MSGene: Derivation and validation of a multistate model for lifetime risk of coronary artery disease using genetic risk and the electronic health record

Sarah M. Urbut, MD, PhD[1,2,3], Ming Yeung, PhD[4], Shaan Khurshid, MD, MPH[2,5,6], So Mi Jemma Cho[1,2,3,7],Jakob German, MSc[8,9], Akl C. Fahed, MD, MPH[1,2,3], Patrick Ellinor, MD, PhD [2,5,6],Ludovic Trinquart, PhD[10], Giovanni Parmigiani, PhD[11,12], Sasha Gusev, PhD[8,13], Pradeep Natarajan, MD, MMSc[1,2,3]

1. Division of Cardiology, Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA
2. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA
3. Center for Genomic Medicine, Department of Medicine, Massachusetts General Hospital, Boston, MA
4. University of Groningen, Netherlands
5. Demoulas Center for Cardiac Arrhythmias, Massachusetts General Hospital, Boston, MA
6. Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA
7. Integrative Research Center for Cerebrovascular and Cardiovascular Diseases, Yonsei University College of Medicine, Seoul, Republic of Korea
8. Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland
9. Eric and Wendy Schmidt Center, Broad Institute of MIT and Harvard, Cambridge, MA, USA
10. Tufts University Medical Center, Boston, MA
11. Dana Farber Cancer Institute, Boston, MA
12. Harvard School of Public Health, Boston, MA
13. Department of Medicine, Harvard Medical School, Boston, MA

Word Count:
Brief Abstract: 362/350
Main text: 4916/5000
Figures and tables: 8
Supplementary figures: 17

Address for correspondence:
Pradeep Natarajan, MD, MMSc
185 Cambridge Street, CPZN 3.184
Boston, MA 02114
Tel: 617-726-1843 I E-mail: pnatarajan@mgh.harvard.edu I Twitter: @pnatarajanmd

Background: Current cardiovascular risk stratification is based largely on minimizing short-term risk, though earlier identification and modification of risk may have a greater impact. Few lifetime risk estimation tools exist and focus on static, extended window estimates that are poorly calibrated. Whether dynamically updated lifetime risk factor trajectories, combined with genetic risk, can better inform lifetime risk is yet undetermined.

Methods: We analyzed 480,638 UK Biobank (UKB) participants enrolled between 2006-2010, including longitudinal follow-up for health-related outcomes ascertained from longitudinal general practice and hospitalization records spanning 1940 to 2021. We designed a multistate model (MSGene) that employs generalized linear models to estimate age-specific transitions across cardiometabolic states. We applied this framework to assess the remaining lifetime risk of coronary artery disease (CAD) for an individual conditional on sex and externally validated CAD polygenic risk score (PRS), while accounting for smoking and time-dependent medication use, beginning at age 40. We applied the expected relative risk reduction from statin therapy trials to estimate lifelong absolute cardiovascular risk reduction. We compared our dynamic predictions to the Framingham 30-year risk score (FRS30) and Pooled Cohort Equation 10-year score (PCE). Model performance was assessed by root mean squared error (RMSE) and concordance (Harrell's C-index) on multiple time-to-event measures. We externally validate precision and discrimination within the Framingham Offspring Cohort (FOS)

Results: Among 480,638 participants, 260,653 (54.2%) were women. Median age at first healthcare utilization was 24.8 [IQR 19.4] years, and 53,460 (11.1%) sustained incident CAD across a median follow-up [IQR] of 44.3 [30-58] years. Using dynamically updated scores improved discrimination in time-to-event with Harrell's C 0.74 vs. 0.65, $p < 0.01$. Similarly, using MSGene to assess the first age at which a lifetime threshold was exceeded improved Harrell's C-index from 0.53 to 0.70 $p < 0.001$, when compared to using FRS30y. MSGene is better calibrated than FRS30 (RMSE 1.1% vs 10.9% lifetime risk, $p < 2.2e{-}16$); these results hold in FOS. We estimate that absolute risk reduction for CAD using trial-based estimates of statin benefit varies from a mean (95% CI) of 5.9% (4.1-7.8%) at age 40 to 1.1% (0.0-3.2%) at age 79.

Conclusions: Our findings underscore the potential value of accurate lifetime estimation of CAD risk, obtained using a novel framework integrating genetic predisposition and dynamic clinical risk factors over time, to enhance CAD prevention.

Background

Coronary artery disease (CAD), remains the leading cause of morbidity and mortality worldwide.[1] Estimating an individual's risk of developing CAD over the lifetime is essential for timely and effective prevention and intervention[2–5]. Traditional risk prediction models, such as the Pooled Cohort Equations (PCE) 10-year risk score, have guided clinical decisions and preventive strategies; however, these models come with inherent limitations.[6–8] A 30-year or 10-year window provides only a fixed, albeit extended, snapshot of risk. It neither captures the entirety of an individual's lifetime risk nor provides dynamic, age-specific insights beyond these arbitrary periods. Most importantly, there is a growing need for models capable of both recognizing undertreated younger patients while reducing over-estimation in older patients.[10,7,9,10]

While short-term risk assessments have traditionally been utilized to predict near-term events, recent evidence suggests that lifetime risk assessment provides a more comprehensive picture of an individual's propensity for developing CAD across time[11]. Such an approach is especially pertinent given that traditional risk factors when present early in life, and genetic risk can confer a disproportionately elevated risk for CAD in the long term.[2,12–14] Moreover, focusing on lifetime risk allows for more effective patient counseling, tailored preventive measures, and earlier interventions that may delay or prevent the onset of CAD altogether.[15,16] Accordingly, understanding and prioritizing lifetime risk becomes paramount for both clinicians and policymakers aiming to reduce the global burden of CAD.

There is an increasing need for models that can provide continuously updated, dynamic and individualized risk assessments that span a patient's entire life, considering the multifactorial nature of CAD[2,17,18] in combination with the wealth of data present from the electronic health record (EHR). Understanding risk from this perspective allows for more informed and timely interventions, potentially even before the conventional risk windows are applicable.

While the EHR provides detailed longitudinal data, modeling observational data is rife with challenges.[19,20] For example, incorporating laboratory data from observational studies can sometimes lead to challenges in determining the direction of an association. It may be unclear whether changes in the biomarker level cause the outcome or if the outcome itself leads to changes in the biomarker level.[21]

A model inherently capable of incorporating a sequence of exposure and treatment and dynamic changes in risk factors over time is needed to facilitate better lifetime risk estimates. Here, we introduce the MSGene model – a multistate model designed to predict the lifetime risk of CAD, conditional on both time-fixed and time-dependent variables. Multistate models allow for the estimation of the risk of an individual transitioning from one health state to another[22–27] through flexible estimation of conditional probabilities by modeling the transitions between states over time. They naturally account for competing risks. Several previous studies have developed limited multistate models applicable to longitudinal populations[25,26] but rarely in cardiology; the development of these all-encompassing models requires comprehensive ongoing data acquisition that is rarely available from cohort studies.

MSGene is capable of modeling the dynamic transitions from risk factor states to CAD with age-specific coefficients. Multistate models avoid the loss of underlying latent risk when the nominal value of a laboratory test declines.[23] Critically, our approach differs from a traditional Markov-based multistate model[22,23] by extending our model to the time inhomogeneous case and by allowing our transitions to vary with age.[22,24]

In the current study we develop and validate the MSGene model. We evaluate the performance compared to the traditionally employed Framingham 30-year[28] and PCE 10-year[5,6] models. We then estimated the potential ability of MSGene to reduce CAD events by guiding timely initiation of statin therapy and demonstrate the benefit of a multistate framework to incorporate dynamic changes in treatment decisions for unique patient profiles.

Methods

Data Source

The UK Biobank (UKB) is a prospective UK population-based study that enrolled approximately half a million adults aged 40-70 between 2006 and 2010 designed to investigate the genetic and lifestyle determinants for a wide range of diseases. Participants underwent genome-wide genotyping, with linkage to longitudinal hospitalization, primary care, and self-report data dating back to 1940 (**Supp. Fig. 1-4**).[29–31] Notably at the time of analysis, linkage to the United Kingdom General Practice Registry[32] was available for a subset of 221,351 individuals. We removed records with invalid dates such as dates before birth or in the future. Using the ***ukbpheno*** package (version 1.0),[31] we assembled detailed longitudinal data from the various sources (**Supp. Fig. 2-4**) documenting events from 1940 onward for 481,927 individuals

after excluding 20,534 who lacked quality control genotyping or risk factor information **(Fig 1; Supp. Fig. 3).** This assembly across data-sources generated phenotypes[33–35] (**Supp. Fig. 2,3**) for hypertension (Htn), diabetes mellitus (DM) (Type 1 or 2), hyperlipidemia (Hld), or coronary artery disease (CAD) based on validated collections of hospitalization (HESIN), diagnostic, operation, general practice clinical and script as well as death information for England, Wales and Scotland. We cross-referenced these generated phenotypes with our own lab's previously generated HESIN-restricted phenotypes[33,34,36] and found a strong overlap (**Supp. Fig. 2**). Informed consent was obtained from all participants, and secondary data analyses were approved by the Mass General Brigham Institutional Review Board 2021P002228. Secondary data analysis of UKB was performed under application number 7089.

