RMExplorer: A Visual Analytics Approach to Explore the Performance and the Fairness of Disease Risk Models on Population Subgroups

Bum Chul Kwon*
IBM Research

Uri Kartoun
IBM Research

Shaan Khurshid
Broad Institute

Mikhail Yurochkin
IBM Research

Subha Maity
University of Michigan

Deanna G Brockman
Broad Institute

Amit V Khera
Broad Institute

Patrick T Ellinor
Broad Institute

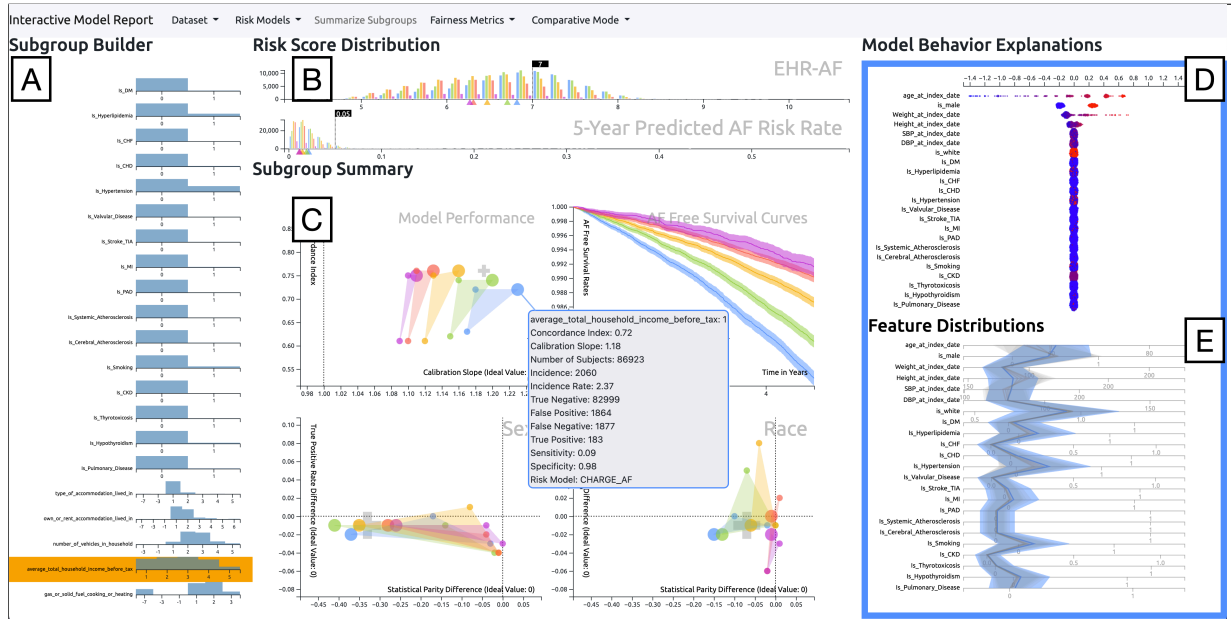Steven A Lubitz
Broad Institute

Kenney Ng†
IBM Research

Figure 1: RMExplorer: (A) Subgroup builder shows the distribution of patients by variables that can be used to create subgroups; (B) Risk score distribution displays the summary of risk scores and the estimated disease onset rate; (C) Subgroup summary presents the model performance (top) and the fairness of subgroups (bottom); (D) Model behavior explanations provide feature contributions to the risk scores; (E) Feature distributions show the mean and confidence intervals of a selected subgroup across variables.

ABSTRACT

Disease risk models can identify high-risk patients and help clinicians provide more personalized care. However, risk models developed on one dataset may not generalize across diverse subpopulations of patients in different datasets and may have unexpected performance. It is challenging for clinical researchers to inspect risk models across different subgroups without any tools. Therefore, we developed an interactive visualization system called RMExplorer (Risk Model Explorer) to enable interactive risk model assessment. Specifically, the system allows users to define subgroups of patients by selecting clinical, demographic, or other characteristics, to explore the performance and fairness of risk models on the subgroups, and to understand the feature contributions to risk scores. To demonstrate the usefulness of the tool, we conduct a case study, where we use RMExplorer to explore three atrial fibrillation risk models by applying them to the UK Biobank dataset of 445,329 individuals. RMExplorer can help researchers to evaluate the performance and biases of risk models on subpopulations of interest in their data.

