Can Improved Data Representation Support AI for Health Equity? A Visual Approach

Steven Vethman[1,2], **Jildau Bouwman**[2,3], **Cor Veenman**[2,4]
[1] Sciences Po Law School
[2] Netherlands Organisation for Applied Scientific Research (TNO)
[3] Leiden Academic Centre for Drug Research (LACDR)
[4] Leiden Institute of Advanced Computer Science (LIACS)
steven.vethman@sciencespo.fr, c.j.veenman@liacs.leidenuniv.nl, j.bouwman@lacdr.leidenuniv.nl

Abstract

Representation is a critical concern in the development, use and decisions about AI. AI fairness and health equity research present frameworks for representation that highlight underlying structural issues (such as systemic discrimination) that go beyond, yet also shape, both data and AI tooling. Still, there is a lack of actionable methods to interrogate data representation and support critical reflection on its role in AI systems.

We propose a visual, data-driven approach that links decisions about data representation to AI model performance across subgroups. Our method consists of two plots: the *representation association plot*, which shows whether subgroup representation affects performance, and the *representation expansion plot*, which simulates how performance disparities may change when expanding subgroup data. We apply our approach to the Lifelines Cohort Study for two health equity use cases: early detection of diabetes and cardiovascular disease. The plots reveal that improving age representation may reduce disparities, while sex and education-based disparities appear unrelated to representation in the dataset. This approach has the potential to guide researchers in identifying when improving data representation may contribute to reducing performance disparities, and when such efforts are unlikely to be effective. As a critical but partial tool, our approach should be embedded in broader inclusive research practices, where representation extends to who defines the data and determines whether and how AI is applied. Future research should validate the actionability of insights with users and priorities of those bearing the health burden.

1 Introduction

Representation is a critical concern in the development and use of AI systems. Without it, patterns of systemic discrimination have been reinforced. For example, the Amazon case illustrates how reliance on a hiring algorithm trained on biased data reflected and reinforced exclusionary practices, especially in male-dominated industries (Dastin 2018). In healthcare, the use of an AI system focusing on treatment costs rather than healthcare needs reinforced the systematic under-treatment of Black patients (Obermeyer et al. 2019). This has been linked to a history of under-documentation and symptom dismissal, which intersects with the under-representation of Black healthcare providers in leadership roles (Obermeyer et al. 2019; Hoffman et al. 2016; van Ryn et al. 2011). These cases demonstrate that representation in AI is not simply about proportional inclusion in datasets, but is deeply shaped by structural inequities that have determined who is included in the data, who creates it, and whose priorities shape how AI systems are optimized and evaluated (Chasalow and Levy 2021; Bergman et al. 2023).

Representation is likewise a central issue in health research. Many global health challenges, from diabetes to cardiovascular disease, disproportionately affect groups that remain under-represented in research (Centers for Disease Control and Prevention 2023; Tobb, Kocher, and Bullock-Palmer 2022; Sardar et al. 2014; Chow et al. 2012). Unrepresentative clinical trials undermine validity and contribute to *health inequities*, which are defined as the avoidable and unjust systemic differences in health outcomes across population subgroups (Sharma and Palaniappan 2021; Turner et al. 2022; World Health Organization 2024, 2025). This disconnect raises concerns about the fairness and generalizability of research outcomes. Scholars emphasize the structural roots of exclusion and discrimination that drive these disparities: from unequal access to clinical trials, which affects who is included in health datasets (Bodicoat et al. 2021; Kelsey et al. 2022; Ibrahim et al. 2021), to the evaluation of how trial findings generalize across subgroups defined by characteristics such as sex, educational attainment, or age (Kent et al. 2020; Saria and Goldenberg 2015).

Both AI fairness and health equity research underscore the need to recognize the multifaceted and contextual nature of representation. While this takes representation beyond data imbalances, data remain significant, as both data and data-driven AI mirror the choices, values, and practices embedded in our society (Vallor 2024). Frameworks such as those by Bergman et al. (2023) and Mccradden et al. (2023), which adopt a broad, critical lens on representation, therefore highlight the importance of attending to data representation in AI. Clearly, such efforts must be situated within sociotechnical contexts and treated as iterative processes, rather than a one-time check or technical fix at the end of development (Selbst et al. 2019).

Yet, operationalizing such an approach remains challenging. For example, Park et al. (2021), who examined the often-overlooked issue of age-related under-representation in AI datasets, stress the need to link critical data inspec-

tions to actual differences in model performance. Receiving systematically lower performance (with mistakes leading to fewer benefits or more harms) is one pathway through which health disparities can be reinforced or exacerbated. When is data representation part of the solution, and when is it not? How can we determine whether issues of discrimination are embedded in the data or model, and whether improving data representation is an effective strategy to address them? This is particularly important in high-stakes domains like healthcare, where the increasing use of AI can reinforce or exacerbate existing inequalities in clinical decision-making (Yu, Beam, and Kohane 2018; Haug and Drazen 2023).

Our work addresses this gap by proposing a visual, actionable method to support responsible decision-making around representation in AI. The approach is grounded in data analysis, yet critically attuned to the broader question of whether data representation is a meaningful part of the solution. It is aimed to enable researchers and practitioners to identify when disparities in training data correspond to differences in model behavior, and when improvements in data representation are unlikely to make a meaningful difference for equitable performance. We draw on insights from both AI fairness and health equity research to develop and demonstrate this approach, responding to calls for context-aware analyses of representation (Bergman et al. 2023; Mccradden et al. 2023; Chasalow and Levy 2021). Concretely, our method centers on two diagnostic plots that support critical reflection on whether and how subgroup representation influences model performance. Each plot corresponds to a guiding research question:

RQ1 To what extent is subgroup representation associated with model performance disparities?

RQ2 How does expansion of training data influence subgroup-level performance disparities?

Understanding whether and how representation affects performance is essential for deciding when improving data is an effective lever for advancing equity, and when other interventions may be required. We illustrate our approach using the Lifelines Cohort Study across two use cases: early detection of Type 2 Diabetes (T2D) and Cardiovascular Disease (CVD). These use cases involve AI models trained on clinical cohort data to identify individuals at risk years before disease onset, enabling early intervention. We examine how subgroup representation in the training data may affect performance disparities related to gender, age, and education.

The layout of this paper is as follows. Section 2 elaborates on related work on representation in both AI fairness and health equity research. Section 3 introduces our approach with the plot design for each research question. Section 4 describes all elements for the empirical studies of early detection of T2D and CVD, i.e. the data set, the operationalization of representation analysis for gender, age and education as well as the models and metrics. Section 5 discusses the results of both plots leading to actionable suggestions for both T2D and CVD. Section 6 reflects on the main findings and presents its limitations. Section 7 concludes.

2 Related Work

In this section, we position our work in relation to two bodies of literature. The first concerns representation in health equity research, particularly in biomedical science, which is the domain of our two empirical use cases. The second focuses on representation in AI fairness and ethics. We explain how our approach relates to existing work in both fields and clarify the specific contribution of our visual method.

2.1 Health Equity

Representation has been widely acknowledged as a critical concern in health research. In their terms, equitable research requires the inclusion of populations that bear the health burden, yet many subgroups (such as those based on socioeconomic status or ethnicity) remain underrepresented in biomedical studies (Chaiyachati et al. 2022; National Academies of Sciences, Engineering, and Medicine 2022; Oh et al. 2015). This under-representation leads to unreliable treatment effects, unknown side effects, and ultimately lower care quality for these groups (Sharma and Palaniappan 2021). Historically, clinical research has focused on relatively homogeneous populations, often white and male, limiting generalizability and forcing clinicians to extrapolate findings where evidence is lacking (Louie and Wilkes 2018; Perez 2019; Santema et al. 2019). The consequences include poorer health outcomes, increased mistrust in medical institutions, and widening disparities (Sharrocks, Camidge, and Papa 2014; Bodicoat et al. 2021). Beyond ethical concerns, addressing such inequities also yields economic benefits: estimates suggest that reducing health disparities could have saved the U.S. hundreds of billions of dollars in recent decades (LaVeist, Gaskin, and Richard 2011; Bhatt et al. 2023; Roldós and Breen 2021; Nanney et al. 2019).

In response, research on health equity has largely taken two directions. One approach seeks to improve representation in health data itself, under the logic that you cannot address what you do not measure (Bodicoat et al. 2021; Kelsey et al. 2022; Ibrahim et al. 2021). However, these studies do not make concrete what level of representation is necessary to reduce disparities? Another common approach focuses on representation in health outcomes. These studies aim to evaluate biomedical findings across subgroups (also known as subtyping), highlighting how treatment effects may differ (Kent et al. 2020; Saria and Goldenberg 2015; Kent et al. 2010). Yet these approaches often do not address whether differences are caused by under-representation or other factors, such as real variation in disease prevalence across subgroups, for instance, due to the role of socioeconomic complexity in disease onset (Rothman and Greenland 2005; Foster et al. 2018; Wensink, Westendorp, and Baudisch 2014; Vinke et al. 2020). Or recall the study of Obermeyer et al. (2019) on structural under-treatment of Black patients; differences in outcomes may also reflect historical biases embedded in the healthcare system. In short, most studies focusing on subgroup performance leave open a critical question: are disparities driven by under-representation in data, or by other social and structural factors?