Because of the longitudinal nature of this cohort, every individual was followed from first encounter with the electronic health record in early adulthood (median age 24.3); an individual remains in the 'healthy' category until censored for additional conditions and is considered in the at-risk set of a given condition until censoring for an alternative condition. At each age, we re-evaluate the risk set as those individuals who have 1) been observed and 2) have not been censored for a given phenotype. A binary indicator is created indicating whether they progress to the end state in the following year **(Supp. Fig. 4),** and the logistic regression for each age state-state transition is computed (**Detailed Supplementary Methods**). As UKB was a cohort study of middle-aged adults recruited between ages 40-70, we restricted our modeling to age 40 and beyond. We exclude individuals with a diagnosis of CAD prior to age 40 years and categorize individuals by their condition at 'enrollment' into our cohort: an individual is at risk for progression to CAD over all periods since first observation in the health record (**Supplementary Fig.s 1 and 4**). We demonstrate the diversity of data sources and the corresponding availability of each data source over time for all considered phenotypes (**Supplementary Fig.s 3).**

Polygenic risk

We use CAD polygenic risk score as released through the UKB resource[37] and compute on individuals with adequate genotype information after quality control and after controlling for the principal component axes obtained from the common genotype data in the 1000 Genomes reference dataset using standard methods.[38] Data supporting these scores were entirely from external GWAS data (the Standard PRS set) as conducted by Genomics PLC (Oxford, UK) under UKB project 9659.[37] Briefly, PRS scores were generated using a Bayesian approach applied to meta-analysed summary statistics GWAS data, obtained entirely from external

GWAS data (the Standard PRS set). A subsequent principal component (PC)-based ancestry centering step[39] was applied to approximately center the score distributions on zero across all ancestries. Score distributions were also standardized to have approximately unit variance within ancestry groups, as determined by a geometric inference in PCspace.[37]

Statistical Analysis

Time-dependent risk states and time-fixed covariates

We build a time-dependent multistate model in which age serves as the time scale. We use 80% of our data as training and 20% as testing (**Fig. 1**) for internal cross-validation and to optimize model fit (**Supp. Methods**). Accordingly, this divides our data into a training set for model fitting using 384,510 samples, and a testing data set of 79,117 unique individuals.

For each age and starting state, we model the one-year probability of transition to CAD as a logistic regression conditional on both time-fixed covariates (sex, CAD PRS), and time-dependent covariates (smoking, use of anti-hypertensives or statins) (**Supp. Methods**). We use these state- and age-specific logistic regressions (**Supp. Methods)** with consideration of a limited set of available parameters for optimal fit for comparison with existing equations (**Supp. Table 1**). We smooth each set of state–state coefficients across ages using a flexible tricube distance weighted least square local regression[40,41] according to the inverse variant weighted raw estimate. We compute the lifetime risk as one minus the product of the complement of each age and state-specific transition to CAD probability (Detailed equations and error calcuations in **Supp. Methods**).

Age as the time scale

Our model inferences are made per-year using the individuals who are in the particular risk group at a given age (**Supp. Fig.s 1 and 4**). Predictions can, therefore be made over a requested time interval using the product of age-specific risks for which coefficients were informed only by individuals who were in the at-risk subset during a given period (detailed equations in **Supp. Methods).** While enrollment in the UK Biobank required that an individual be alive at age 40 to enroll for genotyping, it did not require that the individual be event-free. Therefore, we are able to include all individuals based on phenotyped diagnosis at risk bucket for inference from age 40 on.

Competing risks

The unique nature of our multistate model features eight mutually exclusive states and restricts one-year transitions as follows (**Fig. 1),** with death as final absorbing state from which one cannot exit. Transitions to death within the same year of CAD diagnosis are assigned to the death state to avoid duplicate counting. At any age across the life course, cumulative one-step transitions can be assessed (**Fig. 1**). Possible transitions are as follows:

1. Health to a single risk factor (Htn, Hld, Dm), CAD or death;
2. Singe risk factor to corresponding double risk factor, CAD or death;
3. Double risk factor to triple risk factor, CAD or death;
4. Triple risk factor to CAD or death;
5. CAD to death.

Comparison

For comparison of time-dependent 10-year risk, we use the 2018 Pooled Cohort Equations with baseline covariates (total cholesterol, HDL-cholesterol and systolic blood pressure, current smoking) obtained from UKB enrollment data and update each prediction[28] with time-varying age, diabetes, and medication use according to available records. This technique has been used previously in the Framingham 30 year risk development to validate new longer window estimates when in which age was iteratively updated with all other risk factors based on the baseline values[28].

For comparison of 30-year risk, we rely on the 2009 FRS 30-year equation (FRS30y) and update each prediction[28] with time-varying age, diabetes, and anti-hypertensive use according to available records. Given the differing populations, we recalibrated[42] according to the population mean levels of each covariate and baseline hazard in our population. Given that recalibration is not guaranteed to preserve the overall incidence in the population[43], we also performed a sensitivity analysis in which we further standardized to reflect average predictions in line with overall incidence[44] (**Supp. Fig 5)**. We describe corresponding precision and discrimination analysis in the **Supplementary methods.**

Time-dependent model assessment

For the time-dependent analysis, we use the predicted risk scores at each age as covariates in a time-dependent Cox proportional hazard model, in which an individual is featured in non-overlapping intervals with their respective score and event status. For the time-dependent threshold analysis, for each score, we calculated the minimum age at which an

individual would exceed a range of ten and lifetime thresholds. We then divide every individual's observed trajectory into non-overlapping intervals, indicating when one or all thresholds are achieved and when an event occurs (**Supp methods** for detailed descriptions). We fit independent time-dependent Cox models[45,46] to this expanded data set using each score as a predictor and report the concordance index (Harrell's C) with confidence intervals derived from bootstraping iterations.[47,48]

Model assessment

We internally assess the RMSE (**Supp. Table 1)** of models using a finite number of covariates for eight state-specific transitions built on a training set and independently assessed on our testing set using forward selection. External validation was performed by comparing the model fits estimated in the UKB with ten and lifetime risk estimates from 2697 adults in the Framingham Offspring Study[49,50] (**Supp. Fig. 6)** for whom genetic data is available**.** This is a community-based Northeastern United States cohort that was recruited in 1971, median age [min,max] 33.0 [18.0, 62.0] and followed through the present. We compare these with the PCE and FRS30y estimates calculated at exam 1 and compute the RMSE and AUC over the 30-year follow-up period. Given the size of the cohort, we report age-specific AUC for 5-year age intervals.

Results

Baseline characteristics

We considered 480,638 individuals: 54.2% were female with 43,855 (11.1%) incident coronary artery disease diagnoses (**Table 1**) with a median [IQR 30.1-58.1] years of follow-up and median age of first observation of 24.3 [IQR: 18.0, 37.1]. Because of the unique nature of the multistate model, we visualize the proportional representation by risk factor at each age (**Fig. 1**). Approximately 39.6% are ultimately diagnosed with hypertension, 23.6% with hyperlipidemia, and 9.9% with Diabetes Mellitus (1 or 2). Furthermore, 10.5% report current smoking and 20.3% begin antihypertensive use during the course of our study. 46.1% also contributed to the GP cohort, and the distribution of risk factors was homogenous between subsets (**Supp. Fig. 7**).

Modeling transitions

Using our multistate approach, MSGene, we describe the overall state distribution across the lifespan in our cohort, normalizing to exclude censoring at each age (**Fig. 2)**. At age

40 years, 94.4% of individuals are in the healthy category, with 4.1% in the hypertensive category and 0.3% with a diagnosis of CAD. By age 76 years, CAD occupancy peaks at 12.5% of uncensored individuals, and health is reduced to 27.6% of uncensored individuals. By age 80 years, 7.4% have died.

Improved detection of early events when compared to 10-year risk

When compared to the Pooled Cohort Equations (PCE), a 10% lifetime threshold using MSGene uniquely identifies 5315 (59.3%) cases versus 123 (1.3%) cases using the 10-year PCE alone at age 40. This reduces to <1% of cases at age 68 (vs 81% with PCE) (**Fig. 3a).** At age 40, MSGene had substantially greater sensitivity for CAD events compared to PCE (event reclassification 58.2%, 95% CI 58.1-58.3), at the cost of moderate inappropriate up classification of non-events (non-event reclassification -37.3%, 95% CI 37.2-37.4). At age 70, MSGene had substantially greater specificity compared to PCE (non-event reclassification 32.1%, 95% CI 31.9-32.1), at the cost of some inappropriate down classification of events (event reclassification -12.5%, 95% CI -12.4-12.6%). Overall, reclassification was consistently favorable (median NRI 0.12) over 40 years of consideration. Furthermore, 9.7% (95% CI 9.6–9.8%) of individuals in the top 20% of genetic risk are identified to have greater than 10% MSGene predicted lifetime risk, while only 3.1% (95% CI 2.9–3.2%) of those in the bottom 20% of genetic risk achieve this level of risk (**Fig. 3c).**

Dynamic effects of 10-year, 30-year and remaining lifetime risk

We compute the remaining lifetime risk when compared with FRS30y. First, we depict the predicted survival curve for individuals of six different genetic and sex strata starting in health at age 40. Under this traditional analysis, CAD-free survival is projected to decline monotonically as a function of sex and genetic risk to 96.8% (95% CI 96.78–96.82) for a female in the lowest genetic strata and to 81.26% (95% CI 81.24–81.28) for a male in the highest genetic strata. However, a remaining lifetime risk (RLR) curve reveals opposite behavior: for example, a high genetic-risk male has a 22.9% (95% CI 22.7–23.1%) risk without treatment at age 40, but the same high-risk male has only a 10.21% (95% CI 10.20–10.22%) risk of developing CAD if he remains CAD-free at age 70. (2B). This contradicts the 10-year risk prediction, in which 10-year risk rises from 2.84% at age 40 to 10.21% at age 70 (**Fig. 4, Supp. tables 2–17**). We compare this to FRS30y projections[28] and note that while remaining lifetime risk declines with age, the extended fixed-window (FRS30y) approach shows monotonically increasing risk across genetic strata. In our cohort the FRS30-year risk for a high genetic-risk

male rises from 13.4% at age 40 to 33.0% at age 70 using the recalibrated measure and 31.1% to 65.7% using the original FRS equation (**Fig. 5**). When imputing RCT benefit (Equation 2) with our MSGene lifetime projections, predicted absolute risk under statin treatment for the same high-genetic-risk male at age 40 improves from 22.86 (95% CI 22.85–22.87%) to 18.7% (95% CI 18.69–18.71%) over the 40-year span. This is compared to a smaller decline from 10.21% (95% CI 10.19–10.22%) to 8.25% (95% CI 8.24–8.26%) at age 70.