*e-mail: bumchul.kwon@us.ibm.com
†e-mail: kenney.ng@us.ibm.com

Keywords: visual analytics, health informatics, fairness, subgroup analysis, explainability, interpretability, electronic health records

Index Terms: Human-centered computing—Visualization

1 INTRODUCTION

Disease risk models trained within a specific dataset may not generalize well to new, unseen datasets in which patient characteristics may differ - including observation periods, demographics, and healthcare regulations/policies - from the training dataset. In addition, the model may learn and propagate inherent, systematic biases captured in the training data, thus amplifying fairness issues. For instance, machine learning models trained with data at a university hospital, which often includes a majority of patients with higher income, younger age, and white race, may not be generalizable to more diverse populations [7]. Medical decisions that are based on biased models may lead to worse outcomes disproportionately affecting specific racial, gender, and socioeconomic groups [21]. Therefore, it is important for clinicians to understand the machine learning models in terms of both performance and fairness. Furthermore, because patient populations can be very heterogeneous, it is also important to understand how the model behaves for different subpopulations.

Recently, several approaches were proposed to summarize the fairness of model outcomes for given attributes. For instance, the AI Fairness 360 Toolkit [4] is an open-source framework that includes various model-agnostic fairness metrics. However, it is computa-

tionally challenging to generate performance and fairness measures for all possible combinations of subgroups. Furthermore, it is cognitively demanding for users to analyze the results. Therefore, it is important to develop tools to help clinical investigators build subgroups and evaluate results easily. Recently, researchers investigated the concept of a model report card which summarizes how a model was trained and how it performs on a fixed set of subgroups in a static report [3, 23]. These approaches are significant milestones, yet they do not allow users to dynamically apply the model to their own datasets, define custom subgroups, and interactively explore how the model behaves across the subgroups.

Visual analytics approaches are recognized as a potential solution to explore large-scale data in healthcare [5, 29]. Visual analytics, which combines computational methods with interactive visualizations, can be used to generate insights and knowledge from data and to enhance trust [6, 18, 25, 26]. Interactive visualizations powered by machine learning algorithms on longitudinal health records have demonstrated usefulness for understanding disease subtypes [14], analyzing heterogeneous disease progression patterns [12], and predicting health outcomes [13]. Recently, visual analytics applications are developed to explore fairness in ranking decision [1], image classification [15], and others. However, previous studies have not extensively studied visual analytics approaches to investigate the performance and fairness of risk models on healthcare data.

To address the important, unmet need, we propose a visual analytics approach called RMExplorer (Risk Model Explorer) to facilitate the exploration of performance heterogeneity and fairness. The system includes interactive visualizations combined with various computational approaches. The system can be applied to explore the performance and fairness of any risk models and any appropriate datasets. To demonstrate the applicability and usefulness of the system, we provide a use case on a large prospective cohort dataset (UK BioBank [27]) with validated atrial fibrillation (AF) risk models, namely, EHR-AF [11], CHARGE-AF [2], and C_2HEST [17].

2 VISUAL ANALYTICS SYSTEM DESIGN

RMExplorer allows clinical researchers to apply existing risk models to their patient datasets, to visually explore the performance and fairness of models within subgroups defined by users, and to probe why the model predicts different outcomes for each subgroup. In the following sections, we describe the design of RMExplorer.

2.1 Users, Tasks, and Requirements

All authors of the study in diverse disciplines, including cardiology, machine learning, and visualization, characterized the tasks and requirements of disease risk model analysis. The intended users, clinical researchers and data scientists, want to evaluate disease risk models. First, they assess the performance of the model by computing risk scores and prediction accuracy. The risk scores are converted to estimated disease onset rates so that users can compare them with other previously reported onset rates. Second, users inspect inherent biases of the scores with respect to various patient subpopulations that they define using various demographic and clinical variables. Third, they also comparatively analyze different metrics for different risk models to test the robustness of the measures. The process repeats iteratively until they gain insights into the risk models.