There are few papers that (similar to us) recognize the necessity to bridge the focus on representation in health data and disparities in health outcomes. For instance, Seyyed-Kalantari et al. (2021) show how under-diagnosis in chest X-rays relates to under-representation, though they do not quantify that under-representation to guide design. Similarly, Larrazabal et al. (2020) examine how imbalances in medical image data affects diagnosis performance, but do not distinguish performance effects from underlying differences in disease prevalence (an important factor in trial design (Gross et al. 2022)). Others stress the value of group-specific models for diseases such as type 2 diabetes (Farran et al. 2013), yet fall short of connecting this insight to concrete design recommendations.

Clearly, while these studies take important steps by linking representation in health data to disparities in health outcomes, they fall short in two key respects. First, they miss actionability: they do not provide guidance on how to determine which additional subgroup data could meaningfully reduce under-representation and improve health equity. Second, they rarely reflect on the limits of representation efforts that focus narrowly on data inclusion. They do not explicitly activate researchers to question whether disparities are in fact reducible through data alone. As an answer, this paper offers a data-driven approach to help decision-making while attuned to the limitations of data representation, namely the question posed in our paper's title.

2.2 AI Fairness

Within the domain of AI fairness and ethics, the term representation is frequently put forward as a positive and necessary goal alongside terms like bias and fairness, yet it often remains ill-defined (Chasalow and Levy 2021). The common interpretation follows a statistical lens: whether data and by extension the outcomes are generalizable, alluding to statistical concepts of a representative sample and external validity (Bergman et al. 2023).

However, interpreting representation purely as a statistical issue is increasingly contested. Representation, similar to bias and fairness, only becomes meaningful when situated within a sociotechnical context (Selbst et al. 2019). This broader view challenges the assumption that under-representation should always be remedied by collecting more data or adapting the algorithm (Davis, Williams, and Yang 2021; Buolamwini and Gebru 2018; Miceli, Posada, and Yang 2022).

To approach representation as a contextual, multifaceted and power-laden concept, several important questions need be raised: Representation in what? In the AI development team, in the training data, in the evaluation benchmark, or in the decision-making process? Representation for whom? Do we look at the protected characteristics defined in discrimination law, such as sex, ethnicity, and age, or do we begin by examining the context and consult domain experts? And representation for what purpose? Are we aiming to achieve equal representation across all groups to comply with standards or regulation, or are we seeking to challenge the status quo by intentionally amplifying marginalized perspectives?

In what follows, we discuss four influential papers from the FAccT and AIES conferences that recognize the multifaceted nature of representation in AI and examine the role of data therein. We position our approach in relation to their work, while underscoring the originality of our contribution.

Chasalow and Levy (2021) trace representation to its historical roots in both statistics (representative sample) and politics (representative government). Thereafter, they elaborate on key debates in AI ethics: on data (e.g. selection bias, unfair representation), shifts (divergence between data distribution of training and deployment), participation (inclusion beyond data subjects, e.g., in model design) and power (focusing on data representation may distract from questioning the desirability of the technology). Above all, they lay the conceptual foundation in AI ethics on the multifaceted nature of representation, but do not yet operationalize this.

Bergman et al. (2023) build on the conceptual foundation of the former and zoom in on representation for AI evaluation. They propose the need for *subject-domain representation*, emphasizing that representation should be instrumentally beneficial. They outline two conditions that come from understanding the application domain: the identification of relevant subgroups and a specification of meaningful performance within that context. They also specify tensions and limitations in realizing representation through data and evaluation (e.g. access, surveillance and power), and frame where and how in the AI pipeline subject domain-representation could be included. However, they call for future research to make the evaluation of instrumentally beneficial representation operable.

Next to that, Mccradden et al. (2023) propose an ethical framework to stimulate the operationalization of medical ethics and social justice in clinical machine learning. They advocate for attention to representation in research teams, consultation processes, data, evaluation, and across several stages of the machine learning pipeline; from problem formulation to critical reflection on representation in clinical trials. They position their comprehensive, multi-step guideline as a starting point for value-based decision-making around issues such as bias in healthcare AI.

Finally, Park et al. (2021) investigated the often-overlooked issue of age representation in 92 face (AI) data sets. Their analysis indicated an under-representation of older adults and they elaborate on the specific challenges and contextual needs for the specific subgroup in AI applications. However, they point to future research to make a link between under-representation in data and performance disparities.

Our work builds on the fundamental contributions of Chasalow and Levy (2021) and Bergman et al. (2023) as our visual and critically grounded approach assesses how data representation affects AI outputs, while recognizing that representation is more than data inclusion. We follow their emphasis on context and domain expertise by designing a visual approach through a collaboration between AI and health experts, and demonstrate it through two use cases. The fact that our plots are designed to link data representation to performance disparities makes, them a concrete instantiation of the instrumentally beneficial representation Bergman et al. (2023) advocate for and is a direct call from what Park et al.

(2021) leave to future research.

Our contribution complements the conceptual and step-by-step frameworks of Bergman et al. (2023) and Mccradden et al. (2023) by offering a concrete method for critically examining data representation at various stages. Hence, the originality of our work lies in the concrete operationalization and demonstration of a critical approach to data representation; one that can complement and be integrated into the broader conceptual frameworks proposed by related work.

3 Visual Approach

In this section, we propose two visual aids that support the critical assessment of data representation within the context of using AI for health equity. The first visual aid, called the *representation association plot* centres on answering the first research question (RQ1) on the relation between (under-)representation and subgroup-specific performance. The second visual aid, called the *representation expansion plot*, zooms in on how expanding different subgroup data influences any performance disparities, i.e. the second research question (RQ2).

3.1 Terminology and Setting

To contextualize the proposed visual aids, we briefly define key terms and introduce a practical example from biomedical AI. Imagine a cohort study used to train a machine learning model to detect a disease from clinical health data. The *target variable* indicates whether a participant has or develops the disease. We define *subgroup features* as attributes relevant to known health disparities, such as sex, ethnicity, or socioeconomic status. For instance, studying whether under-representation of female patients affects predictive accuracy requires a subgroup feature that encodes sex. The *base rate* of a subgroup refers to its disease prevalence (the proportion of people that have the disease) or incidence (the proportion of people that contract the disease within a period). We assume a standard split between *training* and *test sets*: the model learns associations from the training set and is evaluated on the test set. For simplicity, we consider binary target variables and binary subgroup features. (More on this choice and its implications in Section 4.)

3.2 Representation Association Plot

We designed the representation association plot to examine whether under-representation in the data is the driving factor, and therefore also a possible solution, for adverse group-specific differences in the model's outcomes. The representation association plot is a curved line-chart where each curve represents the performance on a test set and is accompanied by an error band (two standard errors) to illustrate uncertainty. For a visual example along this explanation, see Figure 1. The y-axis indicates the performance metric. The defining feature of the plot is that this performance is shown for different subgroup compositions in the training set, which are indicated on the x-axis. In other words, the plot is designed to show how different subgroup compositions in the training data affect the model's performance on each subgroup in the test data. To isolate representation as a possible driver of health disparities, the plot varies subgroup proportions in the training set while keeping the base rate and sample size constant.

The four-step procedure below illustrates how the plot is constructed. Where helpful, we refer back to the practical setting involving an AI diagnostic tool that considers two subgroups: female and male participants.

1. *Equalize data*: The dataset is sampled to contain equal amounts of data for each subgroup, while maintaining constant proportions of the target variable. The majority subgroup is under-sampled to match the size of the minority subgroup for both outcomes. For example, the subset of female participants (majority) is under-sampled to match the number of male participants, with each subset having the same number of individuals who did or did not develop the disease.
2. *K-Fold cross-validation*: The equalized dataset is split into k folds, keeping subgroup and target variable proportions constant. In each iteration, $k-1$ folds are used for training and one fold for testing. For example, each fold contains the same number of male and female participants with and without the disease.
3. *Different representations for training*: In each cross-validation iteration, (a predefined set of) varying subgroup compositions are drawn from the training folds. For example, training sets are constructed with 0%, 5%, 10%, ... up to 100% female participants. While the subgroup composition varies, the number of participants with and without the disease remains constant.
4. *Different subgroups for testing*: Each representation setting of the training set is tested on a separate test set for each subgroup as well as a benchmark test set with the proportions of the subgroups as observed in the original data set. For example, each subgroup composition of the train set is evaluated on a test set of only male participants, a test set of only female participants, and a test set with 41% male and 59% female participants.