Dynamic prediction: Model assessment

An updated remaining lifetime prediction conditional on the population's current state can be made per year, using age-specific coefficients. We use these updated predictions as covariates in a time-dependent Cox model to evaluate the performance of our model on predicting time to event. We first consider the age distribution at which an individual first exceeded a lifetime risk threshold of 10% using MSGene or FRS30y, or using a PCE-derived 10-year risk threshold of >5%. Using MSGene to assess lifetime risk, 44.8% percent of individuals exceed this threshold at age 40 while 38.9% never do. With FRS30y, 44.1% exceed this threshold at age 40, but virtually all (99.8%) exceed this threshold by age 80. Using the first age exceeded under each model as a time-dependent predictor of CAD status, we find that MSGene improves model concordance by 17% (C-Index 0.70 vs 0.53, $p < 2e{-}16$) and of the 10-year index by 12.2% (C-index 0.578, *p*-value <2e–16) (**Fig. 5).**

We then use the time and state-varying predictions themselves as predictors in a time-dependent Cox proportional hazard model in which one's score is recorded per year in non-overlapping intervals. Again, the concordance of this time-dependent model using dynamic MSGene predictions exceeds that of the updated FRS30y predictions (**Fig. 6**) by 0.74 vs 0.65, *p*-val 1.3×10^{-11}. (**Fig. 6).** We repeat these results using the GP subset alone for both training (80%) and testing (20%) and the results hold for both the thresholding analysis (C-Index 0.71 vs 0.53, p-val < 2e-16) and continuous time-dependent analysis (C Index 0.73 vs 0.67, p-val <2e-16, **Supp Figs 8 and 9**).

Estimated benefit

Our model incorporates the estimated benefit of a treatment strategy to be assessed conditional on starting age, risk status, and relative risk reduction of therapy. Using an imputed RCT annual risk reduction of 20% for statins on statin-free individuals,[51,52] we observe an inverse relationship between predicted 10-year risk and expected benefit is inverse. An individual with the highest genetic risk at age 40 has a predicted 10-year risk (4.2%, SD 0.01)

roughly equivalent to the lowest genetic-risk individual at age 70 (3.9%, SD 0.01), but an expected lifetime absolute risk reduction of 5% (SD 0.01) at age 40 versus only 0.8% (SD $5x10^{-2}$) at age 70 (**Fig. 7**). When we consider the distribution of all starting states, we see that the mean absolute risk reduction is the greatest for younger individuals (4.6-7.2%; SD 0.01) across risk states at age 40, to a mean absolute risk reduction of 0.3–3.5% (SD 0.01) at age 79.

Calibration

We compared the average predicted risk by sex and genomic risk strata with empirical overall incidence rates. FRS30y increases monotonically across the lifespan. In healthy individuals, the RMSE of MSGene is 1.06% (1.04% males, 1.09% females, SEM 0.06) while FRS is 10.91% (12.12% males, 10.12% females, SEM 0.07, **Supp. Fig. 10)**. When restricting the analysis to ages 40 and 50 for whom 30 years of follow-up is available, the RMSE is 0.98% with MSGene when compared to 5.68% for FRS30y. Further recalibration to overall empirical incidence systematically underestimates risk in younger individuals (**Supp. Fig. 5**). We further compute the RMSE starting from additional single-risk factor states (hypertension, hyperlipidemia, and diabetes) across a grid of covariate choices (**Supp. Table 1)**.

Improvement in discrimination over the cumulative horizon

When considering only the presence or absence of disease over observed time without regard to timing, the AUC-ROC of a model comparing the prediction of cumulative occurrence using updated MSGene lifetime score shows greater performance than that of either FRS30y or FRSrecal early in the life course (**Supp. Fig. 11**) (0.69 vs. 0.65, $p < 2e{-}16$ at age 40) and also based on precision-recall (0.20 vs 0.16 at age 40, $p < 0.01$). Both metrics exceeded the estimation of lifetime risk using genetics as a predictor alone. In general, when comparing individuals captured by MSGene but not by FRS, MSGene identified more women and individuals at higher genetic risk. With time, these differences were more profound (**Supp. Fig. 12**).

External validation

To test with an external cohort, we compare to the Framingham Offspring Cohort using first measurements to ensure optimal follow-up duration. We find that the RMSE of MSGene on 30-year prediction when compared to FRS30y is RMSE 8.4% with MSGene vs 11.3 with FRS30y ($p < 2e{-}16$). To assess discrimination, the AUC-ROC with MSGene ranges from 0.75 (95% CI 0.69–0.82) at age 40 to 0.63 (95% CI 0.42–0.84) at age 55, while FRS30y ranges from 0.73 (95% CI 0.66–0.80) at age 40 to 0.53 (95% CI 0.29–0.76) at age 55 (**Supp. Fig. 13).**

Discussion

Current guidelines[9,53,54] recommend the consideration of primordial risk factors in risk-stratifying patients, and call for better methods of estimating lifetime risk. This study introduces a novel method called MSGene, which aims to assess the risk of developing CAD and other health states over the lifespan. The technique utilizes generalized linear models (GLMs) to compute the transition probabilities between different states (e.g., from a healthy state or risk factor to CAD, death, or between) for every age over the observed life span. The novelty derives from four features: 1) the provision of unique age-dependent models via GLMS that allow the relationship of each covariate on the outcome to vary freely with time; 2) the calculation of risk conditional on time-dependent states; 3) the assessment of a multistate model via time-dependent Cox modeling; and 4) the unique use of the UKB EHR as a comprehensive longitudinal data resource. The study follows individuals from adulthood through their enrollment in the linked health record, considering them as part of the "healthy subset" until censored due to other conditions. By incorporating age and time dependence, this method provides annual risk estimates that include the entire lifespan.

Advancements in Early-Life Risk Predictions

Over a lifetime horizon, the dynamic change in risk makes accurate lifetime risk estimations challenging.[2,4,7,53] However, leveraging genetics, MSGene enhances lifetime risk predictions, effectively identifying individuals previously deemed low-risk. The model's age-dependent features, producing age-sensitive coefficients, negate the need to rely on fixed parametric interactions between each covariate and time, a prevalent limitation in traditional models.[6] We show that using updated estimates conditional on the dynamic state of an individual improves *time to event* prediction overall.

Lifetime Estimations and Implications for Preventive Therapies

Critically, through incorporation of treatment effects via imputation of RCT data on transition probabilities, we show that those individuals with the greatest and least absolute risk reduction expected from statin therapy actually have a similar 10-year risk, and yet this short-term focus is what current clinical methods rely upon.[7]

Our approach facilitates accurate event prediction both for undercaptured young individuals and also lower-risk older individuals who might otherwise be included in a fixed-window approach that extends the time horizon: our median global net reclassification when compared with a ten-year approach is 12.2% [IQR 5.5-18.6%] over 40 years. This in part explains the improvement in overall time-dependent performance when incorporated into a time-

to-event framework. Using a time-dependent evaluation, the distribution of the first age at which a lifetime threshold is exceeded demonstrates that MSGene optimally identifies at-risk individuals without indiscriminately calling all patients 'at-risk'. However, future work is warranted to determine optimal thresholds of lifetime risk to maximize potential benefits among high-risk younger individuals while reducing unnecessary costs and harms to low-risk older individuals.

Strengthening Late-Life Risk Predictions

One of the strengths of our method is the access to a significant history of electronic health records that allow us to derive estimates informed by a greater group of patients throughout the life course. Existing scores[8,28,55] imply that the levels of covariates will stay fixed over the life course or require recalculation, which ignores the information within transitions through the life course. Here, our ability to consider the GP records allows for individuals to be followed over a lifetime and quickly estimates what their updated risk trajectory would look like under an alternative profile.

Lifetime-risk estimators, as compared with conventional time-to-event analyses, use age as the time scale, not follow-up time since entry.[56,57] Participants do not enter the risk set at time zero but rather at their age at study entry, in our case first observation in EHR; the times to event of interest or censoring are defined by their age at study exit, so participants contribute risk time from their entry age until their exit age (event, death, or censoring). Consequently, lifetime risk estimates are not limited by the maximum follow-up time but rather by the distribution of entry and exit ages of study participants (Supp methods).

Critically, estimation of remaining lifetime risk is conducted using age-specific predictions informed only by individuals in the at-risk set at a given age, thus making this a true lifetime estimate. In our work, we choose a conservatively estimated age of 80 as the maximum lifetime age given the density of age estimation with our set. This estimation is possible under the assumption that risk trajectory is similar across shifting windows of age at risk but falls apart if you have strong calendar time trends, i.e., if the risk of the event of interest depends on the study entry time. Given that our cohort was required to be between 40 and 70 years old in 2006, we reduced the variation in calendar effects.[5,58]

Utility for Clinical Application

When combined with genetic information, an emphasis on dynamically updated lifetime risk projections can uncover latent risks in seemingly healthy individuals. Determining an appropriate lifetime risk threshold is a laudable goal.[2,7] Indeed, current guidelines[53] note that

genetic risk scores can identify individuals at birth with a high propensity to develop disease, but few approaches have coupled this information with realized risk stages dynamically. As age increases, short-term risk increases, and the remaining lifetime risk is reduced, meaning that a metric focusing on short-term risk will preferentially focus on disease in older individuals, thwarting the efforts of true prevention. It is not enough to increase the lifetime threshold to account for younger individuals as proposed in European Society of Cardiology guidelines[54]; additional years add additional uncertainty, and thus, having tools capable of dynamically incorporating new information over the life course in combination with more comprehensive time assessments is critical to moving prevention forward. We provide an application for individuals to assess risk in real-time for patients and clinicians **(Supp. Fig. 14**).

