Investigating fair data acquisition for risk prediction in resource-constrained settings

IOANNA THOMA, Health and Medical Sciences, The Alan Turing Institute, United Kingdom
ELISABETH ABHAYARATNA, School of Clinical Medicine, University of New South Wales, Australia
MATTHEW SPERRIN, Centre for Health Informatics, Institute of Population Health, University of Manchester, United Kingdom
KARLA DIAZ ORDAZ, Department of Statistical Science, University College London, United Kingdom
RICARDO SILVA, Department of Statistical Science, University College London, United Kingdom
BRIEUC LEHMANN, Department of Statistical Science, University College London, United Kingdom

Clinical prediction models (CPMs) play a crucial role in precision medicine, enabling the identification of high-risk patients for targeted interventions. In many settings, additional covariates may be collected to improve risk prediction, but doing so for the entire population may not be feasible due to resource constraints. A key challenge is to determine who should receive these additional resource-intensive assessments in an efficient and equitable manner. Here, we explore policies to select which patients should be selected for additional testing based on a baseline risk estimate. We investigate these policies in the context of an integrated risk tool for cardiovascular disease. This explores how the application of a more complex, and expensive, CPM on a subset of the population can improve fairness. The proposed methodological approaches have the potential to guide future application of CPMs to prioritise patient populations who would most benefit from access to additional investigations and access to more complex CPMs.

Keywords: Risk Prediction, Fairness-Aware Data Collection, Clinical Prediction Models

Reference Format:
Ioanna Thoma, Elisabeth Abhayaratna, Matthew Sperrin, Karla Diaz Ordaz, Ricardo Silva, and Brieuc Lehmann. 2025. Investigating fair data acquisition for risk prediction in resource-constrained settings. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25).* Proceedings of Machine Learning Research, 7 pages.

Authors' Contact Information: Ioanna Thoma, Health and Medical Sciences, The Alan Turing Institute, London, United Kingdom, ithoma@turing.ac.uk; Elisabeth Abhayaratna, School of Clinical Medicine, University of New South Wales, Sydney, University of New South Wales, Sydney, Australia, e.abhayaratna@unsw.edu.au; Matthew Sperrin, Centre for Health Informatics, Institute of Population Health, University of Manchester, Manchester, United Kingdom, matthew.sperrin@manchester.ac.uk; Karla Diaz Ordaz, Department of Statistical Science, University College London, London, United Kingdom, karla.diaz-ordaz@ucl.ac.uk; Ricardo Silva, Department of Statistical Science, University College London, London, United Kingdom, ricardo.silva@ucl.ac.uk; Brieuc Lehmann, Department of Statistical Science, University College London, London, United Kingdom, b.lehmann@ucl.ac.uk.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

1 Introduction

Clinical prediction models (CPMs) provide quantitative estimates to support medical decision-making. By forecasting future health outcomes from a set of baseline covariates, CPMs aim to identify those high-risk patients that may be most likely to benefit from a targeted intervention [14]. CPMs are developed by training a statistical or machine learning model on a dataset consisting of the health outcome of interest and the available baseline covariates. Both the training dataset and the choice of model are key factors influencing the predictive utility of a CPM [10].

Often, several CPMs may be available for the same health outcome. For example, many CPMs are available for cardiovascular disease (CVD), such as QRISK [7], the Framingham risk score (FRS) [17], and the atherosclerotic CVD pooled cohort equations (ASCVD-PCE) tool [6]. These estimate the likelihood of developing CVD based on traditional risk factors such as age, sex, blood pressure, cholesterol levels, and smoking status. Meanwhile, several polygenic risk scores (PRSs) have been developed that aim to quantify an individual's genetic susceptibility to developing various forms of cardiovascular disease (e.g., [8, 11]). Moreover, there has been recent interest in *integrated risk tools*, which aim to combine PRS with more traditional CPMs in order to enhance risk prediction [13].

In addition to predictive accuracy, other considerations are important when evaluating a CPM. For instance, the potential benefits of implementing a CPM must be weighed against the costs of collecting the necessary covariates to compute risk estimates. In the case of integrated risk scores, traditional CPMs may rely on relatively inexpensive clinical data, whereas calculating a PRS requires genotyping, which remains a non-negligible cost [9]. In particular, it may not be financially feasible to offer such costly assessments to all individuals in the target population.