We emphasize the specific design choice to vary subgroup representation while keeping both sample size and base rate constant. This isolates the influence of representation from differences in disease prevalence or data availability. To compensate for the under-sampling in step 1, we repeat the analysis 20 times, allowing us to use the full dataset and capture uncertainty in the estimates.

To illustrate the actionable insights this plot can provide, we describe three prototypical visual patterns (see Figure 1). When curves overlap, performance is similar between subgroups (plot A). Vertical separation between curves indicate subgroup performance differences (plot B). A sloped curve indicates that representation impacts performance for the particular subgroup (plot C). The benchmark curve (based on original subgroup proportions) shows how much additional information is gained, or not, by examining the performance for each subgroups relative to just analyzing a randomly sampled train and test set. In more detail, the three prototypical visual patterns are interpreted as follows:

1. *Stable Similar Outcomes*: No difference in performance per subgroup and no association with representation. The

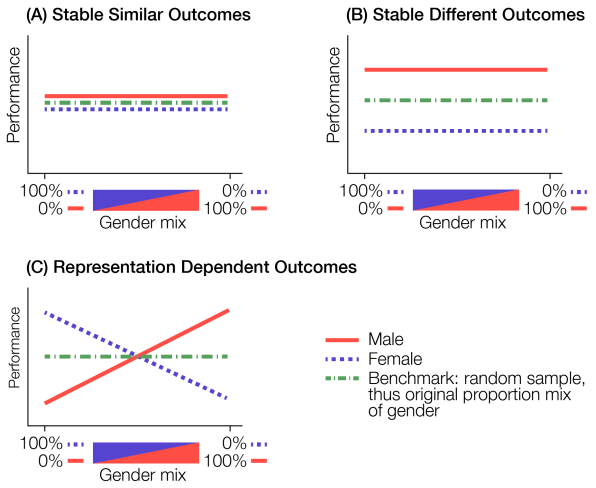

Figure 1: Illustrating the actionable guidance of the representation association plot: three possible visual patterns are demonstrated on two example subgroups of the male and female gender. To clearly show the three scenarios, the uncertainty bandwidths are intentionally omitted.

overlapping horizontal curves provide no indication that the data representation is an issue, for the given sample size.

2. *Stable Different Outcomes*: Differences in performance per subgroup are observed whilst there is no indication that differences stem from representation. The three curves are horizontal with vertical differences. Depending on the magnitude of the vertical difference, this may signal that the model is more capable for one group than the other. However, no evidence suggests that under-representation in the data is the driver of the difference.
3. *Representation Dependent Outcomes*: The sloped curves indicate that performance disparities are dependent on the representation in the training set. Where the curves cross, the performance per subgroup is similar, but the performance for each subgroup deteriorates as representation decreases. The slope of a curve indicates how much representation matters for each subgroup.

3.3 Representation Expansion Plot

The representation association plot indicates whether representation matters for a certain sample size. The next step for a (biomedical) researcher examining data representation is to determine what can be learned from the available data and what additional data is needed to advance health equity. The representation expansion plot therefore relieves the constraint of a fixed sample size and highlights the link between under-representation and health disparities for varying sample sizes. We also relax the constraint of fixed base rates, as acquiring additional subgroup data will follow its natural base rate rather than a fixed experimental one. In the context of AI development for disease detection, this means that differing disease prevalence between subgroups is consciously incorporated into the investigation of which additional data is needed to improve health equity.

To incorporate the additional aspect of varying sample size, the representation expansion plot is a 3D-plot (see Figure 2). A 3D plot allows not only different compositions in terms of subgroups but also different size of training data. In particular, the x- and y-axis ("going left and right") represent the amount of data points in the train set belonging to one subgroup or another. The performance of the models are projected on the vertical z-axis ("going up") that form a plane based on a grid of possible train set compositions. We elaborate below the four steps of the experiment, where each step is made tangible with an example based on two subgroups: female and male participants.

1. Split the data: the data set is split with respect to the subgroups. For example, a subset of male and a subset of female participants.
2. K-Fold cross-validation: both data sets are divided into k folds using stratified sampling for the target variable. In each iteration of the cross-validation process, $k-1$ folds are used for training and one fold is held out for testing. For example, both subsets of male and female participants are randomly split into k folds whilst ensuring that the ratio of participants that develop a disease and those that remain healthy still reflect the same balance as in the full subgroup data.
3. Create a grid of training set compositions: we sample proportional subsets of the available training data for each subgroup, for example, 5%, 10%, 20%, 50%, and 100%. That is, each entry in the grid would correspond to a specific combination, such as 20% female and 10% male, 20% female and 20% male, or 20% female and 100% male. Note that different combinations will result in varying overall sample sizes.
4. Evaluate the combined training sets: for every combination in the training grid, a model is trained and evaluated on separate test sets for each subgroup. For example, each training set is tested on both a male-only and a female-only test set.

As the 3D plot has two horizontal axes (x and y) representing sample sizes, any slice along x and z with a fixed y (or y and z with a fixed x) can be seen as a 2D learning curve (Viering and Loog 2022). The slope of a learning curve at any point indicates how much adding data boosts performance, given the data already available. Similarly, the representation expansion plot shows how much adding data of one subgroup improves performance for that subgroup, depending on existing data. It also reveals whether performance of the subgroup is helped, is indifferent or hurt when data from another subgroup is added. To illustrate the potential actionable insights the representation expansion plot can offer, we present the three corresponding visual patterns in detail in Figure 2.

1. *Adding data of others helps*: If it helps, then although representation might matter, more data regardless of subgroup is more important than subgroup specific data collection. This is observed by a strictly increasing plane

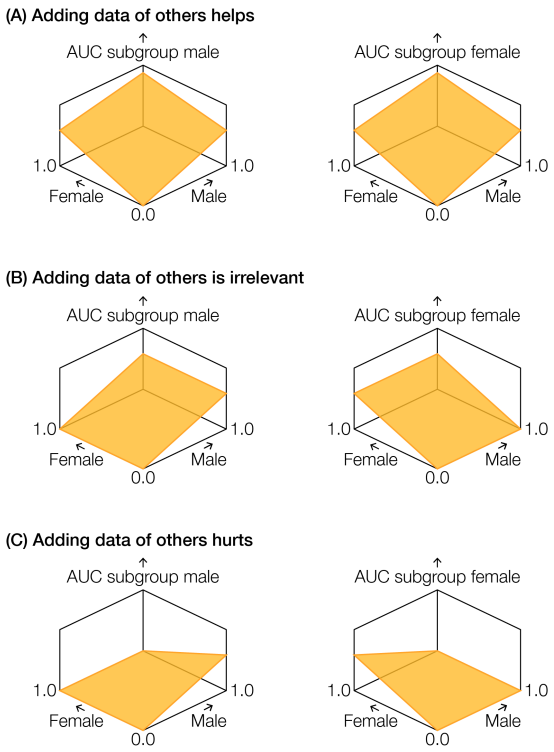

Figure 2: An illustration of the actionable insights provided by the representation expansion plot: three prototypical visual patterns are presented, with performance measured by AUC (ROC) for female and male participant subgroups.

that is highest when both training data sets are combined. In other words, the motto "More is Better" is suggested to fit the data collection procedure.

2. *Adding data of others is irrelevant*: If the performance of a subgroup neither improves nor declines when data from another subgroup is added, this offers valuable guidance for resource allocation. For example, if performance for a subgroup is already near optimal, efforts to collect more data for that group can reduced in favor of subgroups where additional data leads to gains. Visually, this appears as a plane that rises mainly in one direction (e.g. along the x-axis), while increasing data along the other (e.g. y-axis) has little effect, which suggests targeted, selective data collection.
3. *Adding data of others hurts*: If it hurts, pooling data worsens performance for at least one of the subgroups. This is observed by a plane that is not strictly increasing and where combining the full data sets shows sub-optimal performance. In other words, the plot suggests data collection for separate subgroup models, as the subgroup synergy is so low that information from one subgroup makes outcomes for another subgroup less effective.

As observed, each 3D plot shows only one subgroup's performance for all different training data compositions in size and representation. Interpreting two planes in one 3D plot is too complex.

4 Data, Models and Metrics

In this section we describe the empirical study that demonstrates the potential of the proposed plots. We first elaborate on the Lifelines Cohort Study (hereafter Lifelines) and the subgroup-defining features which we apply our approach to (Stolk et al. 2008). Then, we describe the models used for early onset detection, the metrics for their evaluation and the availability of data and materials.