Limitations

In this study, we use a basket of phenotypic codes to define our risk factor states. One of the challenges of developing a lifetime assessment tool surrounds the availability of continuously updated laboratory data. Using EHR data, and unbiased ascertainment of updated biometric variables at uniform intervals is challenging. We added baseline laboratory data from the age of enrollment to our grid search, and this added little to our model (**Supp. Fig. 15)**.

A second limitation surrounds the heterogeneity of phenotyping. We define hyperlipidemia and hypertension according to a collection of physician-validated diagnostic codes.[31] However, there exists heterogeneity in the severity and duration of these conditions, and adding additional states informing the level of progression may improve these predictions. This potential benefit must be balanced with the uncertainty imposed by additional states and the reduction in sample size caused by dispersion across grades of each condition. Our model resolves the loss in underlying latent risk that is often erroneously captured in EHR data when an individual's nominal laboratory value falls secondary to medication use.

One of the advantages of heterogenous data collection is a wealth of available phenotyping modalities: the UKBB has access through linkages to routinely available national health systems enhanced by self-report and previous records.[59] Although not all individuals included had GP data, we demonstrate that the age and prevalence of conditions is homogenous between individuals in the GP subset and otherwise (**Supp. Fig. 3, 11**) and that analysis on this subset alone results in similar model discrimination.

Third, the generalizability of our findings may be impacted by study design and sample specificity. The UK Biobank included healthier and less socioeconomically deprived individuals

who were predominantly White Europeans living in the United Kingdom[60]. Furthermore, given that the minimum age for genotyping was 40 years old, we began our risk modeling at age 40, regardless of age at enrollment. Although individuals who reached age 40 prior to enrollment were appropriately at risk for the primary CAD outcome given their capture in the longitudinal EHR, they were protected from death until the time of enrollment, which may affect estimates related to the competing risk of death. Nevertheless, we note consistent performance in external validation in the Framingham study, where all death and CAD events occurred exclusively after enrollment. Finally, our dynamic logistic regression framework can readily be adapted to any population with minimal computational resources, and we provide a coding framework to do so.

Conclusions

We introduce multistate model for dynamic transitions through the life course using the UKB extended longitudinal data, and apply to lifetime risk estimation of CAD. The method is well-calibrated and discriminates between early and late events using updated health status to inform our conditional predictions. We provide a novel approach for dynamically estimating lifetime risk, with hopes that clinical cardiology can incorporate these interpretable and dynamic estimates in making future therapeutic decisions that help patients throughout the life course.

Acknowledgments

We would like to acknowledge Leslie Gaffney for her invaluable figure and copyediting advice.

SOURCES OF FUNDING

S.M.U. is supported by T32HG010464 from the National Human Genome Research Institute.
**SK XXX
S.J.C. is supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (grant no.: HI19C1330). A.C.F is supported by grants 1K08HL161448 and R01HL164629 from the National Heart, Lung, and Blood Institute.
PTE reported receiving grants from the NIH (1RO1HL092577, 1R01HL157635, and 5R01HL139731), the American Heart Association Strategically Focused Research Networks (18SFRN34110082), the European Union (MAESTRIA 965286), Bayer AG (to the Broad Institute), IBM Health (to the Broad Institute), Bristol Myers Squibb (to Massachusetts General Hospital), and Pfizer (to Massachusetts General Hospital).
LT ***
ASG ***
GP ***
P.N. is supported by grants R01HL1427, R01HL148565, and R01HL148050 from the National Heart, Lung, and Blood Institute, and grant 1U01HG011719 from the National Human Genome Research Institute.

DISCLOSURES

During the course of the project, M.W.Y. became a full-time employee of GSK.
A.C.F. is co-founder of Goodpath.
PTE reports personal fees from Bayer AG, Novartis, and MyoKardia. P.N. reports research grants from Allelica, Apple, Amgen,Boston Scientific, Genentech / Roche, and Novartis, personal fees from Allelica, Apple, AstraZeneca, Blackstone Life Sciences, Foresite Labs, Genentech / Roche, GV, HeartFlow, Magnet Biomedicine, and Novartis, scientific advisory board membership of Esperion Therapeutics, Preciseli, and TenSixteen Bio, scientific co-founder of TenSixteen Bio, equity in MyOme, Preciseli, and TenSixteen Bio, and spousal employment at Vertex Pharmaceuticals, all unrelated to the present work. The remaining authors have nothing to disclose.

Works Cited

1. Tsao CW, Aday AW, Almarzooq ZI, Anderson CAM, Arora P, Avery CL, Baker-Smith CM, Beaton AZ, Boehme AK, Buxton AE, Commodore-Mensah Y, Elkind MSV, Evenson KR, Eze-Nliam C, Fugar S, Generoso G, Heard DG, Hiremath S, Ho JE, Kalani R, Kazi DS, Ko D, Levine DA, Liu J, Ma J, Magnani JW, Michos ED, Mussolino ME, Navaneethan SD, Parikh NI, Poudel R, Rezk-Hanna M, Roth GA, Shah NS, St-Onge M-P, Thacker EL, Virani SS, Voeks JH, Wang N-Y, Wong ND, Wong SS, Yaffe K, Martin SS, Subcommittee on behalf of the AHAC on E and PSC and SS. Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association. *Circulation* [Internet]. 2023 [cited 2023 May 20];Available from: https://www.ahajournals.org/doi/abs/10.1161/CIR.0000000000001123

2. Lloyd-Jones DM, Leip EP, Larson MG, D'Agostino RB, Beiser A, Wilson PWF, Wolf PA, Levy D. Prediction of Lifetime Risk for Cardiovascular Disease by Risk Factor Burden at 50 Years of Age. *Circulation*. 2006;113:791–798.

3. Wilkins JT, Karmali KN, Huffman MD, Allen NB, Ning H, Berry JD, Garside DB, Dyer A, Lloyd-Jones DM. Data Resource Profile: The Cardiovascular Disease Lifetime Risk Pooling Project. *Int J Epidemiol*. 2015;44:1557–1564.

4. Bundy JD, Ning H, Zhong VW, Paluch AE, Lloyd-Jones DM, Wilkins JT, Allen NB. Cardiovascular Health Score and Lifetime Risk of Cardiovascular Disease. *Circulation: Cardiovascular Quality and Outcomes* [Internet]. 2020 [cited 2023 Jun 13];Available from: https://www.ahajournals.org/doi/abs/10.1161/CIRCOUTCOMES.119.006450

5. Grundy SM, Stone NJ, Bailey AL, Beam C, Birtcher KK, Blumenthal RS, Braun LT, de Ferranti S, Faiella-Tommasino J, Forman DE, Goldberg R, Heidenreich PA, Hlatky MA, Jones DW, Lloyd-Jones D, Lopez-Pajares N, Ndumele CE, Orringer CE, Peralta CA, Saseen JJ, Smith SC, Sperling L, Virani SS, Yeboah J. 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/ APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: Executive Summary. *Circulation*. 2019;139:e1082–e1143.

6. Yadlowsky S, Hayward RA, Sussman JB, McClelland RL, Min Y-I, Basu S. Clinical Implications of Revised Pooled Cohort Equations for Estimating Atherosclerotic Cardiovascular Disease Risk. *Ann Intern Med*. 2018;169:20–29.

7. Navar AM, Fine LJ, Ambrosius WT, Brown A, Douglas PS, Johnson K, Khera AV, Lloyd-Jones D, Michos ED, Mujahid M, Muñoz D, Nasir K, Redmond N, Ridker PM, Robinson J, Schopfer D, Tate DF, Lewis CE. Earlier treatment in adults with high lifetime risk of cardiovascular diseases: What prevention trials are feasible and could change clinical practice? Report of a National Heart, Lung, and Blood Institute (NHLBI) Workshop. *American Journal of Preventive Cardiology*. 2022;12:100430.

8. Jaspers NEM, Blaha MJ, Matsushita K, van der Schouw YT, Wareham NJ, Khaw K-T, Geisel MH, Lehmann N, Erbel R, Jöckel K-H, van der Graaf Y, Verschuren WMM, Boer JMA, Nambi V, Visseren FLJ, Dorresteijn JAN. Prediction of individualized lifetime benefit

from cholesterol lowering, blood pressure lowering, antithrombotic therapy, and smoking cessation in apparently healthy people. *Eur Heart J*. 2020;41:1190–1199.

9. Navar AM, Fonarow GC, Pencina MJ. Time to Revisit Using 10-Year Risk to Guide Statin Therapy. *JAMA Cardiol*. 2022;7:785.

10. Zeitouni M, Nanna MG, Sun J-L, Chiswell K, Peterson ED, Navar AM. Performance of Guideline Recommendations for Prevention of Myocardial Infarction in Young Adults. *Journal of the American College of Cardiology*. 2020;76:653–664.

11. Berry JD, Dyer A, Cai X, Garside DB, Ning H, Thomas A, Greenland P, Van Horn L, Tracy RP, Lloyd-Jones DM. Lifetime Risks of Cardiovascular Disease. *New England Journal of Medicine*. 2012;366:321–329.

12. Michos ED, Choi AD. Coronary Artery Disease in Young Adults. *Journal of the American College of Cardiology*. 2019;74:1879–1882.