Fairness is another critical consideration in evaluating a CPM. Broadly, fairness refers to the equal predictive accuracy and utility across individuals or groups. Growing concerns have emerged that predictive algorithms may perform inequitably across different populations, potentially introducing bias into clinical decision-making and exacerbating existing health disparities [12]. Assessing fairness is essential for identifying and addressing health inequities that may arise from biases in data quality or patient selection, particularly for high-cost interventions [4].

In resource-constrained settings, researchers have argued that specific population subgroups should be prioritised for additional tests or interventions to promote health equity [5]. In the context of costly CPMs, the question of precisely which individuals should be prioritised for more resource-intensive assesments, however, has not yet been explored, to the best of our knowledge. In this paper, we present a case study of an integrated risk tool for CVD. We investigate simple policies to determine those individuals who should receive additional testing (i.e. genotyping) in order to obtain a more accurate risk estimate, and explore the impact on accuracy and fairness of these policies.

2 Methods

We begin with some notation and terminology. We are concerned with predicting a binary outcome $Y \in \{0,1\}$. For each individual i, we have access to a set of covariates $X_i \in \mathbb{R}^q$ and a baseline prediction $\hat{f}_i^{(b)} \in \mathbb{R}$. We wish to assess the accuracy and fairness of different risk prediction policies. Let P be a discrete protected characteristic, such as age or gender, against which we would like to evaluate fairness. A policy is a function $\pi(\hat{f}^{(b)}, P)$ that takes as input a baseline prediction and protected characteristic and returns a binary value, indicating whether the individual receives an intervention. For example, given a risk threshold δ, the baseline policy $\pi_\delta^{(b)}$ is given by $\pi^{(b)}(\hat{f}^{(b)}, P) = \mathbb{1}(\hat{f}^{(b)} \geq \delta)$. In words, this policy gives the intervention to those with a baseline prediction greater than or equal to δ, regardless of protected characteristic.

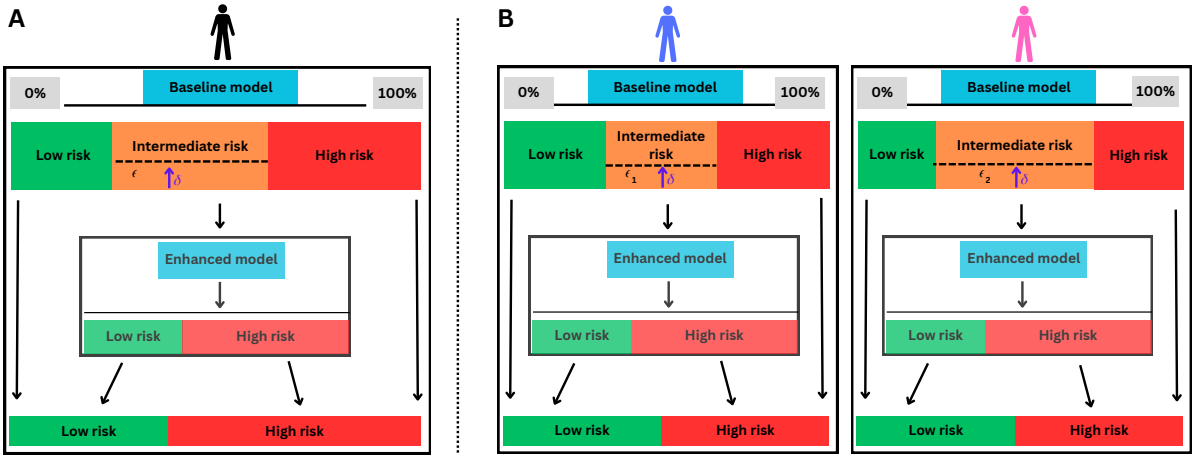

Fig. 1. (A) A schematic representation of a threshold-based policy. All individuals receive a prediction from a baseline model. Those within a tolerance ε of a threshold δ are considered 'intermediate risk' and undergo additional testing required to obtain a prediction from an enhanced model. (B) A stratified threshold-based risk reclassification policy. Here, the tolerance varies by group: blue individuals are considered 'intermediate risk' if their baseline prediction is within ε_1 of the threshold δ, while pink individuals are 'intermediate risk' if their baseline is within ε_2 of δ.