4.1 Data Set

Lifelines is a multi-disciplinary prospective population-based cohort study examining in a unique three-generation design the health and health-related behaviors of 167,729 persons living in the North of the Netherlands. It employs a broad range of investigative procedures in assessing the biomedical, socio-demographic, behavioral, physical and psychological factors which contribute to the health and disease of the general population, with a special focus on multimorbidity and complex genetics. Lifelines has three longitudinal measurements between 2006 and 2023 (Scholtens et al. 2015). The time-span, large sample size and wide variety of potential indicators make Lifelines a suitable fit for early detection of diseases.

Similar to previous studies with Lifelines we include 139 potentially relevant variables for T2D and 142 for CVD (van der Meer, Wolffenbuttel, and Patel 2021; van der Ende et al. 2017). We primarily use the data from the baseline measurement executed in 2006. Only the target variable that measures whether someone develops T2D or CVD uses data from the first follow-up measurement (performed between 2011 - 2013). The target variable is a binary variable that is construed as 1 if a person without T2D or CVD at baseline has *developed T2D or CVD between baseline and follow-up* and 0 if a person without T2D or CVD kept this status. Thus, we assess the capability of early detection by observing whether the model can predict development of T2D or CVD solely based on data from 5 - 8 years prior. For CVD, most people (86%) provided an answer (yes or no) to the survey question on whether they developed CVD between baseline and follow-up. For T2D, only 8% answered the equivalent survey question. Therefore, we have used WHO diagnostic criteria for T2D and clinical measurements of fasting glucose levels (<7 or >7 mmol/l) and HBA1C levels (<48 or >48 mmol/l) to ascertain whether the participants were diabetic at baseline and whether they developed diabetes during the study (World Health Organization and Federation 2006; World Health Organization 2011). Both data sets are highly imbalanced with base rates showing that only 1.02% and 2.36% of the participants have developed T2D or CVD, respectively. For more details on the base rate per subgroup and descriptive tables of the data, see online Appendix A.

For missing data, we use the missing-indicator method which assumes MCAR or MAR (Little and Rubin 2019; Pereira Barata et al. 2019). Missing data in the categorical variables is considered informative and captured as a separate category *Unknown*. Missing data in numerical variables

is encoded with a binary variable that has a 1 if the value is missing and 0 if the value is observed. More information on missing data can be observed in online Appendix A.

Although a variable for ethnicity is available in Lifelines, there is not enough diversity for our approach to reliably estimate an association between data representation and performance. Lifelines only has 1.3% non-Western Migrants in the data set. This demographic is even under-represented relative to the North of the Netherlands (3.2%) (Klijs et al. 2015). On the national level, non-Western migrants constitute nearly 15% of the Dutch population, while the larger cities have over 25% (Centraal Bureau voor de Statistiek 2024).

4.2 Features: Gender, Age and Education

We demonstrate the plots using three types of subgroups: gender, age and education. We stress the need for diligent consideration in selecting subgroups grounded in societal relevance, domain expertise and the specific use case; this approach is also championed by related work (Mccradden et al. 2023; Bergman et al. 2023).

For gender, participants were asked in the baseline and follow-up surveys to self-report their gender, with the available options being: Male, Female, and Unanswered. We limit our analysis to the two self-reported categories: male and female. We acknowledge that focusing only on binary, cisgender categories does not provide a full account of gender-related health disparities. However, treating "Unanswered" as a distinct third category within a health questionnaire, as if it represents a gender identity beyond the binary, is also criticized (Scott et al. 2025); doing justice to gender diversity requires specific recognition and inclusive options that reflect the gender continuum (Eliason 2014; Frohard-Dourlent et al. 2017; Carrotte et al. 2016). Lifelines has recognized this as well and has taken steps toward more inclusive gender representation. For example, in their third assessment (2019–2023), they expanded the gender options in their questionnaire to better reflect gender diversity (University of Groningen; Lifelines 2025). Despite these limitations, analysis using the male and female subgroups remains relevant for illustrating the representation plots, especially given the well-documented male-female biases in biomedical research (Holdcroft 2007; De Castro, Heidari, and Babor 2016; Plevkova et al. 2020).

Education is measured as the highest degree obtained. We choose to inspect differences between participants having obtained tertiary education or not, as people with tertiary education were over-represented in the Lifelines data set (Klijs et al. 2015). A comparison of these two subgroups is also socially relevant through linking education (although imperfectly) as a proxy for socioeconomic status (SES). The health burden of non-communicable diseases such as T2D disproportionately falls on individuals with low SES, while medical researchers and their funders (those in positions of power) tend to come from higher SES (Grintsova, Maier, and Mielck 2014).

Age is registered in the number of years, ranging from 19 to 90. Based on related work, there is no clear cut-off value as studies choose different thresholds such as 45, 50, 60, 65, 75 to denote middle-aged groups or elderly groups (Koo et al. 2016; Yan et al. 2023). However, due to the design of Lifelines the proportion of participants aged 25-50 years are over-represented (Klijs et al. 2015). Therefore, to find out the potential effect of this over-representation on health disparities, we create two age groups: *younger* participants with age below 50 and *older* participants with age above 50 years.

In this work, we therefore use binary subgroup settings for each variable, while acknowledging that gender, age, and education exist on a continuum. The representation association plot could include more lines indicating the performance of more than two subgroups, such as multiple age brackets, education levels and gender identities. However, for the key component of the plot where the training data composition gradually changes, the current implementation compares one subgroup of interest to the rest. Since for two subgroups, the other subgroup constitutes "the rest", only one plot is needed. In the case of more than two subgroups, a separate plot for each subgroup against the rest can be estimated. For age and education, we construed the variable as the one known or likely to be under-represented and plot it against the rest. For gender, we use the two feasible subgroups available in the cohort and have excluded those who did not answer. For many demographic subgroups, such as gender and ethnicity, this may still simplify the complexity of identity. One may identify as Black, Brown and Western simultaneously, or be gender-fluid. Although we acknowledge the drawbacks such as exclusion or misrepresentation that can come from such simplification, we argue it remains sufficiently relevant for a useful illustration, given that our choices are grounded in societal relevance and the specifics of the cohort in question. Naturally, this only holds as long as we remain aware of its limitations; therefore, we return to this point in the Discussion.

4.3 Models and Metrics

We demonstrate our approach with two models for data-driven early detection: Logistic Regression (LR) (Hosmer 2000) and Extreme Gradient Boosting (XGB) (Chen and Guestrin 2016). LR and XGB are commonly used in computational medicine and diagnosis of CVD and T2D (Dinh et al. 2019; Abdalrada et al. 2022; Chaki et al. 2022; Lai et al. 2019).

LR is a classification model with logistic loss. In this paper, we use regularized logistic regression that aims to fit the target variable and features with coefficients in a linear log-likelihood whilst reducing the chance for overfitting with a L2 (quadratic) penalty term on coefficients (Hosmer 2000).

XGB is an efficient implementation of a gradient boosting algorithm where the iterative addition of classification and regression trees (CART) is the core principle (Chen and Guestrin 2016). Regularization is applied to reduce overfitting via a regularization term and a shrinking learning rate.

To evaluate the models, the Receiver Operating Curve (ROC) is used. The area under this curve, AUC ROC or AUC in short, represents the $sensitivity$ (also known as recall) and $1 - specificity$ for all possible thresholds. This measure is commonly used for evaluating machine learning

models, as well as the detection of diseases (Alaa et al. 2019; Dinh et al. 2019).

Moreover, to ensure that the performance metric estimates are based on sufficient cases of people developing the disease in the test set, we choose a relative low number $k = 3$ for the k-fold cross-validation (He and Ma 2013).

4.4 Availability of Data and Materials

Lifelines data are not publicly available. Researchers can apply to use the Lifelines data used in this study. The specific dataset and code of this study are preserved at Lifelines Biobank server, section (OV19_0514). More information about how to request Lifelines data and the conditions of use can be found on: https://www.lifelines-biobank.com/researchers/working-with-us. The general code for the plots, additional results and descriptive statistics are all available in the online Appendix A at: https://github.com/vethman-s/aies-2025-data-representation.

5 Results

This section presents findings from using the plots to analyze early detection of T2D and CVD. We begin by highlighting key insights from the representation association plot and explain how its interpretation helps to address RQ1: does shifting subgroup representation in the training data affect performance? We then turn to findings for RQ2, drawn from the representation expansion plot, showing how it can inform potential next steps.

5.1 Representation Association Plot

For the discussion of results coming from the representation association plots, we showcase the results for both use cases CVD and T2D. Given the similarity of results, we showcase only the results of XGBoost, whilst results of Logistic Regression are found in the online Appendix A.