13. O'Sullivan JW, Raghavan S, Marquez-Luna C, Luzum JA, Damrauer SM, Ashley EA, O'Donnell CJ, Willer CJ, Natarajan P, on behalf of the American Heart Association Council on Genomic and Precision Medicine; Council on Clinical Cardiology; Council on Arteriosclerosis, Thrombosis and Vascular Biology; Council on Cardiovascular Radiology and Intervention; Council on Lifestyle and Cardiometabolic Health; and Council on Peripheral Vascular Disease. Polygenic Risk Scores for Cardiovascular Disease: A Scientific Statement From the American Heart Association. *Circulation* [Internet]. 2022 [cited 2023 Oct 6];146. Available from: https://www.ahajournals.org/doi/10.1161/CIR.0000000000001077

14. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, Lai FY, Kaptoge S, Brozynska M, Wang T, Ye S, Webb TR, Rutter MK, Tzoulaki I, Patel RS, Loos RJF, Keavney B, Hemingway H, Thompson J, Watkins H, Deloukas P, Di Angelantonio E, Butterworth AS, Danesh J, Samani NJ. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults. *Journal of the American College of Cardiology*. 2018;72:1883–1893.

15. Sniderman AD, Furberg CD. Age as a modifiable risk factor for cardiovascular disease. *Lancet*. 2008;371:1547–1549.

16. Wang N, Woodward M, Huffman MD, Rodgers A. Compounding Benefits of Cholesterol-Lowering Therapy for the Reduction of Major Cardiovascular Events: Systematic Review and Meta-Analysis. *Circulation: Cardiovascular Quality and Outcomes*. 2022;15:e008552.

17. Lloyd-Jones DM, Larson MG, Beiser A, Levy D. Lifetime risk of developing coronary heart disease. *The Lancet*. 1999;353:89–92.

18. Marma AK, Berry JD, Ning H, Persell SD, Lloyd-Jones DM. Distribution of 10-year and lifetime predicted risks for cardiovascular disease in US adults: findings from the National Health and Nutrition Examination Survey 2003 to 2006. *Circ Cardiovasc Qual Outcomes*. 2010;3:8–14.

19. Darke P, Cassidy S, Catt M, Taylor R, Missier P, Bacardit J. Curating a longitudinal research resource using linked primary care EHR data—a UK Biobank case study. *Journal of the American Medical Informatics Association*. 2022;29:546–552.

20. Denaxas S, Shah AD, Mateen BA, Kuan V, Quint JK, Fitzpatrick N, Torralbo A, Fatemifar G, Hemingway H. A semi-supervised approach for rapidly creating clinical biomarker phenotypes in the UK Biobank using different primary care EHR and clinical terminology systems. *JAMIA Open*. 2020;3:545–556.

21. Farmer R, Mathur R, Bhaskaran K, Eastwood SV, Chaturvedi N, Smeeth L. Promises and pitfalls of electronic health record analysis. *Diabetologia*. 2018;61:1241–1248.

22. Le-Rademacher JG, Therneau TM, Ou F-S. The Utility of Multistate Models: A Flexible Framework for Time-to-Event Data. *Curr Epidemiol Rep*. 2022;9:183–189.

23. Wreede LC de, Fiocco M, Putter H. **mstate** : An *R* Package for the Analysis of Competing Risks and Multi-State Models. *J Stat Soft* [Internet]. 2011 [cited 2022 Dec 1];38. Available from: http://www.jstatsoft.org/v38/i07/

24. Brookmeyer R, Abdalla N. Multistate models and lifetime risk estimation: Application to Alzheimer's disease. *Statistics in Medicine*. 2019;38:1558–1565.

25. Neumann JT, Thao LTP, Callander E, Carr PR, Qaderi V, Nelson MR, Reid CM, Woods RL, Orchard SG, Wolfe R, Polekhina G, Williamson JD, Trauer JM, Newman AB, Murray AM, Ernst ME, Tonkin AM, McNeil JJ. A multistate model of health transitions in older people: a secondary analysis of ASPREE clinical trial data. *The Lancet Healthy Longevity*. 2022;3:e89–e97.

26. Jack CR, Therneau TM, Lundt ES, Wiste HJ, Mielke MM, Knopman DS, Graff-Radford J, Lowe VJ, Vemuri P, Schwarz CG, Senjem ML, Gunter JL, Petersen RC. Long-term associations between amyloid positron emission tomography, sex, apolipoprotein E and incident dementia and mortality among individuals without dementia: hazard ratios and absolute risk. *Brain Communications*. 2022;4:fcac017.

27. Jack CR, Therneau TM, Wiste HJ, Weigand SD, Knopman DS, Lowe VJ, Mielke MM, Vemuri P, Roberts RO, Machulda MM, Senjem ML, Gunter JL, Rocca WA, Petersen RC. Rates of transition between amyloid and neurodegeneration biomarker states and to dementia among non-demented individuals: a population-based cohort study. *Lancet Neurol*. 2016;15:56–64.

28. Pencina MJ, Ralph B. D'Agostino S, Larson MG, Massaro JM, Vasan RS. Predicting the 30-Year Risk of Cardiovascular Disease. *Circulation* [Internet]. 2009 [cited 2023 Sep 20];Available from: https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.108.816694

29. : External Info : Data_providers_and_dates [Internet]. [cited 2023 Sep 25];Available from: https://biobank.ctsu.ox.ac.uk/crystal/exinfo.cgi?src=Data_providers_and_dates

30. : External Info : HESupdate_2019_09 [Internet]. [cited 2023 Sep 25];Available from: https://biobank.ndph.ox.ac.uk/ukb/exinfo.cgi?src=HESupdate_2019_09

31. Yeung MW, Van Der Harst P, Verweij N. ukbpheno v1.0: An R package for phenotyping health-related outcomes in the UK Biobank. *STAR Protocols*. 2022;3:101471.

32.

33. Patel AP, Wang M, Ruan Y, Koyama S, Clarke SL, Yang X, Tcheandjieu C, Agrawal S, Fahed AC, Ellinor PT, Tsao PS, Sun YV, Cho K, Wilson PWF, Assimes TL, van Heel DA, Butterworth AS, Aragam KG, Natarajan P, Khera AV. A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease. *Nat Med*. 2023;29:1793–1803.

34. Klarin D, Zhu QM, Emdin CA, Chaffin M, Horner S, McMillan BJ, Leed A, Weale ME, Spencer CCA, Aguet F, Segrè AV, Ardlie KG, Khera AV, Kaushik VK, Natarajan P, CARDIoGRAMplusC4D Consortium, Kathiresan S. Genetic analysis in UK Biobank links insulin resistance and transendothelial migration pathways to coronary artery disease. *Nat Genet*. 2017;49:1392–1397.

35. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, Kathiresan S. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018;50:1219–1224.

36. Natarajan P, Young R, Stitziel NO, Padmanabhan S, Baber U, Mehran R, Sartori S, Fuster V, Reilly DF, Butterworth A, Rader DJ, Ford I, Sattar N, Kathiresan S. Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention Setting. *Circulation*. 2017;135:2091–2101.

37. Thompson DJ, Wells D, Selzam S, Peneva I, Moore R, Sharp K, Tarran WA, Beard EJ, Riveros-Mckay F, Palmer D, Seth P, Harrison J, Futema M, Consortium GER, McVean G, Plagnol V, Donnelly P, Weale ME. UK Biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits [Internet]. 2022 [cited 2023 Oct 2];2022.06.16.22276246. Available from: https://www.medrxiv.org/content/10.1101/2022.06.16.22276246v1

38. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38:904–909.

39. Khera AV, Chaffin M, Zekavat SM, Collins RL, Roselli C, Natarajan P, Lichtman JH, D'Onofrio G, Mattera J, Dreyer R, Spertus JA, Taylor KD, Psaty BM, Rich SS, Post W, Gupta N, Gabriel S, Lander E, Ida Chen Y-D, Talkowski ME, Rotter JI, Krumholz HM, Kathiresan S. Whole-Genome Sequencing to Characterize Monogenic and Polygenic Contributions in Patients Hospitalized With Early-Onset Myocardial Infarction. *Circulation*. 2019;139:1593–1602.

40. Cleveland WS. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*. 1979;74:829–836.

41. Cleveland WS, Devlin SJ. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting.

42. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P, Marchini J. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562:203–209.

43. Wessler BS, Ruthazer R, Udelson JE, Gheorghiade M, Zannad F, Maggioni A, Konstam MA, Kent DM. Regional Validation and Recalibration of Clinical Predictive Models for Patients With Acute Heart Failure. *J Am Heart Assoc*. 2017;6:e006121.

44. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PRO, Bernstam EV, Lehmann HP, Hripcsak G, Hartzog TH, Cimino JJ, Saltz JH. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care*. 2013;51:S30-37.

45. Therneau T, Crowson C, Atkinson E. Using Time Dependent Covariates and Time Dependent Coe cients in the Cox Model. :31.

46. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model [Internet]. New York, NY: Springer New York; 2000 [cited 2022 Aug 8]. Available from: http://link.springer.com/10.1007/978-1-4757-3294-8

47. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the Yield of Medical Tests. *JAMA*. 1982;247:2543–2546.

48. Schmid M, Wright M, Ziegler A. On the use of Harrell's C for clinical risk prediction via random survival forests [Internet]. 2016 [cited 2023 Oct 8];Available from: http://arxiv.org/abs/1507.03092

49. Feinleib M, Kannel WB, Garrison RJ, McNamara PM, Castelli WP. The Framingham Offspring Study. Design and preliminary data. *Prev Med*. 1975;4:518–525.

50. KANNEL WB, FEINLEIB M, McNAMARA PM, GARRISON RJ, CASTELLI WP. AN INVESTIGATION OF CORONARY HEART DISEASE IN FAMILIES: THE FRAMINGHAM OFFSPRING STUDY. *American Journal of Epidemiology*. 1979;110:281–290.