To support the analysis, the system requires interactive visualizations, where users can view the summary of various metrics about performance and fairness. In particular, researchers need to define subgroups and compute various metrics on each of them so that they understand the risk models that are applied to the subpopulations.

2.2 Subgroup Builder

Subgroup Builder shows the distribution of patients over a specified set of variables, such as comorbidities, age, sex, and race. The initial view shows the overall summary of the entire population. Users then can select a set of variables which will be used to create subgroups. For instance, if users want to create a subgroup based on hyperlipidemia disease status in the health records, they can select the variable 'is_Hyperlipidemia'. If users select more than one variable, the system will generate all possible combinations of the selected variables, each of which becomes a mutually exclusive group of patients that satisfy that combination of conditions. Subgroups without any patients are automatically discarded. Figure 1 (A) shows that the user chose one variable with five discrete levels, which subsequently generated five subgroups for analysis in other views.

2.3 Risk Score Distribution

RMExplorer requires users to provide risk models that are previously trained. The models are provided as functions that generate risk scores and probabilities of diseases during a follow-up time window (e.g., 5 years since the baseline). RMExplorer can then compute the metrics for selected patients using the model covariates on the fly.

Risk Score Distribution (Figure 1 (B)) shows distributions of a chosen risk score (top) and its predicted risk rate at a future date (bottom) for the selected subpopulation. Each of the two charts provides a threshold handle that represents a constant value to convert the continuous risk score to the binary predicted label (high-risk vs low-risk). The two thresholds are synchronized: if users adjust one threshold then the other is adjusted accordingly. As users update the threshold, the system recalculates the predicted labels, which are used by the performance and fairness metrics in Subgroup Summary. Users can at any point switch to other risk models, which resets the view with the chosen scores and predicted risk rates. Figure 1 (B) shows the distribution of EHR-AF, a risk model that predicts the time to onset of atrial fibrillation since the baseline of covariates measured, and 5-year predicted disease risk rate.

2.4 Subgroup Summary: Performance

Subgroup Summary presents the performance (the two charts on top) and the fairness (the other two charts at the bottom) of the risk score model on the subgroups defined by users as Figure 1 (C) shows.

The risk model performance is shown in two different plots located on top. The first chart on the left initially shows a scatter plot of subgroups as points, colored according to subgroups, on two-dimensional space with calibration slope on the x-axis and concordance index on the y-axis. Concordance index, also known as Harrell's C-index, is a measure that evaluates the discrimination performance of risk models in survival analysis, where data are censored (e.g., lost to follow up) [9]. Calibration slope is a measure of the relation between predicted risk and observed incidence, where a value of one indicates optimal calibration. Among models that are reasonably calibrated-in-the-large, values higher than 1 indicate underestimation while values lower than 1 indicate overestimation of the actual risk [28]. With these two measures, users can gain insights into how well the model discriminates disease risk in each subgroup and how well the model estimates are calibrated within each subgroup. Users can expand the view by showing other risk models in the view, which creates polygons, colored according to subgroups (Figure 1 (C)). The vertices of each polygon represent the respective risk models. By switching to the comparative mode, users can investigate the similarities and differences of multiple risk models across and within subgroups.

The second chart on the right shows a disease-free survival plot stratified by subgroups. The plot is generated by fitting the patients' disease-free survival data using the Kaplan-Meier method. By viewing the chart, users can understand the relative differences in estimated risks over time among different subgroups.

