput. Sci. Inf. Technol. (IJCSIT)* **10**, 27–37 (2018).

56. Efron, B. & Tibshirani, R. J. *An introduction to the bootstrap* (CRC press, 1994).

57. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.: Ser. B (Methodol.)* **57**, 289–300 (1995).

58. Chen, C.-N., Tsaur, T.-W. & Rhai, T.-S. The gini coefficient and negative income. *Oxf. Econ. Pap.* **34**, 473–478 (1982).

59. Caton, S. & Haas, C. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053* (2020).

ACKNOWLEDGEMENTS

AUTHOR CONTRIBUTIONS

R.C. and P.N. jointly developed the approach and wrote the manuscript, based on ideas formulated in discussion with I.S., J.K.R., D.H., A.N., D.B., D.F-R., and G.R. R.C. performed the experiments. All authors contributed to revising the work critically for important intellectual content, approved the completed version, and are accountable for all aspects of the work.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-022-00716-4.

Correspondence and requests for materials should be addressed to Robert Carruthers or Parashkev Nachev.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.