51. Cholesterol Treatment Trialists' (CTT) Collaborators, Mihaylova B, Emberson J, Blackwell L, Keech A, Simes J, Barnes EH, Voysey M, Gray A, Collins R, Baigent C. The effects of lowering LDL cholesterol with statin therapy in people at low risk of vascular disease: meta-analysis of individual data from 27 randomised trials. *Lancet*. 2012;380:581–590.

52. Chou R, Cantor A, Dana T, Wagner J, Ahmed AY, Fu R, Ferencik M. Statin Use for the Primary Prevention of Cardiovascular Disease in Adults: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. *JAMA*. 2022;328:754.

53. Lloyd-Jones DM, Albert MA, Elkind M. The American Heart Association's Focus on Primordial Prevention. *Circulation*. 2021;144:e233–e235.

54. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice | European Heart Journal | Oxford Academic [Internet]. [cited 2023 Oct 6];Available from: https://academic.oup.com/eurheartj/article/42/34/3227/6358713

55. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, Brindle P. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ*. 2008;336:1475–1482.

56. Vasan RS, Enserro DM, Beiser AS, Xanthakis V. Lifetime Risk of Heart Failure Among Participants in the Framingham Study. *J Am Coll Cardiol*. 2022;79:250–263.

57. Vasan RS, Beiser A, Seshadri S, Larson MG, Kannel WB, D'Agostino RB, Levy D. Residual lifetime risk for developing hypertension in middle-aged women and men: The Framingham Heart Study. *JAMA*. 2002;287:1003–1010.

58. Arnett DK, Blumenthal RS, Albert MA, Buroker AB, Goldberger ZD, Hahn EJ, Himmelfarb CD, Khera A, Lloyd -Jones Donald, McEvoy JW, Michos ED, Miedema MD, Mu ñoz D, Smith SC, Virani SS, Williams KA, Yeboah J, Ziaeian B. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease. *Journal of the American College of Cardiology*. 2019;74:e177–e232.

59. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med*. 2015;12:e1001779.

60. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, Collins R, Allen NE. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American Journal of Epidemiology*. 2017;186:1026–1034.

Figure Legends

Figure 1. Study Overview
Using the UK Biobank data on half a million participants (54% female) with access to health record from 1940, we harmonize hospitalization, prescription and primary care records from the electronic health record and train our model on individuals free of CAD at age 40. The UKB required participants to be between ages 40-69 between 2006-2010 for genotyping. In our model, individuals join disease-free in the 'health' state and progress to additional states upon censoring. We use 80% of the eligible data for training and the remaining 20% for testing. For the testing subset we require that individuals have variables necessary for computation of FRS30 year and the pooled cohort equations, which require laboratory (HDL, TC) and biometric (SBP) measurements.

TC: Total Cholesterol, **SBP**: systolic blood pressure, **HDL**: high density lipoprotein. **CAD**: coronary artery disease. **FRS30y**: Framingham 30 year, **PCE:** Pooled cohort equation 10 year risk.

Figure 2. Multistate transitions over time
A. We depict the potential one-step transitions in our multistate framework. Per year, an individual can progress from health to single risk factor states, CAD or death. Similarly, an individual can progress from single risk factor states, to double risk factor states, to CAD or death; from double risk factor states, to triple risk factor, CAD or death**. B.** We display the proportional occupancy excluding censored individuals at each state.

CAD: Coronary Artery Disease, **Ht**: Hypertension, **HyperLip**: Hyperlipidemia, **Dm**: Type 2 Diabetes Mellitus.

Figure 3. Comparison to ten-year pooled cohort equations.
A. We display the proportion of cases captured using a pooled-cohort equation (PCE) threshold of 5%, a lifetime threshold of 10% as computed by MSGene, or both. At age 40, 58% of individuals who ultimately develop CAD demonstrate an MSGene lifetime threshold greater than 10% while less than 1.3% demonstrate a PCE 10-year threshold than 5% alone. B. The net proportion of events (NRI case) detected by a lifetime score exceeds that of a ten year score at age 40 and the net proportion of non-events exceeds that of a ten year measure after age 60. Median NRI over the 40 year period is 12.2% (5.4%-18.6%) C. High lifetime risk individuals not captured by ten year equation is enrichemed in high-genomic risk individuals. After age 68, there are no individuals with lifetime score over 10% who lack a short term risk greater than 5%.

PCE: pooled cohort equations. **PRS**: Polygenic Risk Score

Figure 4. Survival, Ten year and Lifetime Risk Curves
In **A.**, we demonstrate the singular projected survival curve by MSGene for an individual at age 40 of low, medium or high genomic risk. In **B.** we demonstrate the MSGene predicted ten-year risk for individuals at each age along the x-axis, showing that in general, for fixed window approaches, ten-year risk is monotonically increasing. In **C**, we demonstrate the MSGene predicted lifetime risk curve for individuals at each age featured along the x-axis under an untreated (dashed) or treated (solid) strategy. The conditional remaining lifetime risk declines with age, from 24% for a high genomic risk individual in our cohort to <5% for an individual at the same risk level by age 70. In **D**, using the FRS30y recalibrated equation, like 10-year risk

and unlike the remaining lifetime risk approach, 30 year risk calculation is monotonically increasing, from 13.4 (13.2-13.6%) at age 40 to 32.9% at age 70 for an individual of the highest genomic risk.

FRS30: Framingham 30 year recalibrated.

Figure 5: Time-dependent threshold analysis
We consider the distribution of the first age at which an individual exceeds the PCE-derived ten year threshold of 5%, or lifetime threshold or 10% using FRS 30 recalibrated (B) or the MSGene lifetime prediction (C). We then use this age as a time-dependent predictor of time to event in a time-dependent Cox PH (**Supp Methods)** in which an individual's time followed is stratified by start time and periods in which a threshold is passed, and final censoring time with an indicator variable demarcating whether or not each threshold has been surpassed. We report Harrell's C-index (p-val<2e-16) for discrimination on how well a model predicts events that tend to occur earlier versus later. Light blue arrow heads indicate individuals who surpass the threshold at first prediction, and open red arrow head indicates individuals who never surpass a threshold for a given metric. Framingham recalibrated score is shown here with C-index 0.53 (original with C index 0.50)

FRS30: Framingham 30 year recalibrated. **PCE**: Pooled Cohort equations. Cox PH: Cox Proportional Hazards moel

Figure 6: Improved time-to-event prediction using time-dependent evaluation
(A) We compute the lifetime prediction at each age under one of 8 potential risk starting states, with bootstrapped confidence intervals for a sample individual. **(B)** Using the electronic health record, we extract state position for each individual per year. We then use MSGene to compute predicted risk for each individual at each state in time, displayed here for a sample individuals. **(C)** We use these as predictors in a time-dependent cox model in which we expand the data set into non overlapping intervals for each individual (**Supp Methods; Supp Figure 17)** and evaluate the concordance when compared to FRS30y RC and PCE-derived ten year., p-val <2.2e-16.

FRS30y: Framingham 30year Recalibrated. **PCE**: Pooled Cohort equations.

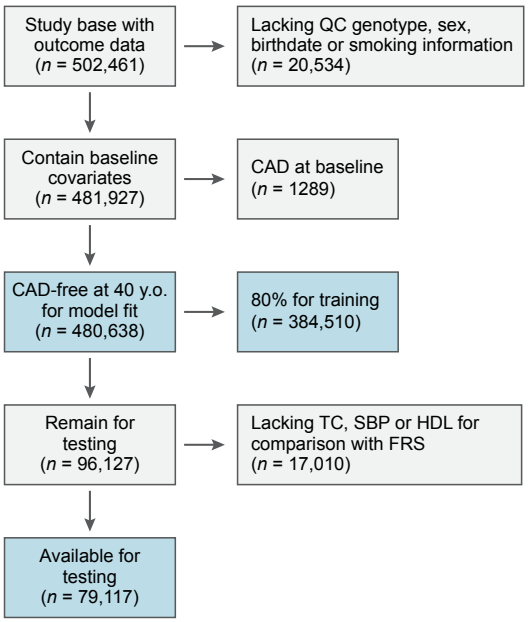
Figure 7: Absolute Risk Reduction: Short Term and Lifetime Risk
We display the relationship between remaining lifetime and ten-year risk. Each ray represents an age group, in which individuals are parameterized by their short (10-year) and long-term (lifetime) risk, and colored by genomic risk in SD from mean. We display the lifetime absolute risk reduction as computed in Equation RR and stratified by age rays, and colored by genetic risk. (**A**) For an individual at the top genetic risk at age 40, ten-year risk is roughly equivalent to an individual at the lowest genetic risk at age 70 (3.8% vs 4.2%, SE 0.01). However, the projected lifetime benefit is directly proportional to lifetime risk **(B)**, and more than twice that of a high risk individual at age 70 (5.0 vs 2.3%, SEM 0.02). (**C**) Marginalized across starting states and covariate profiles, we project mean expected difference in risk. At age 40, this ranges from a median of 5.8% (SD 0.01) to 0.8% (SD 0.01) at age 79.