2.5 Subgroup Summary: Fairness

The two charts at the bottom show fairness metrics for subgroups that users have chosen. Our system supports group and individual

fairness metrics, and users can choose any two at a time. As Figure 1 (C) shows, the system provides two group fairness metrics: statistical parity difference on the x-axis and true positive rate difference on the y-axis computed using the AI Fairness 360 Toolkit [4]. For group fairness metrics, each of the charts presents the fairness in terms of a protected attribute. A protected attribute (e.g., gender, race, socioeconomic level) is an application-specific, user-selected attribute of patients across the categories of which it is desirable to achieve equality in terms of benefit received. A value of 0 for these measures indicates an equal performance of the model for any given protected attribute. Let's say we define race as a target protected attribute, then statistical parity difference computes the difference between the proportion of the predicted patients diagnosed (by the risk model) with the target disease for patients with nonwhite race and that for patients with white race within each subgroup. True positive rate difference measures the difference in the accuracy of risk models that predict the disease of patients between the nonwhite and the white race groups within each subgroup. For individual fairness, we compute the individual fairness violation rate, which provides the proportion of individuals whose outcome predictions are sensitive to the protected attributes [20, 30]. By viewing the fairness charts, users can understand the differences between subgroups in terms of key fairness metrics for each protected attribute.

Similar to the first chart of the model performance, the fairness plot initially shows a scatterplot of points, colored by subgroups, over the two dimensions of selected fairness measures. Users can switch to the comparative mode so that the view shows polygons, colored by subgroups, over the two dimensions.

2.6 Model Behavior Explanations

Model Behavior Explanations aim to show why the model behaves the way it does for a selected subgroup. The view opens when users select a subgroup in the Subgroup Summary. The view includes two visualizations: the SHAP (SHapley Additive exPlanations) beeswarm plot and the feature distribution chart.

SHAP is a model-agnostic interpretability approach that generates a feature importance value, called a SHAP value, for each prediction [19]. Once the tool generates the SHAP values for all predictions made for the patients in the selected subgroup, it plots them using the beeswarm plot. The plot shows dots (patients) per each row (a feature). The color of each dot indicates the normalized feature value, ranging from minimum (blue) to maximum (red) for the corresponding feature. The horizontal position of each dot represents the SHAP value. The vertical position is jittered to show the distribution of points along the x-axis without occlusion. To avoid the visual clutter from overplotting and to speed up computation, the tool randomly samples a fraction (i.e., 10%) of the subgroup population to generate the beeswarm plot. Figure 1 (D) shows the relationship between SHAP values and feature values of 23 variables.

To summarize the feature distributions of the selected subgroups, we use a variation of parallel coordinates, called Parallel Trends, which shows a central measure (e.g., mean) and a dispersion (e.g., standard deviation) of a group of instances with band crossing axes [14]. The view highlights the selected subgroup while displaying other subgroups in a grey color. Using the visualizations, users can compare the feature distributions across multiple subgroups. Figure 1 (E) shows the parallel trends of the blue subgroup in the context of other subgroups in the background.

3 CASE STUDY: ATRIAL FIBRILLATION RISK MODELS

In this section, we provide a case study on risk models of atrial fibrillation (AF), a common cardiovascular condition that is a leading cause of stroke and heart failure in the elderly. The case study was performed by all authors, where the leading author mainly operated the tool, and all authors, including the five domain experts who are actively involved in clinical research in cardiac disease risk
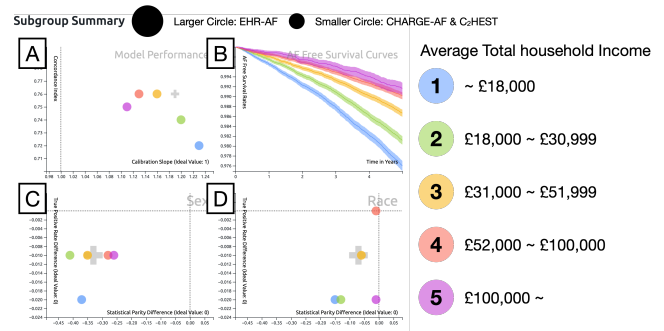

Figure 2: The subgroup summary shows differences in model performance and fairness among the five income subgroups: (a) model performance; (b) atrial fibrillation (AF) free survival curves; (c) fairness on sex; (d) fairness on race.

models, analyzed the results using visualizations. In this case study, we used three representative AF risk score models: EHR-AF [11], CHARGE-AF [2], and C_2HEST [17], each of which provides scores for individual patients and the 5-year risk of AF using their respective covariates, such as demographics (e.g., age, race, sex) and AF-risk factors (e.g., hypertension, hyperlipidemia, heart failure).

