Derivation and validation of Health Insights polygenic risk scores and integrated risk tools

Melisa Chuong, Iliana Peneva, Sophie Landon, Daniel Wells,
Robert Bettles, Mike Weale, Seamus Harrison, Peter Donnelly,
Gil McVean, Vincent Plagnol, Rachel Moore

November 2023

Introduction

Health Insights is a product that provides adults living in the US, with their predicted risk of developing selected common diseases over the next 10 years, or by the age of 80 as appropriate. This risk is based on an individual's self-reported current age, self-reported race or ethnicity (SIRE), self-reported biological sex at birth, and the relevant polygenic risk score (PRS), which is derived from genetic data obtained via an at-home, saliva-based, self-use DNA sampling kit. Polygenic risk scores (PRSs) provide a personalised measure of an individual's disease risk by combining genetic risk information from across the genome [1]. They have been shown to be useful tools for risk stratification [2, 3, 4, 5, 6]. In addition, they can have similar or greater predictive power compared to established clinical risk factors, such as LDL cholesterol for coronary artery disease [7] and can identify individuals at equivalent or higher risk than carriers of clinically relevant rare variants [8, 9].

Within Health Insights, individuals receive disease risk estimates for up to seven diseases: atrial fibrillation (AF), cardiovascular disease (CVD), hypercholesterolemia (HCH; high low-density lipoprotein cholesterol), hypertension (HT; high blood pressure), osteopenia (low bone density), type 2 diabetes (T2D) and one of breast (BrC) or prostate cancer (PrC). In addition to receiving the risk estimates, a digital service can be used by individuals to support lifestyle actions to mitigate the risk of diseases. A laboratory report is also provided for use by health professionals, who can use the risk estimates to inform their recommendations to an individual regarding engagement with state of the art disease prevention strategies and screening programs. This product builds on a successful pilot study[1], which found that understanding personal risk of disease supports engagement with healthy behaviours, with 69.9% of all participants planning to take, or having already taken, action within four months of participating in the study, rising to 78.4% of individuals amongst those classified as high risk for at least one disease (from a possible three diseases within the pilot).

A PRS measures the relative disease risk for an individual from inherited, common genetic variants. This becomes more powerful when combined with other relevant risk factors, in an integrated risk tool (IRT), to generate estimates of absolute risk: the probability of being diagnosed with a disease within a specified time period [7, 10]. The use of absolute risk scores is already established in several clinical contexts, such as the pooled cohort equation risk score, which measures risk for cardiovascular disease, used in the context

[1]The pilot study returned a genetic risk report to 1044 individuals. Up to 3 disease risks were returned per participant: 10 year risk of cardiovascular disease, 10 year risk of type 2 diabetes, lifetime risk of breast cancer (women), and lifetime risk of prostate cancer (men). Participant satisfaction and feedback was assessed via an initial survey (Questionnaire A, 489 respondents), a post genetic counselling survey (Questionnaire B, 18 respondents), and a 4-month survey (Questionnaire C, 532 respondents).

© 2023 Genomics plc — All rights reserved

of shared decision making about preventative medication with statins and other cholesterol-lowering drugs [11, 12].

In this paper, we describe how the PRSs and IRTs used within Health Insights are generated and validated. Health Insights combines optimised, proprietary PRS, which are well-powered across all ancestry groups, with estimates of disease incidence that are up-to-date and relevant for the specific country, sex, age, and race or ethnicity of an individual (as defined by descriptors included within the US census, augmented for individuals of East Asian, South Asian and South East Asian race or ethnicity). We show that the Health Insights IRTs are well calibrated, meaning, for example, that 30% of individuals identified as having a risk of 30% do go on to have a diagnosis of the relevant disease in the specified time period. Such strong calibration is essential for providing users and health-care professionals with the confidence to act on the information provided. We also show that inclusion of the PRS improves predictive performance compared to the baseline model (which does not include the PRS), across multiple ethnicities for which there is data, within well-powered US and UK validation cohorts. These results establish the added value of using PRS to inform about disease risks.

We apply this approach across eight diseases included in the Health Insights product. These diseases were selected on the basis that they are common within the population (average lifetime risk of at least 10%), across sexes (excluding sex-specific diseases) and racial/ethnic groups, costly to society, and can be mitigated by lifestyle changes and/or accessing readily-available medical pathways and screening programs. It is expected that 80% of individuals using Health Insights will be informed of being at increased risk for at least one disease.