Cardio Vascular Disease The results for CVD are shown in panels A to C in Figure 3, which showcase two visual patterns. In panel A concerning Age the plots show crossing curved lines and therefore indicate a visual pattern close to that of *Representation Dependent Outcomes*. Where a training set of mainly older (younger) patients leads to substantially better early detection of cardiovascular diseases for older (younger) people relative to younger (older) people. The curvature of the lines displays that the marginal benefit of changing the composition of training data is dependent on the level representation.

To further interpret that finding, we highlight that the early detection on older people (the red line) has the upward slope between training data compositions with 0% to 50% of older people, whilst any representation above 50% shows a nearly horizontal line. The blue curve (younger people) is stable between training data compositions with 0% and 60% older people, and only descends thereafter. This indicates that the model does not learn additional subgroup-specific information for older people when at least 50% of the data constitutes of older people and for younger when at least 40% of the data constitutes of younger people.

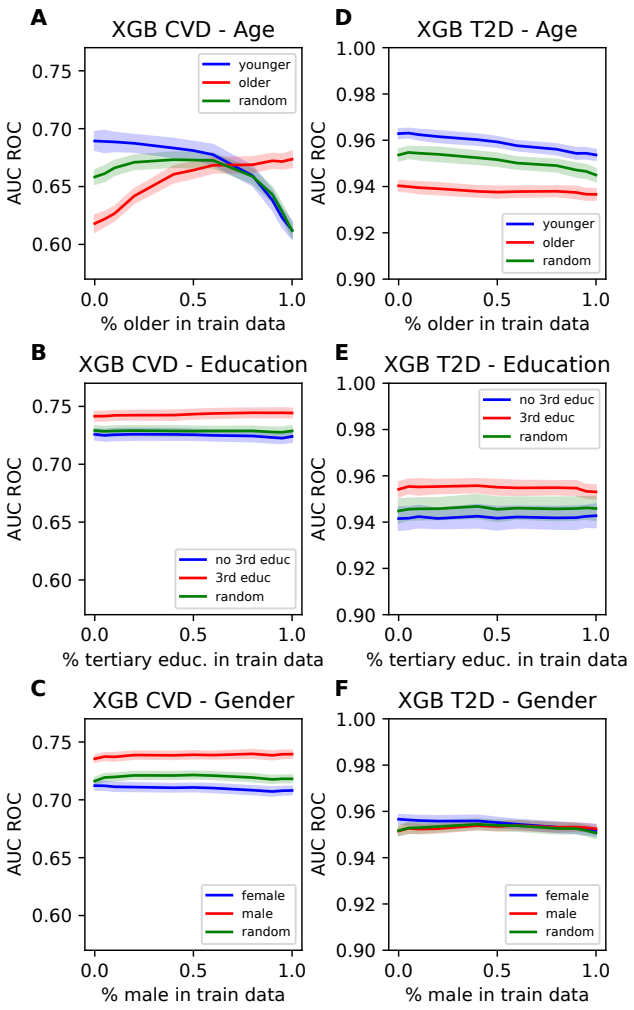

Figure 3: Representation association plots showing AUC-ROC performance for CVD (left) and T2D (right).

Furthermore, plot A illustrates how including a benchmark curve for a random sample highlights the importance of evaluating subgroup-specific outcomes. For example, if model performance had only been assessed on average rather than per subgroup, important disparities would be obscured. In a case where only 10% of participants are older adults, the model's ability to detect CVD early would be substantially overestimated for older individuals and underestimated for younger ones.

For panels B and C, concerning Education and Gender, the curves are horizontal and therefore show no indication that training data composition drives performance differences between subgroups. There is, however, a vertical gap between early detection of males and females (Panel C) and slightly smaller in magnitude between the detection for those with and without tertiary education (Panel B). This resembles most closely the visual pattern of *Stable Different Outcomes*. In other words, the plots visualize that there is no indication that under-representation of the under-performing

subgroup is the driver for these gaps.

Thus, in terms of what data representation may offer, the plots suggest that for age, an ideal representation constitutes of 50-60% people above 50 and 40-50% people below 50. For Gender and Education, there is no indication that improving data representation may offer solution for the differences in subgroup performance.

Type 2 Diabetes The results for T2D in panel D to F in Figure 3 show all three visual patterns. The representation association plot in Panel F displays the horizontal overlapping curves of *Stable Similar Outcomes*, i.e. no indication that performance differs per subgroup nor that representation in training data affects subgroup-specific outcomes.

The visual pattern of *Stable Different Outcomes* in panel E indicates that the models are better in detecting T2D early for people with tertiary education relative to those without, based on the measurement of AUC ROC. Similarly, plot D demonstrates that detecting T2D in younger people has a higher AUC ROC regardless of the representation in training data. Plot D and E also are suitable examples to stress that researchers need to interpret confidence intervals within the context of the use case. To elaborate, we observe in plot D that the confidence interval for younger around 0.96 does not overlap the confidence interval for older around 0.94. However, one needs to consider whether this difference in performance is substantial enough to be clinically relevant when using the model in practice (Wasserstein, Schirm, and Lazar 2019). Especially given the fact that the interpretation of absolute values of the AUC ROC require caution as it is slightly dependent on the class imbalance (only 1.03% developped T2D) (Davis and Goadrich 2006; Abdalrada et al. 2022). In other words, the patterns are a guideline for interpretation rather than that identifying the "right" visual pattern is a goal on its own.

To sum up, based on the representation association plots for T2D, no clear suggestions are raised that improving data representation is priority for limiting health disparities in terms of age, education or gender.

5.2 Representation Expansion Plot

For the representation expansion plot, we highlight its potential based on health disparities for age as they were indicated to be dependent on representation.

Cardio Vascular Disease Figure 4 shows four sets of representation expansion plots. Panel A and B visualize the AUC ROC of early detection for younger and older people who develop a CVD, respectively. Here we see an indication that group dependent information is key for achieving the highest possible performance. The AUC plane predominantly rises for the subgroup older (younger) when training data concerning older (younger) participants is added. We see the plane only slightly rising, the color only slightly darker, when adding training data from the other subgroup. The performance on the younger (older) people does not increase nor decrease when adding training data of older (younger) patients. In other words, the plots demonstrate the visual pattern of *Adding data of others is irrelevant*.

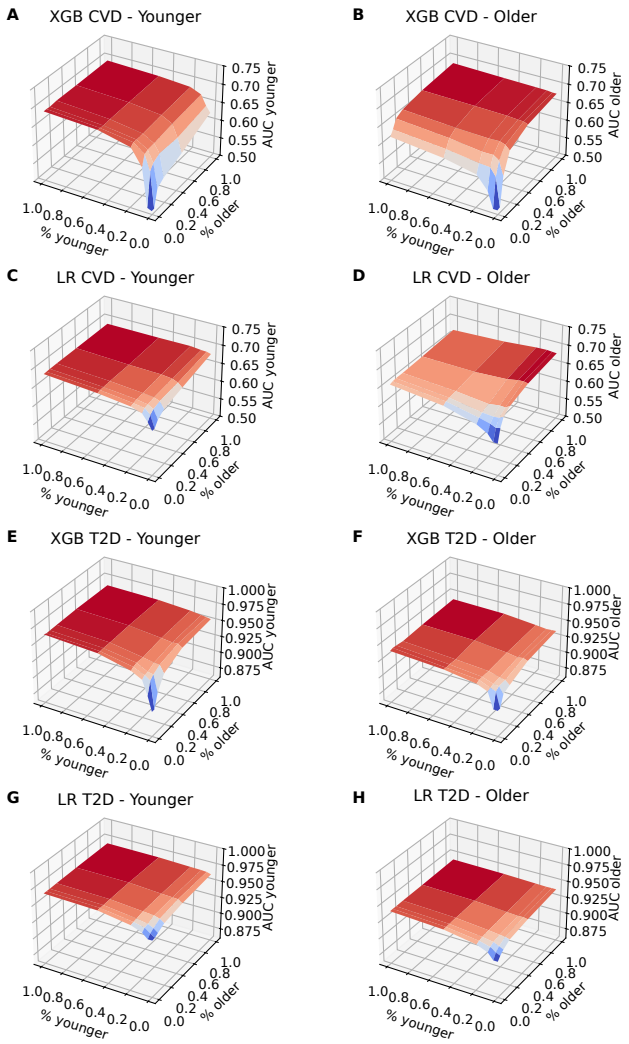

Figure 4: Representation Expansion Plots for age subgroups: CVD (top) and T2D (bottom). AUC shown by plane color, from blue (low) to dark red (high).

Panel C shows a dark-red peak when all data is added to the model which indicates the visual pattern of *Adding data of others helps*. Whether you move left or right along the axes, the plane rises. In other words, adding more data of younger or older people helps the model to increase performance for early detection for younger people.

In Panel D, we see the visual pattern where *Adding data of others hurts*. This is visible by the dark red peak in the right upper corner of the plot, where the training data constitutes the maximum amount of older people and minimum amount of younger people. In other words, the ability of Logistic Regression to aid early detection for older people deteriorates when adding training data of younger people.