For the dataset, we extracted a patient cohort from the UK Biobank (research application 50658), an observational study that enrolled over 500,000 individuals between 2006 and 2010 [27]. To be consistent with the original design of CHARGE-AF, we excluded individuals under age 45 at enrollment, individuals with prevalent AF, and individuals with missing data for height, weight, systolic blood pressure (SBP), and diastolic blood pressure (DBP) at baseline (Table A.1). The final cohort contains 445,329 patients with 7,407 (1.66%) having incident AF within 5 years after their enrollment.

We hypothesized that the risk model may estimate AF risk disproportionately with respect to an individual patient's socioeconomic status because there may be association between disease incidence and household income [8, 16, 22]. The variable includes five income ranges as discrete integer levels: 1) less than £18,000; 2) £18,000 to £30,999; 3) £31,000 to £51,999; 4) £52,000 to £100,000; 5) Greater than £100,000. We discarded those who did not report their income by selecting "Do not know" and "Prefer not to answer".

A user launches RMExplorer to explore the three AF risk models. Using Subgroup Builder, they selected the variable called "Annual Total Household Income Before Tax". We select EHR-AF as the risk model to begin the analysis. As the user clicks on "Summarize Subgroups", Subgroup Summary provides four visualizations of the five subgroups as Figure 2 shows. First, the model performance view shows that concordance index is lower for patients in lower income groups (blue and green) and generally higher for those with higher income (Figure 2 (A)). Similarly, the calibration slope for the higher income group is closer to the ideal value (the dotted, vertical line), 1, but the lower income group is more distant from 1. In other words, users can hypothesize that the EHR-AF risk model makes more accurate decisions for higher income groups. In addition, the AF free survival curves in Figure 2 (B) show that the higher income groups have longer AF-free survival than the others.

The fairness plots summarize the fairness metrics of the models for the income subgroups. The fairness plot on Sex (Figure 2 (C)) shows that all subgroups have negative values for both true positive rate differences and statistical parity differences. This indicates that the true positive rates for females are lower than those for males, and the proportion of predicted AF for females is lower than males for all subgroups. Similarly, the fairness plot on Race (Figure 2 (D)) shows that subgroups tend to have negative values for both metrics as well. Interestingly, the lower income groups show greater disparities in the statistical parity difference between white and non-white patients than the higher income groups.

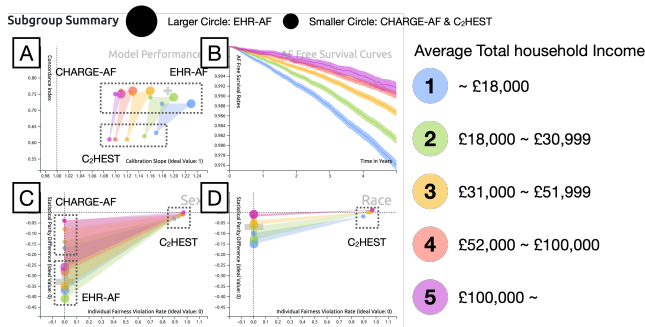What about the two other risk models? The user switches to the

Figure 3: The subgroup summary compares EHR-AF, CHARGE-AF, and C$_2$HEST. The polygon areas with distinct colors show differences in the scores across subgroups, and the coordinate location of vertices per each polygon summarizes the differences in scores among risk models within each subgroup.

Figure 4: The fairness plots show differences when the 5-year predicted risk threshold is adjusted from 0.05 to 0.08.

comparative mode to explore the differences among risk models. Figure 3 (A) illustrates that the five polygons are shown next to each other, starting from low income groups (right) to high income groups (left). It also shows that each triangle has its vertices in relatively consistent positions, as CHARGE-AF on top-left, EHR-AF on top-right, and C$_2$HEST at the bottom. It shows that C$_2$HEST has lower performance than the other risk models consistently across subgroups. As Figure 3 (C) shows, the fairness scores of the risk models are separated in their relative positions within each subgroup with CHARGE-AF on top left, EHR-AF on bottom left, and C$_2$HEST on top right. C$_2$HEST shows very high values for individual fairness violation rates, close to 0.9, compared to other risk scores, close to 0. This indicates that C$_2$HEST is likely to change predictions for individuals if the input values are perturbed with respect to sex. The similar insight can be drawn for the race attribute in Figure 3 (D).