Threshold-based policies

In this work, we will focus on threshold-based policies. In short, a *simple* threshold-based policy $\pi_{\delta,\varepsilon}$ returns an enhanced prediction for individuals whose baseline risk is within a tolerance ε of a threshold δ. Let $\hat{f}_i^{(e)} \in \mathbb{R}$ denote an enhanced prediction for individual i. Then,

$$\pi_{\delta,\varepsilon}(\hat{f}^{(b)}, P) = \pi_{\delta,\varepsilon}(\hat{f}^{(b)}) = \begin{cases} \mathbb{1}(\hat{f}^{(e)} \geq \delta), & |\hat{f}^{(b)} - \delta| \leq \varepsilon \\ \mathbb{1}(\hat{f}^{(b)} \geq \delta), & |\hat{f}^{(b)} - \delta| > \varepsilon \end{cases} \tag{1}$$

See Figure 1A for a schematic of a simple threshold-based policy. In words, baseline predictions are partitioned into three categories: 'low risk' if $\hat{f}^{(b)} < \delta - \varepsilon$, 'high risk' if $\hat{f}^{(b)} > \delta - \varepsilon$, and 'intermediate risk' if $|\hat{f}^{(b)} - \delta| \leq \varepsilon$. Those considered intermediate risk then receive additional testing required to obtain an enhanced prediction $\hat{f}(e)$, which is then used to assign the intervention. Note that we do not initially have access to $\hat{f}(e)$; this is only obtained eventually for a subset of individuals according to the policy.

A *stratified* threshold-based policy is one that applies different tolerances for different values of the protected characteristic. Let n_P be the number of distinct values that the protected characteristic takes and let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_{n_P})$ be a vector of tolerance values for each value of the protected characteristic. Then,

$$\pi_{\delta,\varepsilon}(\hat{f}^{(b)}, P) = \begin{cases} \mathbb{1}(\hat{f}^{(e)} \geq \delta), & |\hat{f}^{(b)} - \delta| \leq \varepsilon_P, \\ \mathbb{1}(\hat{f}^{(b)} \geq \delta), & |\hat{f}^{(b)} - \delta| > \varepsilon_P. \end{cases} \tag{2}$$

See Figure 1B for a schematic of a stratified threshold-based policy with a binary protected characteristic. The introduction of different tolerances enables considerable flexibility in designing policies that vary according to the protected characteristic. For example, setting $\varepsilon_1 = 0$ and $\varepsilon_2 = \infty$ uses the baseline prediction for all individuals with $P = 1$ and the enhanced prediction for all individuals with $P = 2$.

3 Case study: an integrated risk tool for cardiovascular disease

We examine the fairness of different threshold-based policies using an integrated risk tool for cardiovascular disease (CVD) in the UK Biobank cohort [3], a large prospective study of approximately 500,000 individuals aged 40 to 79, with clinical and demographic data collected at baseline. Our analysis is restricted to participants in the 'White British' cohort with non-missing genotype data ($n = 414{,}964$). We also excluded individuals with a history of cardiovascular disease at baseline. The primary outcome is the first occurrence of a fatal or non-fatal CVD event within 10 years of baseline. Following the broad definition used by Alaa et al. [2], CVD events were identified using specific International Classification of Diseases (ICD) codes: ICD-10 codes for vascular dementia (F01), coronary and ischaemic heart disease (I20–I25), heart failure (I50), and cerebrovascular diseases (I60–I69); and ICD-9 codes for ischaemic heart disease (410–414) and cerebrovascular disease (430–434, 436–438).