In conclusion, the plots suggest that group-specific information is vital such that separate models for subgroups is a viable option. When Logistic Regression is chosen, a separate model for older people is even strongly advised.

Type 2 Diabetes For T2D, the plots E, F, G and H indicate that the value of additional data regardless of subgroup membership is greater than the value of subgroup-specific information. Similar to plot C for CVD, the visual pattern suggests that *Adding data of others helps*. This is not surprising, given the weak to no indication of representation dependent outcomes for T2D in Figure 3 panel D.

Thus, for T2D, the plots suggest that no immediate subgroup-specific data collection is needed from a health equity standpoint. Instead, expansion strategies may focus on increasing overall data volume (as this still improves model performance) or on acquiring data for relevant subgroups beyond age, education, and gender that were too underrepresented to be included in the analysis. In the case of early detection using Lifelines data, this includes ethnic minority groups in the Netherlands.

6 Discussion

This section discusses overarching insights, recommendations and limitations of our research.

First of all, we emphasize that the visual approach is aimed to support more critical decision-making about data representation, not a stand alone solution for AI fairness or health equity. The plots may indicate that an AI substantially under-performs and therefor may adversely affect a certain subgroup, and highlights whether this is related to data under-representation. Such a signal could be sufficient reason to take action, question the use of the AI or reconsider the data used to train it. However, the reverse is not necessarily true: the absence of an signal (the plots indicating there are no subgroup performance disparities) is not sufficient to conclude that no issues exist. From this, we could only infer that no health disparities are measured within the AI outcomes or with the specific metric chosen. Hence, for application of our visual approach, we stress the need for complementing mixed-method approaches where qualitative methods such as interviews and focus groups can capture the nuance that quantitative data cannot (Bates et al. 2020). Think also of the calls for participatory methods to co-define the goals of the AI, which data to include, and how the AI is evaluated (Bondi et al. 2021; Costanza-Chock 2020) or more particular in the domain of clinical trials, the participation of subgroups to identify blind spots in their design (Wallerstein et al. 2017; Dion et al. 2021). Such embedding is crucial to circumvent that our visual approach may distract from the important zero-question also highlighted by (Mccradden et al. 2023; Bergman et al. 2023; Chasalow and Levy 2021), does it serve under-represented groups at all that this AI is developed?

Furthermore, within appropriate embedding, our visual approach has the potential to support concrete action. For instance, the severe under-representation of people with an immigration background in our use cases made it impossible to estimate any performance disparities, let alone their association with under-representation. This could motivate collaboration with clinical cohorts in more diverse regions of the Netherlands, as conclusions based on Lifelines may be extrapolated to the whole of Netherlands. In addition, when sufficient data for the visual approach is available for a still under-represented subgroup, the plots can suggest a concrete proportion to strive for, reducing guesswork in next steps. Given the persistent challenges for achieving clinical diversity (Bodicoat et al. 2021), our approach offers representation targets tied to potential performance gains, which provides an alternative to the potential paralysis of striving for unattainable equality across all subgroups. However, we stress that claims about supporting concrete action should be strengthened by further research. This could include user studies with researchers, or involving representation advocates working on specific challenges such as diabetes and cardiovascular disease, to assess whether visual examination of data representation is a useful addition to their toolkit.

Next to that, while we have elaborated on our reasoning behind the binary subgroup comparisons, we acknowledge that this remains a simplification. The representation-association plots may not fully capture the complexity or intersectional nature of identity groups. Future work should explore whether and how this method continues to add value when applied to real-world cases with more granular or interwoven subgroup dynamics. It may be that data-driven methods are not suitable for certain small minority groups, as these approaches rely on estimating averages and certainties based on sample size. In such cases, inclusion in design and evaluation processes (with actual power in decision-making) may be more important than additional data collection (Costanza-Chock 2020; Ovalle et al. 2023).

Finally, we also encountered potential technical and operational limitations. As the plots evaluate multiple training data compositions, applying them to larger datasets or more complex models may require significant computing resources. Nonetheless, our implementation on Lifelines data (within a secure, offline environment with bounded computing power) demonstrated that our approach was still feasible.

7 Conclusion

Addressing under-representation in AI, especially in health contexts, requires looking beyond proportional inclusion and calls for reflection on what has impact. This study examines how data representation relates to AI model outcomes that affect people's lives. Is the model's under-performance on a subgroup related to their representation? Is expanding the data to include more examples from the subgroup an effective strategy? Our visual approach allows users to reflect on their decisions about data representation by answering these questions. Using two use cases, early detection of diabetes and cardiovascular disease, we showed how the plots can help identify when data representation may reduce performance disparities, and when adaptations in representation are unlikely to be effective. Notably, the plots revealed that for cardiovascular disease detection, improving age representation may reduce disparities and suggest group-specific data expansion, while sex- and education-based disparities appeared unrelated to data representation.

We also recognize that meaningful inclusion in terms of representation goes beyond data. Therefore, we position the visual aids as part of a broader toolkit that takes data representation as a central issue, yet does not blindly assume that solutions are found by altering data composition.

Acknowledgments

To start, we thank the funders of this study, as well as the team behind the Lifelines Cohort Study, its contributing research facilities, and all the study participants.

The study is partially funded by the Appl.AI program of TNO and the DIVERSIFAIR project from the Erasmus+ grant (101107969), and is therefore co-funded by the European Union. However, the views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Access to Lifelines is funded by the Health, Living and Work department of TNO. The Lifelines initiative, in turn, has been made possible by a subsidy from the Dutch Ministry of Health, Welfare and Sport, the Dutch Ministry of Economic Affairs, the University Medical Center Groningen (UMCG), the University of Groningen, and the northern provinces of the Netherlands (Drenthe, Friesland, Groningen). The data pertaining to the Lifelines Cohort Study used in this article have been obtained in agreement with the principles of the Declaration of Helsinki and the research code of the University Medical Center Groningen (UMCG). The Lifelines protocol was approved by the UMCG Medical Ethical Committee under number 2007/152. Informed consent was obtained from all individuals in the cohort study. The funding parties and programs for Lifelines had no role or influence in the cohort study design and results.

Furthermore, we acknowledge all colleagues at TNO who generously read, discussed, and provided intermediate feedback during the research. A special thank you to Erik Boertjes for transforming our initial sketches into the clear illustrations in Figures 1 and 2. Finally, we thank the anonymous AIES reviewers for their valuable input.

The authors declare that they have no competing interests.

Positionality Statement

As our contribution engages with representation at the brink of AI fairness and health equity research, we also reflect on our own position as writers. We are a research team where expertise in statistics, computer science, biomedical science, personalized health care, and experience in AI fairness research come together. We recognize a personal stake in this research, shaped by experiences of diabetes and cardiovascular disease in our families, and by encounters with societal bias in healthcare for deviating from the norm. Still, we acknowledge our many privileges and the blind spots they bring. We are all Dutch and have academic backgrounds. As a result, our research focuses on visual tools for an academic audience of AI and health researchers/practitioners, and draws on assumptions rooted in Dutch practices. It does not directly engage with representation beyond an academic focus and may be skewed toward Dutch (Western European) views and values. As such, our contribution lacks relevant international perspectives and the lived experiences of diverse affected communities. We are aware of these consequences. For example, the author's presence in employee resource groups and queer and activist communities helped clarify the limits of gender analysis based on data with binary labels. However, more expertise and agency from within the community would be needed to identify which actions are most urgent for meaningful inclusion of gender-diverse people.

References

Abdalrada, A. S.; Abawajy, J.; Al-Quraishi, T.; and Islam, S. M. S. 2022. Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: a retrospective cohort study. *Journal of Diabetes and Metabolic Disorders*, 21(1): 251–261.

Alaa, A. M.; Bolton, T.; Di Angelantonio, E.; Rudd, J. H.; and Van der Schaar, M. 2019. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PloS one*, 14(5): e0213653.

Bates, J.; Cameron, D.; Checco, A.; Clough, P.; Hopfgartner, F.; Mazumdar, S.; Sbaffi, L.; Stordy, P.; and De La Vega De León, A. 2020. Integrating FATE/critical data studies into data science curricula: where are we going and how do we get there? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 425–435. Barcelona Spain: ACM. ISBN 978-1-4503-6936-7.

Bergman, A. S.; Hendricks, L. A.; Rauh, M.; Wu, B.; Agnew, W.; Kunesch, M.; Duan, I.; Gabriel, I.; and Isaac, W. 2023. Representation in AI Evaluations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 519–533. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.

Bhatt, D. J.; Gerhardt, W.; Andy Davis, F.; Batra, N.; Dhar, A.; and Brian Rush, A. 2023. US Health Care Can't afford health inequities. https://www2.deloitte.com/us/en/insights/industry/health-care/economic-cost-of-health-disparities.html. Accessed: 2025-05-19.