SEM: Standard Error of Mean, RR: Relative Risk, SD: CAD-PRS SD

	Low Genomic Risk (N=115288)	Intermediate Genomic Risk (N=294978)	High Genomic RIsk (N=70372)	Overall (N=480638)
Sex				
Female	62304 (54.0%)	160122 (54.3%)	38227 (54.3%)	260653 (54.2%)
Male	52984 (46.0%)	134856 (45.7%)	32145 (45.7%)	219985 (45.8%)
Age first observed				
Mean (SD)	29.2 (13.2)	29.2 (13.2)	29.1 (13.2)	29.2 (13.2)
Median [Min, Max]	24.5 [18.0, 78.6]	24.3 [18.0, 79.1]	24.2 [18.0, 78.0]	24.3 [18.0, 79.1]
Years Followed				
Median (IQR)	44.4 (30.3-58.3)	44.4 ((30.3-58.4)	44.2 (30.1-58.3)	44.4 (30.3-58.3)
Birthdate				
Mean (SD)	1950 (8.12)	1950 (8.11)	1950 (8.10)	1950 (8.11)
Median [Min, Max]	1950 [1940, 1970]	1950 [1930, 1970]	1950 [1940, 1970]	1950 [1930, 1970]
Develop Hypertension				
0	75780 (65.7%)	176772 (59.9%)	37634 (53.5%)	290186 (60.4%)
1	39508 (34.3%)	118206 (40.1%)	32738 (46.5%)	190452 (39.6%)
Develop Coronary Disease				
0	107466 (93.2%)	262513 (89.0%)	57199 (81.3%)	427178 (88.9%)
1	7822 (6.8%)	32465 (11.0%)	13173 (18.7%)	53460 (11.1%)
Develop Hyperlipidemia				
0	94296 (81.8%)	224469 (76.1%)	48279 (68.6%)	367044 (76.4%)
1	20992 (18.2%)	70509 (23.9%)	22093 (31.4%)	113594 (23.6%)
Develop Diabetes (Type 1 or 2)				
0	105679 (91.5%)	266049 (90.0%)	62042 (87.6%)	433770 (90.1%)
1	9763 (8.4%)	29633 (10.0%)	8761 (12.4%)	48157 (9.9%)
Start an anti-Hypertensive				
Mean (SD)	0.180 (0.384)	0.204 (0.403)	0.232 (0.422)	0.203 (0.402)
Current Smoker				
Mean (SD)	0.101 (0.301)	0.106 (0.308)	0.110 (0.312)	0.105 (0.307)
General Practice Primary Care Data				
Proportion (SD)	0.455 (0.498)	0.462 (0.499)	0.466 (0.499)	0.461 (0.498)
Proportion White				
Mean (SD)	0.832 (0.373)	0.844 (0.363)	0.828 (0.377)	0.839 (0.368)

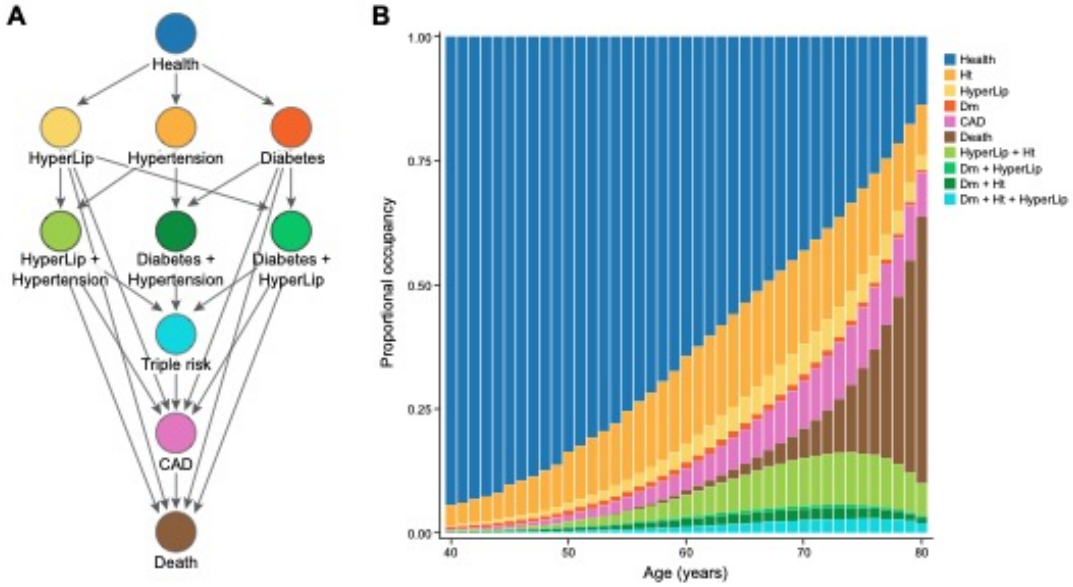
Table 1. Distribution of Overall Cohort. We use approximately 80% (385541) individuals in the training, and 79,119 in the testing set, of which approximately 45% represent members of the general practice primary care data. Of note, low genomic risk connotes individuals in the lowest (<20%) of genomic risk by PRS percentile, intermediate (20-80%) PRS percentile, and high denotes >80% PRS percentile.

Figure 1


Study base with outcome data (n = 502,461)
Lacking QC genotype, sex, birthdate or smoking information (n = 20,534)
Contain baseline covariates (n = 481,927)
CAD at baseline (n = 1289)
CAD-free at 40 y.o. for model fit (n = 480,638)
80% for training (n = 384,510)
Remain for testing (n = 96,127)
Lacking TC, SBP or HDL for comparison with FRS (n = 17,010)
Available for testing (n = 79,117)