3.1 Risk Prediction Models

As baseline model, we used the Framingham risk score (FRS), a well-established sex-specific algorithm used to estimate 10-year cardiovascular risk [17]. We recalibrated the FRS to reflect UK Biobank population-specific risks using separate Cox proportional hazards models for men and women, accounting for gender differences in CVD incidence [15]. For the enhanced model, we combined the FRS with a polygenic risk score (PRS) for cardiovascular disease, available from the UKB PRS Release [16]. Following Riveros-Mckay et al. [13], the enhanced model was calculated as $\hat{f}^{(e)} = \hat{f}^{(FRS)} + \beta \hat{f}^{(PRS)}$, where $\hat{f}^{(FRS)}$ is the Framinham risk score, $\hat{f}^{(PRS)}$ is the polygenic risk score, and β is the parameter estimate obtained from sex-stratified logistic regression performed on a held-out 30% subset of the 'White British' cohort. For both models, we use a threshold of $\delta = 0.1$ to assign the binary intervention, corresponding to the United Kingdom National Institute for Health and Care Excellence (NICE) guidelines recommending statins for individuals with estimated 10-year CVD risk greater than 10% [1].

3.2 Evaluation of policies

We compared a range of threshold-based policies focusing on age and sex as protected characteristics. Specifically, we considered both simple and sex-stratified threshold-based policies with $\delta = 0$ and $\varepsilon = 0, 0.01, 0.05, \infty$. Recall that $\varepsilon = 0$ corresponds to the baseline policy whereby every individual receives the baseline prediction, while if $\varepsilon = \infty$ then each individual receives the enhanced prediction. We evaluated each policy on the remaining set of 'White British' individuals that were not used to fit the enhanced model. The PRS is already available in the UKB dataset for each of these individuals, allowing us to evaluate each of the policies above in the hypothetical scenario in which we only genotype a subset of the individuals in order to obtain their enhanced predictions.

We focus on sensitivity, the proportion of true cases correctly identified, as our primary predictive metric because our goal is to minimize missed cases of disease, particularly in a setting where the intervention is selectively offered due to resource constraints. In such contexts, failing to identify high-risk individuals can have serious clinical consequences, whereas false positives may be less detrimental if they simply lead to further testing or monitoring. For each policy, we compute overall sensitivity, as well as sensitivity for sex-specific (male, female) subgroups and (sex, age)-specific subgroups, splitting age at baseline into three categories: $(40-49)$, $(50-59)$, and $(60-69)$. Differences in sensitivity between subgroups indicates a lack of fairness, specifically in terms of 'true positive rate parity', also referred to as 'equal opportunity fairness'. For each policy, we also report the number of individuals who receive the enhanced model, providing an indication of the overall cost of each policy.

Table 1. Sensitivity across a subset of policies. ε_1 and ε_2 are the tolerances for men and women respectively. n_e is the number of individuals who receive the enhanced model. Left: sex-stratified sensitivity. Right: (age,sex)-stratified sensitivity.

ε_1	ε_2	Sex	Age	Sensitivity	n_e
0	0	Female	All	0.45	0
0	0	Male	All	0.85	0
0.01	0.01	Female	All	0.46	14219
0.01	0.01	Male	All	0.86	10351
0.05	0.05	Female	All	0.49	72775
0.05	0.05	Male	All	0.85	51572
∞	∞	Female	All	0.49	126333
∞	∞	Male	All	0.84	104923
0	0.01	Female	All	0.46	14219
0	0.01	Male	All	0.85	0
0	0.05	Female	All	0.49	72775
0	0.05	Male	All	0.85	0
0	∞	Female	All	0.49	126333
0	∞	Male	All	0.85	0

ε_1	ε_2	Sex	Age	Sensitivity	n_e
0	0	Female	40-49	0.05	0
0	0	Female	50-59	0.23	0
0	0	Female	60-69	0.55	0
0	0	Male	40-49	0.22	0
0	0	Male	50-59	0.72	0
0	0	Male	60-69	0.97	0
0.05	0.05	Female	40-49	0.10	4077
0.05	0.05	Female	50-59	0.31	26773
0.05	0.05	Female	60-69	0.59	41925
0.05	0.05	Male	40-49	0.37	13307
0.05	0.05	Male	50-59	0.76	25194
0.05	0.05	Male	60-69	0.95	13071

3.3 Results

Table 1 reports the sensitivity and the number of people receiving the enhanced model for each policy. Sensitivity was substantially higher for men than for women under the baseline policy ($\varepsilon_1 = \varepsilon_2 = 0$), with values of 0.85 and 0.45 respectively, indicating a lack of true positive rate parity. As the tolerance parameters ε_1 and ε_2 increased, sensitivity improved for women, reaching 0.49 when $\varepsilon_1 = \varepsilon_2 = \infty$, in which all individuals receive the enhanced model, while remaining relatively stable for men, thus resulting in improved fairness. The lack of true positive rate parity was even more stark when comparing across age groups, with younger groups exhibiting much lower sensitivity at baseline. A simple threshold-based policy with $\varepsilon = 0.05$ resulted in higher sensitivity for all but the Male, 60-69 group but did not substantially impact true positive rate parity.