Bodicoat, D. H.; Routen, A. C.; Willis, A.; Ekezie, W.; Gillies, C.; Lawson, C.; Yates, T.; Zaccardi, F.; Davies, M. J.; and Khunti, K. 2021. Promoting inclusion in clinical trials—a rapid review of the literature and recommendations for action. *Trials*, 22(1): 1–11.

Bondi, E.; Xu, L.; Acosta-Navas, D.; and Killian, J. A. 2021. Envisioning Communities: A Participatory Approach Towards AI for Social Good. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 425–436. Virtual Event USA: ACM. ISBN 978-1-4503-8473-5.

Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Friedler, S. A.; and Wilson, C., eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 77–91. PMLR.

Carrotte, E. R.; Vella, A. M.; Bowring, A. L.; Douglass, M. S.; Hellard, M. E.; Stoové, M. A.; and Pedrana, A. E. 2016. "I am yet to encounter any survey that actually reflects

my life": a qualitative study of inclusivity in sexual health research. *BMC Medical Research Methodology*, 16(1): 86.

Centers for Disease Control and Prevention. 2023. Diabetes - Advancing Health Equity. https://www.cdc.gov/diabetes/health-equity/index.html. Accessed: 2025-03-20.

Centraal Bureau voor de Statistiek. 2024. Hoeveel mensen met een migratieachtergrond wonen in Nederland? https://www.cbs.nl/nl-nl/dossier/dossier-asiel-migratie-en-integratie/hoeveel-mensen-met-een-migratieachtergrond-wonen-in-nederland-. Accessed: 2025-05-19.

Chaiyachati, K. H.; Beidas, R. S.; Lane-Fall, M. B.; Rendle, K. A.; Shelton, R. C.; and Kaufman, E. J. 2022. Weaving equity into the fabric of medical research. *Journal of general internal medicine*, 37(8): 2067–2069.

Chaki, J.; Ganesh, S. T.; Cidham, S.; and Theertan, S. A. 2022. Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. *Journal of King Saud University-Computer and Information Sciences*, 34(6): 3204–3225.

Chasalow, K.; and Levy, K. 2021. Representativeness in Statistics, Politics, and Machine Learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 77–89. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8309-7.

Chen, T.; and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–794. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342322.

Chow, E. A.; Foster, H.; Gonzalez, V.; and McIver, L. 2012. The disparate impact of diabetes on racial/ethnic minority populations. *Clinical Diabetes*, 30(3): 130–133.

Costanza-Chock, S. 2020. *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

Dastin, J. 2018. Insight - Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG. Accessed: 2025-05-19.

Davis, J.; and Goadrich, M. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, 233–240.

Davis, J. L.; Williams, A.; and Yang, M. W. 2021. Algorithmic reparation. *Big Data & Society*, 8(2).

De Castro, P.; Heidari, S.; and Babor, T. F. 2016. Sex and gender equity in research (SAGER): reporting guidelines as a framework of innovation for an equitable approach to gender medicine. *Annali dell'Istituto superiore di sanita*, 52(2): 154–157.

Dinh, A.; Miertschin, S.; Young, A.; and Mohanty, S. D. 2019. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making*, 19(1): 211.

Dion, A.; Klevor, A.; Nakajima, A.; and Andersson, N. 2021. Evidence-based priorities of under-served pregnant and parenting adolescents: addressing inequities through a participatory approach to contextualizing evidence syntheses. *International Journal for Equity in Health*, 20(1): 118.

Eliason, M. J. 2014. An Exploration of Terminology Related to Sexuality and Gender: Arguments for Standardizing the Language. *Social Work in Public Health*, 29(2): 162–175.

Farran, B.; Channanath, A. M.; Behbehani, K.; and Thanaraj, T. A. 2013. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study. *BMJ open*, 3(5).

Foster, H. M.; Celis-Morales, C. A.; Nicholl, B. I.; Petermann-Rocha, F.; Pell, J. P.; Gill, J. M.; O'Donnell, C. A.; and Mair, F. S. 2018. The effect of socioeconomic deprivation on the association between an extended measurement of unhealthy lifestyle factors and health outcomes: a prospective analysis of the UK Biobank cohort. *The Lancet Public Health*, 3(12): e576–e585.

Frohard-Dourlent, H.; Dobson, S.; Clark, D. B.; Doull, M.; and Saewyc, E. M. 2017. "I would have preferred more options": Accounting for non-binary youth in health research. *Nursing Inquiry*, 24(1): e12150.

Grintsova, O.; Maier, W.; and Mielck, A. 2014. Inequalities in health care among patients with type 2 diabetes by individual socio-economic status (SES) and regional deprivation: a systematic literature review. *International journal for equity in health*, 13(1): 1–14.

Gross, A. S.; Harry, A. C.; Clifton, C. S.; and Della Pasqua, O. 2022. Clinical trial diversity: An opportunity for improved insight into the determinants of variability in drug response. *British Journal of Clinical Pharmacology*, 88(6): 2700–2717.

Haug, C. J.; and Drazen, J. M. 2023. Artificial intelligence and machine learning in clinical medicine, 2023. *New England Journal of Medicine*, 388(13): 1201–1208.

He, H.; and Ma, Y. 2013. *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons.

Hoffman, K. M.; Trawalter, S.; Axt, J. R.; and Oliver, M. N. 2016. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences of the United States of America*, 113(16): 4296–4301.

Holdcroft, A. 2007. Gender bias in research: how does it affect evidence based medicine? *Journal of the Royal Society of Medicine*, 100(1): 2–3.

Hosmer, D. 2000. Lemeshow S. Applied logistic regression. *New York*.

Ibrahim, H.; Liu, X.; Zariffa, N.; Morris, A. D.; and Denniston, A. K. 2021. Health data poverty: an assailable barrier to equitable digital health care. *The Lancet Digital Health*, 3(4): e260–e265.

Kelsey, M. D.; Patrick-Lake, B.; Abdulai, R.; Broedl, U. C.; Brown, A.; Cohn, E.; Curtis, L. H.; Komelasky, C.;

Mbagwu, M.; Mensah, G. A.; et al. 2022. Inclusion and diversity in clinical trials: Actionable steps to drive lasting change. *Contemporary Clinical Trials*, 106740.

Kent, D. M.; Rothwell, P. M.; Ioannidis, J.; Altman, D. G.; and Hayward, R. A. 2010. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials*, 11(1): 1–11.

Kent, D. M.; Van Klaveren, D.; Paulus, J. K.; D'Agostino, R.; Goodman, S.; Hayward, R.; Ioannidis, J. P.; Patrick-Lake, B.; Morton, S.; Pencina, M.; et al. 2020. The predictive approaches to treatment effect heterogeneity (PATH) statement: explanation and elaboration. *Annals of internal medicine*, 172(1): W1–W25.

Klijs, B.; Scholtens, S.; Mandemakers, J. J.; Snieder, H.; Stolk, R. P.; and Smidt, N. 2015. Representativeness of the LifeLines cohort study. *PloS one*, 10(9): e0137203.

Koo, B. K.; Roh, E.; Yang, Y. S.; and Moon, M. K. 2016. Difference between old and young adults in contribution of β-cell function and sarcopenia in developing diabetes mellitus. *Journal of diabetes investigation*, 7(2): 233–240.

Lai, H.; Huang, H.; Keshavjee, K.; Guergachi, A.; and Gao, X. 2019. Predictive models for diabetes mellitus using machine learning techniques. *BMC endocrine disorders*, 19(1): 1–9.

Larrazabal, A. J.; Nieto, N.; Peterson, V.; Milone, D. H.; and Ferrante, E. 2020. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23): 12592–12594.

LaVeist, T. A.; Gaskin, D.; and Richard, P. 2011. Estimating the economic burden of racial health inequalities in the United States. *International Journal of Health Services*, 41(2): 231–238.

Little, R. J.; and Rubin, D. B. 2019. *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Louie, P.; and Wilkes, R. 2018. Representations of race and skin tone in medical textbook imagery. *Social Science & Medicine*, 202: 38–42.

Mccradden, M.; Odusi, O.; Joshi, S.; Akrout, I.; Ndlovu, K.; Glocker, B.; Maicas, G.; Liu, X.; Mazwi, M.; Garnett, T.; Oakden-Rayner, L.; Alfred, M.; Sihlahla, I.; Shafei, O.; and Goldenberg, A. 2023. What's fair is... fair? Presenting JustEFAB, an ethical framework for operationalizing medical ethics and social justice in the integration of clinical machine learning: JustEFAB. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 1505–1519. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.

Miceli, M.; Posada, J.; and Yang, T. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP).