The user clicks through subgroups to explore their behaviors in the SHAP plot and the feature summary. The SHAP plots show that the age of patients is the most influential factor in risk estimation across the income subgroups. By their distributions, they can see that the first four variables, age, sex, weight, and height, have non-zero values of SHAP, which indicates their relative importance to the EHR-AF score. As the plot shows, the estimated risk is higher for older patients. The feature summary plots show that age is indeed the variable that differentiates the income subgroups, as lower income groups have relatively higher mean age. The feature distribution shows that height is lower for lower income groups (1,2) than other income groups. With the consistent mean weight across subgroups, the user speculates that lower income groups might have higher body mass indices (BMIs), which may lead to higher AF risk scores.

To explore how fairness measures vary as a function of 5-year predicted AF risk thresholds and income levels, the user decides to experiment with multiple threshold values on Risk Score Distribution and observe their results (Figure 4). By default, the system shows summary results with 0.05 as the 5-year risk score threshold because it is a standard threshold used in previous studies [11]. The user increased the value to 0.08. The updated fairness plots (Figure 4 bottom) show some differences in the three different risk models among the five income-based subgroups after the threshold adjustment has been made. In EHR-AF, the true positive rate differences for females are now higher than or equal to those for males in the income subgroups 1, 2, 3, and 4. This means that the model with the updated threshold makes decisions for female patients more accurately than or at a similar level to male patients in those subgroups. The fairness plot on Race shows the gap that previously existed on statistical parity (x-axis) is now narrower than before as all subgroups are closer to 0. On the other hand, in CHARGE-AF, the updated threshold vertically widened the spread of subgroups with respect to true positive rate differences in sex and race. However, it also narrowed the spread in statistical parity difference. For
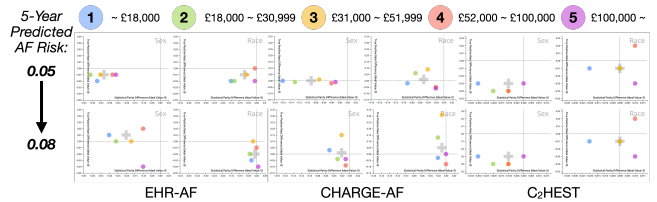
C$_2$HEST, the user does not find any differences before and after the updated thresholds for both fairness measures with respect to sex or race. The experiment shows that the updated threshold can affect the fairness metrics differently for different risk models across the income subgroups.

4 DISCUSSION AND FUTURE WORK

This study presents RMExplorer that allows users to define subgroups, assess risk model performance and fairness based on the subgroups, and understand the model behavior with feature distributions and feature importance measures. The multiple, coordinated visualizations can help users gain insights into how the model performs differently for different subgroups. The case study on atrial fibrillation risk models demonstrates the usefulness of the interactive visualizations combined with various computational approaches for generating clinically relevant findings. The interactive visualizations provide some design implications for tools that support subgroup creation and exploration with respect to risk models. Subgroups are the keys to examining the risk models, and visualizations need to provide users with abilities to create, modify, and compare subgroups by various means. In this study, we provided the subgroup builder, where users can create customizable mutually exclusive subgroups. Future work can investigate various methods to computationally discover and recommend subgroups for further analysis. Interactive visualizations need to summarize the subgroups using patient characteristics and allow users to modify them if necessary. In addition, we present polygons in scatter plots to enable comparison of multiple risk models. The approach can be revisited and improved for a greater number of subgroups.

The visualizations show the performance and fairness of subgroups using various metrics. Our work provides an example using three risk score models for survival analysis, which are commonly used in clinical research. The survival curve is a specific visualization that is tailored for this type of analysis. However, the same visualization strategy can be applied to other prediction models with different performance metrics (e.g., AUC score). The fairness plots are designed to show disparity of subgroups in terms of two group fairness and an individual fairness metrics. Future work can investigate the usefulness of other fairness metrics.