Methods

Genetic ancestry and self-identified race or ethnicity

Previous work has established the importance of, and distinction between, genetically inferred ancestry (the fraction of someone's genome that has descended from ancestors who lived in a particular region of the world at a particular point in time) and self-identified race or ethnicity (a social identity relating to shared background, behaviours, beliefs and practices) in calculating IRTs [13]. Within Health Insights, both aspects of an individual's identity are used in the construction of risk estimates.

For genetic ancestry, individuals are assigned to one of five high-level ancestry groups based on the principle genetic similarity between their genome and a set of reference genomes: European (EUR), East Asian (EAS), South Asian (SAS), Sub-Saharan African (AFR), and Native/Indigenous American (AMR) [6, 13, 14]. These groups are used to estimate the predictive weight assigned to the PRS in each ancestry group (technically measured as odds ratio associated with each standard deviation of the PRS; see methods section below for more details).

Self-identified race or ethnicity (SIRE) as reported by individuals, along with self-reported age, sex assigned at birth and country-of-origin is used to define context-specific baselines for disease risk. For the US, racial/ethnicity categories are based on the US census, and accordingly we use White Hispanic, White non-Hispanic, White prefer-not-to-say (PNTS), Black Hispanic, Black non-Hispanic, Black prefer-not-to-say (PNTS), South Asian, East Asian, South East Asian, American Indian or Alaskan Native (AIAN), Native Hawaiian or other Pacific Islander (NHPI) and Hispanic SIREs. For the UK (when required for validation purposes), we use White, Black African, Black Caribbean, Black Other, South Asian and East Asian SIREs. If the individual does not provide their SIRE and/or sex, SIRE-

© 2023 Genomics plc — All rights reserved

averaged and/or sex-averaged baselines are used.

Polygenic risk scores

PRS are derived by meta-analysing [15] disease-appropriate genome-wide association study (GWAS) summary statistics, ensuring that there is no overlap between samples used for PRS training, calculating PRS effect sizes within an IRT, baseline derivation and PRS or IRT validation. We use proprietary methods [6] to calculate PRS, including a computational approach related to the published LDpred method [16]. Our methodology estimates the genome-wide contribution to disease risk, and spreads this risk across a large number of genetic variants (between two and five million variants, depending on the trait). PRS are computed for each individual using the resulting PRS weights, then centered and standardised to ensure consistent mean and standard deviation across the major ancestry groups described above [2].

PRS computed using Genomics' methods for 25 diseases and 28 quantitative traits have been previously released in the UK Biobank resource [6]. Detailed assessment of these scores shows them to outperform most published PRS algorithms [6]. This improved performance of the Genomics methods is a consequence both of the use of Genomics' data resources (the internal data platform, which integrates summary statistics from tens of thousands of GWAS studies, augmented with summary statistics from collaborators), and of its improved methodology for deriving PRS from GWAS summary statistics.

Assessment of PRS performance

We assess the ability of PRS to distinguish between people with and without disease, or with different values of a quantitative trait, using the Area Under the receiver operator channel Curve (AUC) [17] for binary traits and the proportion of trait variance explained (R^2) for quantitative traits, considering models where we regress the PRS against the disease status in the absence of any covariates as a measure of standardised performance. The Health Insights PRS AUC displayed in Figures 1 and 3 are the meta-analysed results across all available cohorts (apart from osteopenia where we use the UK Biobank cohort for all ancestries, excepting Native Americans, since this is the only cohort to contain both genders, and the Women's Health Initiative cohort for Native Americans, as this is the only cohort available for that ancestry). The comparator AUC numbers are as published by several other organisations (also considering models where the PRS is regressed against disease status in the absence of any covariates); see Supplementary Table 1 for AUC values and sources.

Estimation of absolute risk

We estimated absolute risk using an approach similar to that presented by Gail et al. [18], appropriately accounting for the probability that someone has not been previously diagnosed with the disease and has not died from other causes. Calculation of an absolute risk score that does not account for an individual's PRS, which we call a 'baseline model', requires appropriate baseline rates of disease (see below). In addition, to calculate an absolute risk score that accounts for an individual's PRS, which we refer to as an 'integrated risk tool (IRT) model', we require an effect size associated with each unit increase of the PRS.

Baselines

US appropriate disease incidence, all-cause mortality and disease-specific mortality baselines were compiled for each disease for all of the relevant US SIREs and likewise for a subset of diseases for the UK, where UK prospective cohort data was used for validation purposes. Baselines are constructed to account for the fact that an individual's age is rounded down to the nearest integer.