4 Discussion

In this study, we propose a framework for evaluating the fairness of risk prediction policies in resource-constrained settings. This framework can help determine which individuals should receive additional testing to improve fairness. We applied it to an integrated cardiovascular disease risk tool using data from the UK Biobank. Specifically, we examined a threshold-based policy in which individuals whose baseline risk falls within a specified tolerance range are offered the combined risk model (CPM). Future work will consider alternative selection strategies such as top-n ranking policies and stochastic policies that assign the enhanced model probabilistically, potentially reflecting the uncertainty inherent in clinical decision-making. Additionally, fairness could be improved by tailoring thresholds or tolerance levels across demographic groups, though this raises ethical and practical challenges. A comprehensive cost-effectiveness analysis, incorporating health outcome impacts, would further support comparisons of the health and economic value of different policies [9].

Acknowledgments

This work was supported by the Turing–Roche strategic partnership. We are grateful to Chris Harbron, Ben MacArthur, Rohan Chakraborty, Deepak Parashar, Leandra Brauninger, and other members of the partnership for many useful discussions. We also thank three anonymous reviewers for their highly constructive feedback.

References

[1] 2023. NICE Guidelines for Cardiovascular disease: risk assessment and reduction, including lipid modification. https://www.nice.org.uk/guidance/ng238/chapter/Recommendations. Accessed: 2025-05-24.

[2] Ahmed M Alaa, Thomas Bolton, Emanuele Di Angelantonio, James HF Rudd, and Mihaela Van der Schaar. 2019. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PloS one* 14, 5 (2019), e0213653.

[3] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 7726 (Oct. 2018), 203–209. https://doi.org/10.1038/s41586-018-0579-z

[4] Gary S Collins, Paula Dhiman, Jie Ma, Michael M Schlussel, Lucinda Archer, Ben Van Calster, Frank E Harrell, Glen P Martin, Karel G M Moons, Maarten Van Smeden, Matthew Sperrin, Garrett S Bullock, and Richard D Riley. 2024. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ* (Jan. 2024), e074819. https://doi.org/10.1136/bmj-2023-074819

[5] Panagis Galiatsatos, Allen Kachalia, Harolyn M E Belcher, Mark T Hughes, Jeffrey Kahn, Cynda H Rushton, Jose I Suarez, Lee Daugherty Biddison, and Sherita H Golden. 2020. Health equity and distributive justice considerations in critical care resource allocation. *The Lancet Respiratory Medicine* 8, 8 (Aug. 2020), 758–760. https://doi.org/10.1016/S2213-2600(20)30277-0

[6] David C. Goff, Donald M. Lloyd-Jones, Glen Bennett, Sean Coady, Ralph B. D'Agostino, Raymond Gibbons, Philip Greenland, Daniel T. Lackland, Daniel Levy, Christopher J. O'Donnell, Jennifer G. Robinson, J. Sanford Schwartz, Susan T. Shero, Sidney C. Smith, Paul Sorlie, Neil J. Stone, and Peter W. F. Wilson. 2014. 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* 129, 25_suppl_2 (June 2014). https://doi.org/10.1161/01.cir.0000437741.48606.98

[7] Julia Hippisley-Cox, Carol Coupland, and Peter Brindle. 2017. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* (May 2017), j2099. https://doi.org/10.1136/bmj.j2099

[8] Michael Inouye, Gad Abraham, Christopher P. Nelson, Angela M. Wood, Michael J. Sweeting, Frank Dudbridge, Florence Y. Lai, Stephen Kaptoge, Marta Brozynska, Tingting Wang, Shu Ye, Thomas R. Webb, Martin K. Rutter, Ioanna Tzoulaki, Riyaz S. Patel, Ruth J.F. Loos, Bernard Keavney, Harry Hemingway, John Thompson, Hugh Watkins, Panos Deloukas, Emanuele Di Angelantonio, Adam S. Butterworth, John Danesh, and Nilesh J. Samani. 2018. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults. *Journal of the American College of Cardiology* 72, 16 (Oct. 2018), 1883–1893. https://doi.org/10.1016/j.jacc.2018.07.079