The SHAP plot and the feature distribution summary provide explanations as to why the model behaves in certain ways for different subgroups. There are various other explanation approaches (e.g., LIME [24]), which can be added to the visualization pipeline with additional coding. In addition to the model-agnostic techniques, one may derive some explanations directly from the model when it uses the attention mechanism for deep learning models [10]. Then, the visualizations need to be adapted to show the attention scores for relevant input features instead. Future work can investigate various ways to combine the attention-based model outputs with other important metrics for the subgroup analysis. Future work can also evaluate RMExplorer against other visual analytics systems, such as FairSight [1], for model performance and fairness inspection tasks.

REFERENCES

[1] Y. Ahn and Y.-R. Lin. FairSight: Visual Analytics for Fairness in Decision Making. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1086–1095, 2020.

[2] A. Alonso, B. P. Krijthe, T. Aspelund, K. A. Stepas, M. J. Pencina, C. B. Moser, M. F. Sinner, N. Sotoodehnia, J. D. Fontes, A. C. J. W. Janssens, R. A. Kronmal, J. W. Magnani, J. C. Witteman, A. M. Chamberlain, S. A. Lubitz, R. B. Schnabel, S. K. Agarwal, D. D. McManus, P. T. Ellinor, M. G. Larson, G. L. Burke, L. J. Launer, A. Hofman, D. Levy, J. S. Gottdiener, S. Kääb, D. Couper, T. B. Harris, E. Z. Soliman, B. H. C. Stricker, V. Gudnason, S. R. Heckbert, and E. J. Benjamin. Simple Risk Model Predicts Incidence of Atrial Fibrillation in a Racially and Geographically Diverse Population: the CHARGE-AF Consortium. *Journal of the American Heart Association*, 2(2):e000102, 2013. Publisher: American Heart Association. doi: 10.1161/JAHA.112.000102

[3] M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. N. Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, and K. R. Varshney. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5):6:1–6:13, July 2019. Conference Name: IBM Journal of Research and Development. doi: 10.1147/JRD.2019.2942288

[4] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4:1–4:15, 2019.

[5] J. J. Caban and D. Gotz. Visual analytics in healthcare – opportunities and research challenges. *Journal of the American Medical Informatics Association*, 22(2):260–262, Mar. 2015. doi: 10.1093/jamia/ocv006

[6] A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren. The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations. *Computer Graphics Forum*, 39(3):713–756, 2020.

[7] M. A. Gianfrancesco, S. Tamang, J. Yazdany, and G. Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178(11):1544–1547, 2018.

[8] E. Guhl, A. Althouse, M. Sharbaugh, A. M. Pusateri, M. Paasche-Orlow, and J. W. Magnani. Association of income and health-related quality of life in atrial fibrillation. *Open Heart*, 6(1):e000974, Apr. 2019. Publisher: Archives of Disease in childhood Section: Arrhythmias and sudden death. doi: 10.1136/openhrt-2018-000974

[9] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

[10] B. Hoover, H. Strobelt, and S. Gehrmann. exBERT: A visual analysis tool to explore learned representations in Transformer models. pp. 187–196, 2020.

[11] S. Khurshid, U. Kartoun, J. M. Ashburner, L. Trinquart, A. Philippakis, A. V. Khera, P. T. Ellinor, K. Ng, and S. A. Lubitz. Performance of Atrial Fibrillation Risk Prediction Models in Over 4 Million Individuals. *Circulation: Arrhythmia and Electrophysiology*, 14(1):e008997, Jan. 2021. Publisher: American Heart Association. doi: 10.1161/CIRCEP.120.008997

[12] B. C. Kwon, V. Anand, K. A. Severson, S. Ghosh, Z. Sun, B. I. Frohnert, M. Lundgren, and K. Ng. DPVis: Visual analytics with hidden markov models for disease progression pathways. *IEEE transactions on visualization and computer graphics*, 27:3685–3700, 2020.