PRS effect sizes

Where available training data has at least 100

© 2023 Genomics plc — All rights reserved

cases and 100 controls, the odds ratio (OR) per standard deviation of the PRS was directly estimated within each of the five high-level ancestry groups (EUR, EAS, SAS, AFR, AMR). For ancestry groups with insufficient data, we linearly interpolated the OR based on the principal component projection of an individual's genome.

Validation cohorts

We evaluated the performance of our integrated risk tools (IRTs) in independent prospective cohorts. We used subsets of UK Biobank (UKB) non-overlapping with training samples [19] for cancer traits, the Study of Osteoporotic Fractures (SOF) [20] and the Osteoporotic Fracture in Men Study (MrOS) [21] for osteopenia, and Atherosclerosis Risk in Communities (ARIC) [22] for the remaining traits. For UKB, ARIC, SOF and MrOS, we use self-reported race or ethnicity information to map to appropriate SIRE categories. For the SOF, MrOS and ARIC cohorts, we use White PNTS and Black PNTS baselines (as Hispanic information for individuals of White and Black SIRE is not available), which we shorten henceforth to "White" and "Black", respectively.

IRT calibration methods and statistics

We use a variety of approaches and metrics to measure IRT calibration. The following definitions are used:

Censored follow up time

The follow-up time available (with a limit at the relevant prediction period) for each individual in validation data. If the full follow up time is not available, we use the maximum follow up time that is available. For osteopenia, where we predict risk at a given age, there is no follow up time.

Censored absolute risk score

The individual absolute risk score, adjusted to match the relevant censored follow up time. For osteopenia, where there is no follow up time, we predict the risk for the age at which the individual had their femoral neck bone mineral density measurement taken.

O/E

The ratio of the observed number of disease onset events to the expected number of events: $\frac{\text{Observed number of incident cases}}{\text{Expected number of incident cases}}$ [23, 24]. The observed number of cases is the number of individuals that are incident cases within the censored follow up time periods. The expected number of cases is given by summing the censored absolute risk scores. Ignoring statistical fluctuations in estimation, if O/E$=$1, this means that the number of observed events matches the number of expected events. If O/E$>$1 or $<$1, this means that risk of developing a disease is being under- or over-estimated, respectively.

Calibration intercept and slope

The slope and intercept of the curve obtained by fitting a logistic regression of the observed case control status at the end of the censored follow up time against the censored absolute risk score. Ignoring statistical fluctuations in estimation, if the calibration slope $>$1, then there is insufficient spread in the risk predictions whereas the converse is true if the calibration slope $<$1 [24]. The intercept can be interpreted analogously to the O/E statistic (though, in this case, an intercept of 0 means that the expected and observed numbers of cases agree).

95% confidence intervals (CI's)

95% CIs were obtained using 2,000 bootstrap samples.

IRT performance metrics

We assessed the IRT's ability to distinguish between individuals with and without disease using Harrell's C statistic, which is a rank-based concordance metric that measures how well a risk score (specifically the uncensored risk score) is able to predict the observed sequence of events [25]. Values range between 0

© 2023 Genomics plc — All rights reserved

and 1, with 0.5 the expected value if risk scores were assigned at random, and 1 representing a score that is perfectly able to predict the order of events. Harrell's C statistic is not an appropriate statistical measure to use for remaining lifetime risk [26, 27] and as a result for this risk duration we report Harrell's C for 10 year and total lifetime risk. Survival time is the minimum of the time to diagnosis and the censored follow up time as defined above; cases for which the age at diagnosis is unknown are excluded from this metric. In the case of osteopenia, where there is no follow up time, Harrell's C statistic is calculated using the predicted risk for the age at which the femoral neck bone mineral density measurement was taken, along with their disease status at this age.

Results

Health insights PRS are predictive of disease in all major ancestry groups and outperform alternative risk scores

Assessment of the predictive performance of the PRSs derived for use in Health Insights shows that these are predictive of disease in all five major ancestry groups (where we have available data) for each of the eight Health Insights diseases (Figures 1 and 2). We note that for AF and T2D, where there are gaps in data availability, the PRS is predictive in individuals of AFR ancestry and so very likely to be predictive across all ancestries based on broader patterns observed in the portability of PRS