[9] Michel Kiflen, Ann Le, Shihong Mao, Ricky Lali, Sukrit Narula, Feng Xie, and Guillaume Paré. 2022. Cost-Effectiveness of Polygenic Risk Scores to Guide Statin Therapy for Cardiovascular Disease Prevention. *Circulation: Genomic and Precision Medicine* 15, 5 (Oct. 2022). https://doi.org/10.1161/CIRCGEN.121.003423

[10] Ziad Obermeyer and Ezekiel J. Emanuel. 2016. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine* 375, 13 (Sept. 2016), 1216–1219. https://doi.org/10.1056/NEJMp1606181

[11] Aniruddh P. Patel, Minxian Wang, Yunfeng Ruan, Satoshi Koyama, Shoa L. Clarke, Xiong Yang, Catherine Tcheandjieu, Saaket Agrawal, Akl C. Fahed, Patrick T. Ellinor, Genes & Health Research Team; the Million Veteran Program, Philip S. Tsao, Yan V. Sun, Kelly Cho, Peter W. F. Wilson, Themistocles L. Assimes, David A. Van Heel, Adam S. Butterworth, Krishna G. Aragam, Pradeep Natarajan, and Amit V. Khera. 2023. A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease. *Nature Medicine* 29, 7 (July 2023), 1793–1803. https://doi.org/10.1038/s41591-023-02429-x

[12] Jessica K. Paulus and David M. Kent. 2020. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *npj Digital Medicine* 3, 1 (July 2020), 99. https://doi.org/10.1038/s41746-020-0304-9

[13] Fernando Riveros-Mckay, Michael E. Weale, Rachel Moore, Saskia Selzam, Eva Krapohl, R. Michael Sivley, William A. Tarran, Peter Sørensen, Alexander S. Lachapelle, Jonathan A. Griffiths, Ayden Saffari, John Deanfield, Chris C.A. Spencer, Julia Hippisley-Cox, David J. Hunter, Jack W. O'Sullivan, Euan A. Ashley, Vincent Plagnol, and Peter Donnelly. 2021. Integrated Polygenic Tool Substantially Enhances Coronary Artery Disease Prediction. *Circulation: Genomic and Precision Medicine* 14, 2 (April 2021), e003304. https://doi.org/10.1161/CIRCGEN.120.003304

[14] Ewout W. Steyerberg. 2019. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-16399-0

[15] Maneesh Sud, Atul Sivaswamy, Anna Chu, Peter C. Austin, Todd J. Anderson, David M.J. Naimark, Michael E. Farkouh, Douglas S. Lee, Idan Roifman, George Thanassoulis, Karen Tu, Jacob A. Udell, Harindra C. Wijeysundera, and Dennis T. Ko. 2022. Population-Based Recalibration of the Framingham Risk Score and Pooled Cohort Equations. *Journal of the American College of Cardiology* 80, 14 (Oct. 2022), 1330–1342. https://doi.org/10.1016/j.jacc.2022.07.026

[16] Deborah J. Thompson, Daniel Wells, Saskia Selzam, Iliana Peneva, Rachel Moore, Kevin Sharp, William A. Tarran, Edward J. Beard, Fernando Riveros-Mckay, Carla Giner-Delgado, Duncan Palmer, Priyanka Seth, James Harrison, Marta Futema, Genomics England Research Consortium, Gil McVean, Vincent Plagnol, Peter Donnelly, and Michael E. Weale. 2024. A systematic evaluation of the performance and properties of the UK Biobank Polygenic Risk Score (PRS) Release. *PLOS ONE* 19, 9 (Sept. 2024), e0307270. https://doi.org/10.1371/journal.pone.0307270

[17] Peter W. F. Wilson, Ralph B. D'Agostino, Daniel Levy, Albert M. Belanger, Halit Silbershatz, and William B. Kannel. 1998. Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation* 97, 18 (May 1998), 1837–1847. https://doi.org/10.1161/01.CIR.97.18.1837