[13] B. C. Kwon, M.-J. Choi, J. T. Kim, E. Choi, Y. B. Kim, S. Kwon, J. Sun, and J. Choo. Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE transactions on visualization and computer graphics*, 25(1):299–309, 2019.

[14] B. C. Kwon, B. Eysenbach, J. Verma, K. Ng, C. De Filippi, W. F. Stewart, and A. Perer. Clustervision: Visual supervision of unsupervised clustering. *IEEE transactions on visualization and computer graphics*, 24(1):142–151, 2018.

[15] B. C. Kwon, J. Lee, C. Chung, N. Lee, H.-J. Choi, and J. Choo. DASH: Visual Analytics for Debiasing Image Classification via User-Driven Synthetic Data Augmentation. In M. Agus, W. Aigner, and T. Hoellt, eds., *EuroVis 2022 - Short Papers*. The Eurographics Association, 2022. doi: 10.2312/evs.20221099

[16] A. R. LaRosa, J. Claxton, W. T. O'Neal, P. L. Lutsey, L. Y. Chen, L. Bengtson, A. M. Chamberlain, A. Alonso, and J. W. Magnani. Association of household income and adverse outcomes in patients with atrial fibrillation. *Heart*, 106(21):1679–1685, Nov. 2020. Publisher: BMJ Publishing Group Ltd and British Cardiovascular Society Section: Healthcare delivery, economics and global health. doi: 10.1136/heartjnl-2019-316065

[17] Y.-G. Li, D. Pastori, A. Farcomeni, P.-S. Yang, E. Jang, B. Joung, Y.-T. Wang, Y.-T. Guo, and G. Y. H. Lip. A Simple Clinical Risk Score (C2HEST) for Predicting Incident Atrial Fibrillation in Asian Subjects: Derivation in 471,446 Chinese Subjects, With Internal Validation and External Application in 451,199 Korean Subjects. *Chest*, 155(3):510–518, Mar. 2019.

[18] S. Liu, X. Wang, M. Liu, and J. Zhu. Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*, 1(1):48–56, 2017.

[19] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pp. 4768–4777. Neural Information Processing Systems Foundation, Inc., Long Beach, CA, USA, 2017.

[20] S. Maity, S. Xue, M. Yurochkin, and Y. Sun. Statistical inference for individual fairness. *arXiv preprint arXiv:2103.16714*, 2021.

[21] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[22] J. R. Misialek, K. M. Rose, S. A. Everson-Rose, E. Z. Soliman, C. J. Clark, F. L. Lopez, and A. Alonso. Socioeconomic Status and the Incidence of Atrial Fibrillation in Whites and Blacks: The Atherosclerosis Risk in Communities (ARIC) Study. *Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease*, 3(4):e001159, Aug. 2014. doi: 10.1161/JAHA.114.001159

[23] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 220–229. Association for Computing Machinery, New York, NY, USA, Jan. 2019. doi: 10.1145/3287560.3287596

[24] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. Association for Computing Machinery, New York, NY, USA, 2016.

[25] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim. The role of uncertainty, awareness, and trust in visual analytics. *IEEE transactions on visualization and computer graphics*, 22(1):240–249, 2016.

[26] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim. Knowledge generation model for visual analytics. *IEEE transactions on visualization and computer graphics*, 20(12):1604–1613, 2014.

[27] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, and R. Collins. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, 12(3):e1001779, Mar. 2015. Publisher: Public Library of Science. doi: 10.1371/journal.pmed.1001779

[28] B. Van Calster, D. Nieboer, Y. Vergouwe, B. De Cock, M. J. Pencina, and E. W. Steyerberg. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of clinical epidemiology*, 74:167–176, 2016.

[29] V. L. West, D. Borland, and W. E. Hammond. Innovative information visualization of electronic health record data: a systematic review. *Journal of the American Medical Informatics Association*, 22(2):330–339, Mar. 2015. doi: 10.1136/amiajnl-2014-002955

[30] M. Yurochkin, A. Bower, and Y. Sun. Training individually fair ml models with sensitive subspace robustness. 2020.