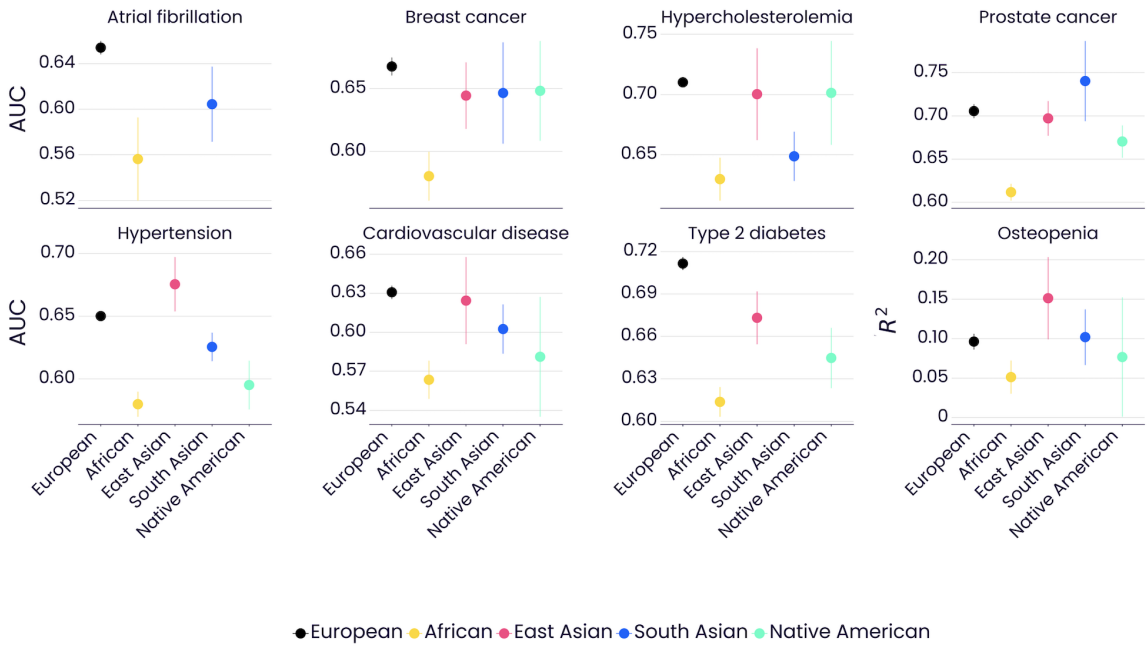

Figure 1: Predictive performance of Health Insights PRS within the five high level ancestry groups (European in black, African in yellow, East Asian in pink, South Asian in blue and Native American in green) where data is available. Predictive performance is measured using proportion of variance explained (R^2) for osteopenia where we are measuring the predictive performance of the underlying quantitative trait (bone mineral density) and area under the receiver operating channel curve (AUC) for all other traits. Error bars correspond to the 95% confidence intervals.

© 2023 Genomics plc — All rights reserved

across populations [28, 29].

Thus the PRS included within Health Insights have utility for all individuals regardless of ancestral background.

In addition, the Health Insights PRS outperform those developed by third parties (where available) within individuals of EUR ancestry (Figure 3), demonstrating the added value of the Genomics plc data resource and proprietary methods.