1
2

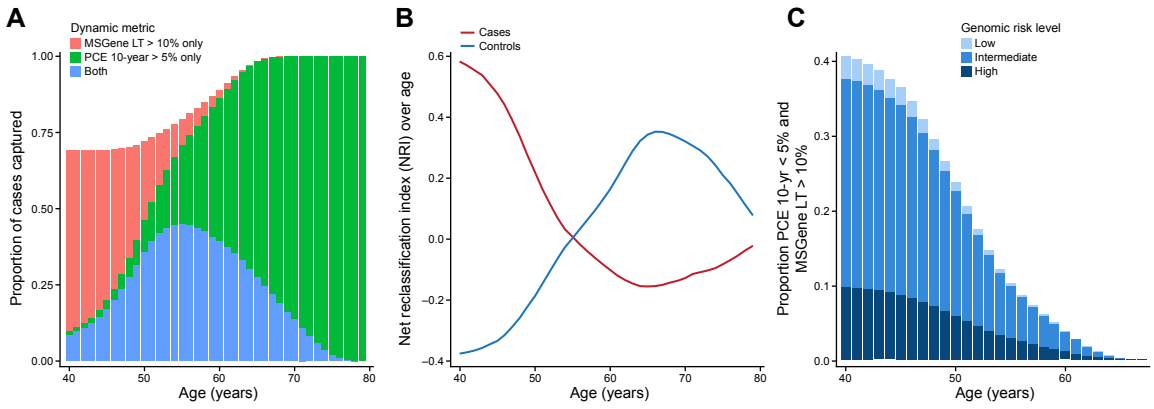
Figure 2



1
2
3
4

1
2
3

Figure 3



4

Figure 4

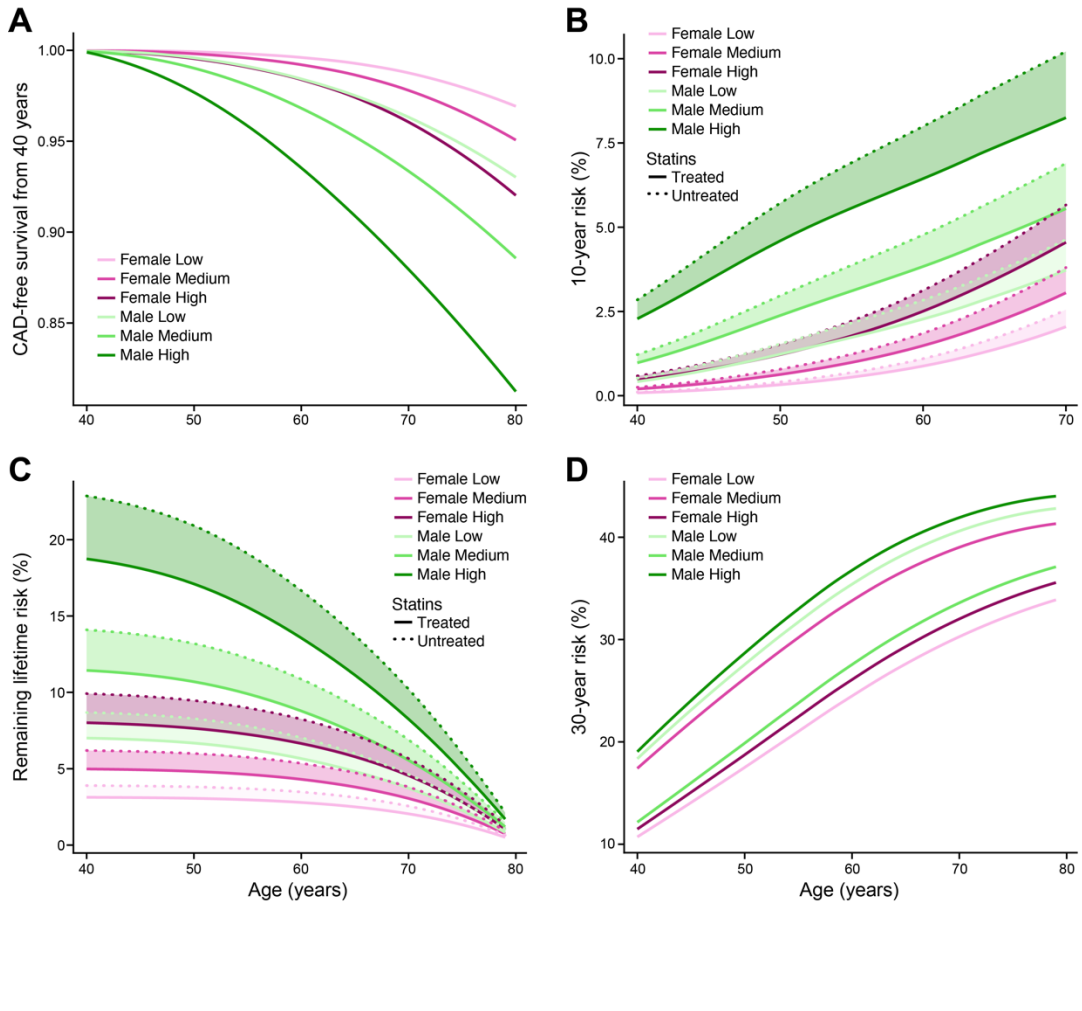
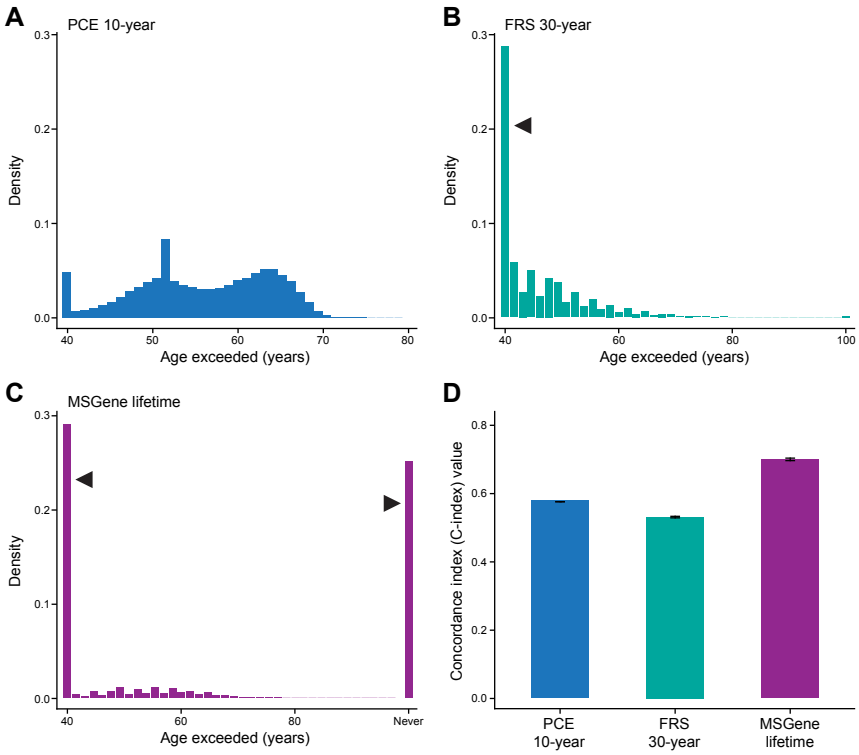
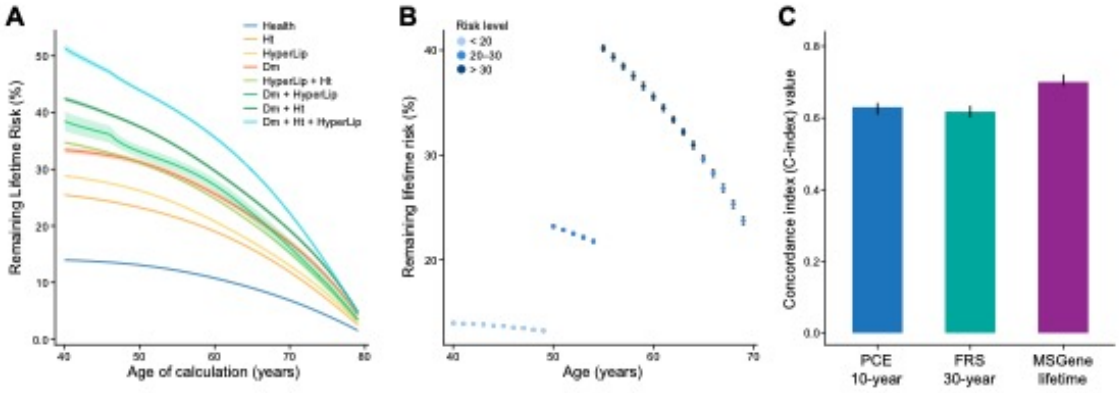

A
CAD-free survival from 40 years
1.00
0.95
0.90
0.85
40
50
60
70
80
Female Low
Female Medium
Female High
Male Low
Male Medium
Male High
B
10-year risk (%)
10.0
7.5
5.0
2.5
0.0
40
50
60
70
Female Low
Female Medium
Female High
Male Low
Male Medium
Male High
Statins
Treated
Untreated
C
Remaining lifetime risk (%)
20
15
10
5
0
40
50
60
70
80
Age (years)
Female Low
Female Medium
Female High
Male Low
Male Medium
Male High
Statins
Treated
Untreated
D
30-year risk (%)
40
30
20
10
40
50
60
70
80
Age (years)
Female Low
Female Medium
Female High
Male Low
Male Medium
Male High

Figure 5

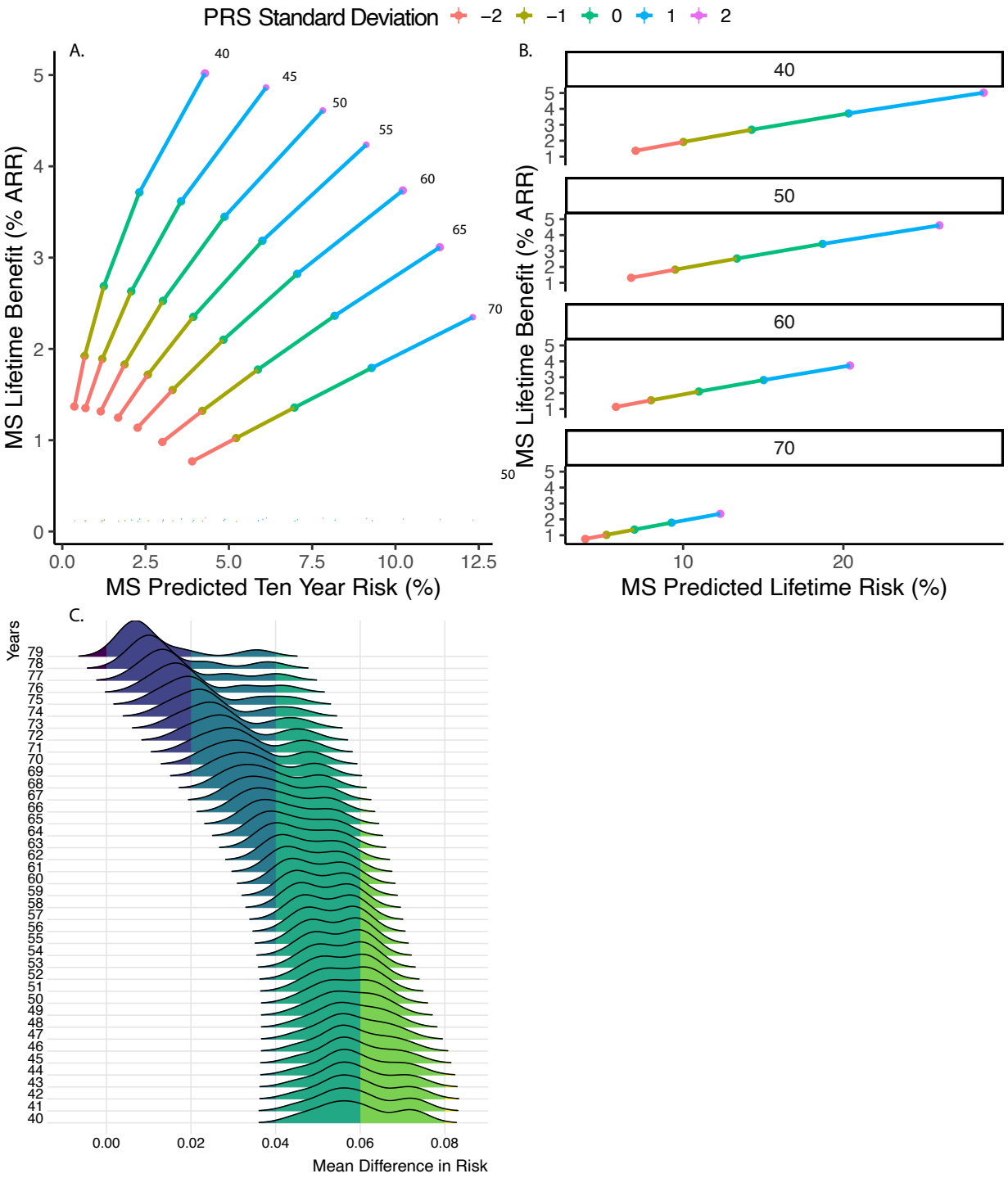

A
PCE 10-year
Density
Age exceeded (years)
B
FRS 30-year
Density
Age exceeded (years)
C
MSGene lifetime
Density
Age exceeded (years)
Never
D
Concordance index (C-index) value
PCE
10-year
FRS
30-year
MSGene
lifetime

1

Figure 6


A
Remaining Lifetime Risk (%)
Age of calculation (years)
Health
Ht
HyperLip
Dm
HyperLip + Ht
Dm + HyperLip
Dm + Ht
Dm + Ht + HyperLip
B
Remaining lifetime risk (%)
Age (years)
Risk level
< 20
20–30
> 30
C
Concordance index (C-index) value
PCE
10-year
FRS
30-year
MSGene
lifetime

1


PRS Standard Deviation
−2
−1
0
1
2
A.
MS Lifetime Benefit (% ARR)
MS Predicted Ten Year Risk (%)
40
45
50
55
60
65
70
B.
MS Lifetime Benefit (% ARR)
MS Predicted Lifetime Risk (%)
C.
Years
Mean Difference in Risk