Nanney, M. S.; Myers, S. L.; Xu, M.; Kent, K.; Durfee, T.; and Allen, M. L. 2019. The Economic Benefits of Reducing Racial Disparities in Health: The Case of Minnesota. *International Journal of Environmental Research and Public Health*, 16(5): 742.

National Academies of Sciences, Engineering, and Medicine. 2022. *Improving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups*. Washington, DC: The National Academies Press. ISBN 978-0-309-27820-1.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

Oh, S. S.; Galanter, J.; Thakur, N.; Pino-Yanes, M.; Barcelo, N. E.; White, M. J.; de Bruin, D. M.; Greenblatt, R. M.; Bibbins-Domingo, K.; Wu, A. H.; et al. 2015. Diversity in clinical and biomedical research: a promise yet to be fulfilled. *PLoS medicine*, 12(12): e1001918.

Ovalle, A.; Subramonian, A.; Gautam, V.; Gee, G.; and Chang, K.-W. 2023. Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 496–511. Montr\'{e}al QC Canada: ACM. ISBN 9798400702310.

Park, J. S.; Bernstein, M. S.; Brewer, R. N.; Kamar, E.; and Morris, M. R. 2021. Understanding the Representation and Representativeness of Age in AI Data Sets. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, 834–842. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384735.

Pereira Barata, A.; Takes, F. W.; van den Herik, H. J.; and Veenman, C. J. 2019. Imputation Methods Outperform Missing-Indicator for Data Missing Completely at Random. In *2019 International Conference on Data Mining Workshops (ICDMW)*, 407–414.

Perez, C. C. 2019. *Invisible women: Data bias in a world designed for men*. Abrams.

Plevkova, J.; Brozmanova, M.; Harsanyiova, J.; Sterusky, M.; Honetschlager, J.; and Buday, T. 2020. Various aspects of sex and gender bias in biomedical research. *Physiological research*, 69(Suppl 3): S367.

Roldós, M. I.; and Breen, N. 2021. Using Economic Evaluation to Hasten Health Equity. *Health Equity*, 5(1): 627–632.

Rothman, K. J.; and Greenland, S. 2005. Causation and causal inference in epidemiology. *American journal of public health*, 95(S1): S144–S150.

Santema, B. T.; Ouwerkerk, W.; Tromp, J.; Sama, I. E.; Ravera, A.; Regitz-Zagrosek, V.; Hillege, H.; Samani, N. J.; Zannad, F.; Dickstein, K.; et al. 2019. Identifying optimal doses of heart failure medications in men compared with women: a prospective, observational, cohort study. *The Lancet*, 394(10205): 1254–1263.

Sardar, M. R.; Badri, M.; Prince, C. T.; Seltzer, J.; and Kowey, P. R. 2014. Underrepresentation of women, elderly patients, and racial minorities in the randomized trials used for cardiovascular guidelines. *JAMA internal medicine*, 174(11): 1868–1870.

Saria, S.; and Goldenberg, A. 2015. Subtyping: What it is and its role in precision medicine. *IEEE Intelligent Systems*, 30(4): 70–75.

Scholtens, S.; Smidt, N.; Swertz, M. A.; Bakker, S. J.; Dotinga, A.; Vonk, J. M.; Van Dijk, F.; van Zon, S. K.; Wijmenga, C.; Wolffenbuttel, B. H.; et al. 2015. Cohort Profile: LifeLines, a three-generation cohort study and biobank. *International journal of epidemiology*, 44(4): 1172–1180.

Scott, D.; Derrett, S.; Rupel, V. P.; Jelsma, J.; Gurung, G.; Oduro, G. Y.; and Withey-Rila, C. 2025. He/She/They - Gender Inclusivity in Developing and Using Health-Related Questionnaires: A Scoping Review. *Quality of Life Research*, 34(1): 67–87.

Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. Atlanta GA USA: ACM. ISBN 978-1-4503-6125-5.

Seyyed-Kalantari, L.; Zhang, H.; McDermott, M. B. A.; Chen, I. Y.; and Ghassemi, M. 2021. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12): 2176–2182.

Sharma, A.; and Palaniappan, L. 2021. Improving diversity in medical research. *Nature Reviews Disease Primers*, 7(1): 74.

Sharrocks, K.; Camidge, D.; and Papa, S. 2014. The impact of socioeconomic status on access to cancer clinical trials. *British journal of cancer*, 111(9): 1684–1687.

Stolk, R. P.; Rosmalen, J. G.; Postma, D. S.; de Boer, R. A.; Navis, G.; Slaets, J. P.; Ormel, J.; and Wolffenbuttel, B. H. 2008. Universal risk factors for multifactorial diseases: LifeLines: a three-generation population-based study. *European journal of epidemiology*, 23: 67–74.

Tobb, K.; Kocher, M.; and Bullock-Palmer, R. P. 2022. Underrepresentation of women in cardiovascular trials-it is time to shatter this glass ceiling. *American Heart Journal Plus: Cardiology Research and Practice*, 100109.

Turner, B. E.; Steinberg, J. R.; Weeks, B. T.; Rodriguez, F.; and Cullen, M. R. 2022. Race/ethnicity reporting and representation in US clinical trials: a cohort study. *The Lancet Regional Health–Americas*, 11.

University of Groningen; Lifelines. 2025. General Cohort & Factsheets. https://wiki-lifelines.web.rug.nl/doku.php?id=cohort. Accessed: 2025-04-28.

Vallor, S. 2024. *The AI mirror: How to reclaim our humanity in an age of machine thinking*. Oxford University Press.

van der Ende, M. Y.; Hartman, M. H. T.; Hagemeijer, Y.; Meems, L. M. G.; de Vries, H. S.; Stolk, R. P.; de Boer, R. A.; Sijtsma, A.; van der Meer, P.; Rienstra, M.; and van der Harst, P. 2017. The LifeLines Cohort Study: Prevalence and treatment of cardiovascular disease and risk factors. *International Journal of Cardiology*, 228: 495–500.

van der Meer, T. P.; Wolffenbuttel, B. H. R.; and Patel, C. J. 2021. Data-driven assessment, contextualisation and implementation of 134 variables in the risk for type 2 diabetes: an analysis of Lifelines, a prospective cohort study in the Netherlands. *Diabetologia*, 64(6): 1268–1278.

van Ryn, M.; Burgess, D. J.; Dovidio, J. F.; Phelan, S. M.; Saha, S.; Malat, J.; Griffin, J. M.; Fu, S. S.; and Perry, S. 2011. The Impact of Racism on Clinician Cognition, Behavior, and Clinical Decision Making. *Du Bois review : social science research on race*, 8(1): 199–218.

Viering, T.; and Loog, M. 2022. The shape of learning curves: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 7799–7819.

Vinke, P. C.; Navis, G.; Kromhout, D.; and Corpeleijn, E. 2020. Socio-economic disparities in the association of diet quality and type 2 diabetes incidence in the Dutch Lifelines cohort. *EClinicalMedicine*, 19: 100252.

Wallerstein, N.; Duran, B.; Oetzel, J. G.; and Minkler, M. 2017. *Community-Based Participatory Research for Health: Advancing Social and Health Equity*. John Wiley & Sons. ISBN 978-1-119-25885-8.

Wasserstein, R. L.; Schirm, A. L.; and Lazar, N. A. 2019. Moving to a world beyond "$p < 0.05$". *The American Statistician*, 73(sup1): 1–19.

Wensink, M.; Westendorp, R. G.; and Baudisch, A. 2014. The causal pie model: an epidemiological method applied to evolutionary biology and ecology. *Ecology and Evolution*, 4(10): 1924–1930.

World Health Organization. 2011. Use of glycated haemoglobin (HbA1c) in diagnosis of diabetes mellitus: abbreviated report of a WHO consultation. https://apps.who.int/iris/handle/10665/70523. Accessed: 2023-01-30.

World Health Organization. 2024. Health equity. https://www.who.int/health-topics/health-equity. Accessed: 2025-05-19.

World Health Organization. 2025. World Report on Social Determinants of Health Equity. https://www.who.int/publications/i/item/9789240107588. Accessed: 2025-05-19, License: CC BY-NC-SA 3.0 IGO.

World Health Organization; and Federation, I. D. 2006. Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia : report of a WHO/IDF consultation. https://apps.who.int/iris/handle/10665/43588. Accessed: 2023-01-30.

Yan, Z.; Cai, M.; Han, X.; Chen, Q.; and Lu, H. 2023. The Interaction Between Age and Risk Factors for Diabetes and Prediabetes: A Community-Based Cross-Sectional Study. *Diabetes, Metabolic Syndrome and Obesity*, 85–93.

Yu, K.-H.; Beam, A. L.; and Kohane, I. S. 2018. Artificial intelligence in healthcare. *Nature biomedical engineering*, 2(10): 719–731.