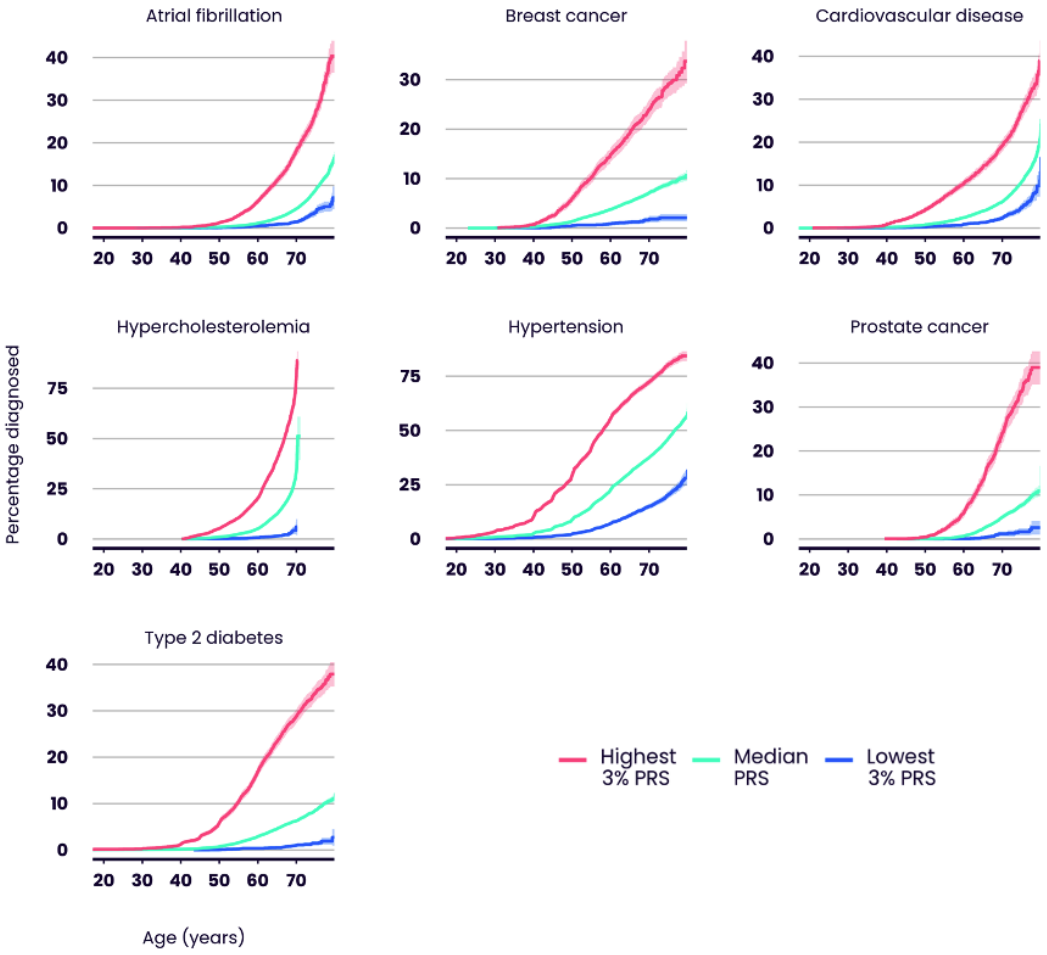

Figure 2: Cumulative disease incidence plot illustrating the predictive performance of the PRS in individuals of European ancestry in the UK Biobank cohort (using an independent subset of individuals from those used to train the PRS; ~85K individuals for breast cancer, ~75K individuals for prostate cancer and ~160K individuals for the other diseases). Colours indicate individuals in the highest 3% (red), median 40-60% (green) and lowest 3% (blue) of the PRS distribution. Shadings indicate 95% confidence intervals. All individuals of European ancestry with appropriate bone mineral density measurements within UK Biobank were used to train the osteopenia PRS, hence no plot for osteopenia is included.

© 2023 Genomics plc — All rights reserved

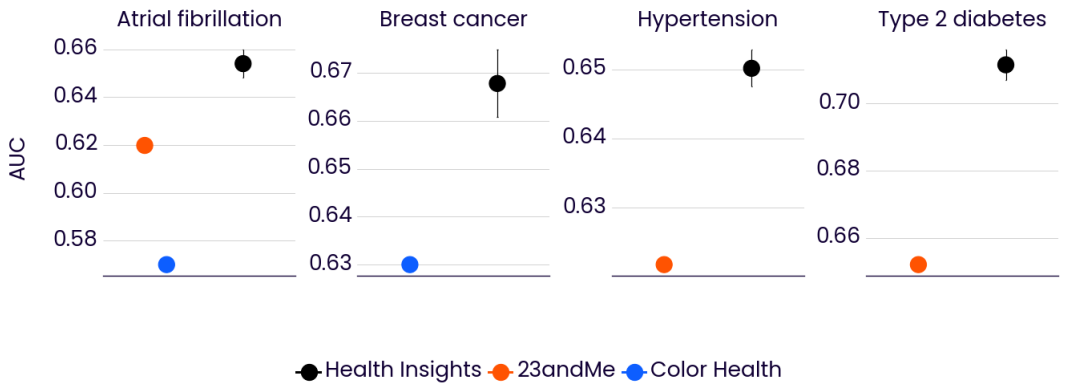

Figure 3: Predictive performance of Health Insights PRS (black) compared to 23andMe[1] (orange) and Color Health[2] (blue; see Methods and Supplementary Table 1 for further details) for traits where these are available evaluated within individuals of European ancestry. Predictive performance is measured using area under the receiver operating channel curve (AUC). Error bars correspond to the 95% confidence intervals for Health Insights PRS and are not available for 23andMe and Color Health.

[1]Atrial fibrillation and hypertension: https://permalinks.23andme.com/pdf/23_21-PRSMethodologyAppendix_May2020.pdf, May 2020
Type 2 diabetes: https://permalinks.23andme.com/pdf/23_19-Type2Diabetes_March2019.pdf, March 2019
[2]Atrial fibrillation and breast cancer: https://www.color.com/wp-content/uploads/2020/04/2019_Homburger_et_al_Genome_Medicine.pdf, November 2019

Health Insights disease risk estimates account for variation in disease risk by country, age, sex, race/ethnicity and adjust for competing mortality

Health Insights PRS are combined with baseline rates of disease through an IRT such that individuals receive their absolute risk of developing a disease. Baseline rates of disease capture socioeconomic determinants of health and as such vary over time, between countries and by ethnic groups (Figure 4). We therefore construct US-specific baselines using appropriate data sources that are race or ethnicity specific, at a minimum matching the granularity of the US census (including minority ethnic groups such as AIAN and NHPI) and, where appropriate, further separating the Asian ethnic group into East, South and South East Asian categories (Figure 4b).

Health Insights baselines capture disease incidence, disease specific mortality and all cause mortality broken down into one year age intervals, accounting for disease-specific mortality where appropriate using a competing synthetic risk model [18] and thus avoiding underestimation of disease risk. This effect is particularly important for diseases such as CVD, where a substantial fraction of individuals die from the disease before age 80; such deaths would not be captured using an approach that simply uses disease prevalence baselines.

Finally, to ensure that the results individuals receive as part of Health Insights are aligned with the United States Preventive Services Task Force guidelines, we provide risk scores tailored to the individual's current age.

© 2023 Genomics plc — All rights reserved

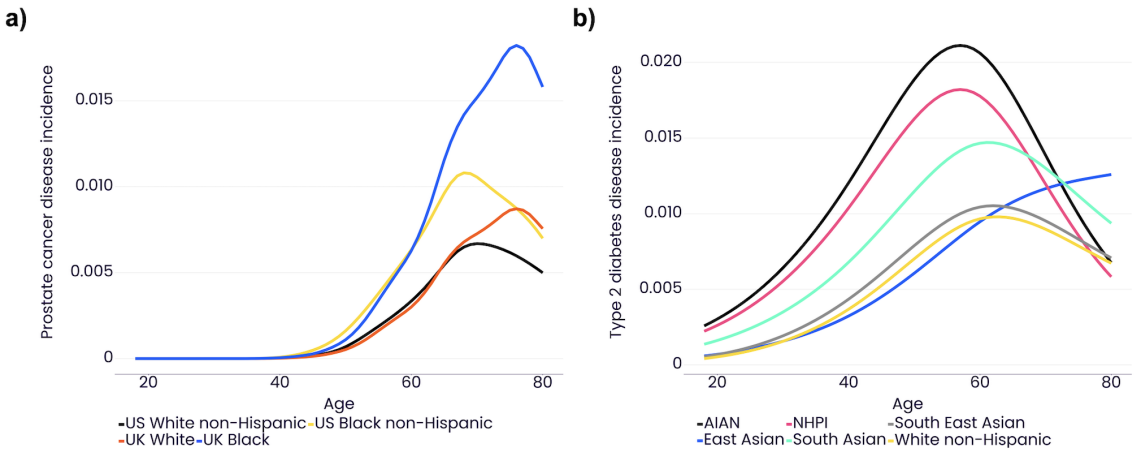

Figure 4: Baseline disease incidence for each one year age interval for two Health Insights diseases for selected self-identified races or ethnicities (SIREs). a) Prostate cancer in US individuals of White non-Hispanic (black) and Black non-Hispanic (yellow) SIRE based on data obtained from Surveillance, Epidemiology, and End Results [32] and UK individuals of White (orange) and Black (blue) SIRE based on data obtained from Cancer Research UK and from Delon et al. [33] and b) type 2 diabetes in female US individuals of American Indian or Alaskan Native (AIAN, black), Native Hawaiian or Other Pacific Islander (NHPI, pink), South East Asian (grey), East Asian (blue), South Asian (green) and White non-Hispanic (yellow) SIRE based on data obtained from the National Health Interview Survey.

Health Insights IRT models are strongly calibrated

We assess the accuracy of the risk scores generated by the Health Insights IRT models by comparing the predicted absolute risk scores to actual outcomes over the same time period within appropriate prospective cohort data (independent of any data used to train the models) across a range of available SIREs. Calibration of predictive models is challenging to assess due to the limited availability of long-term prospective data, coupled with cohort specific biases and disease incidence varying over time and by location due to differences in the environment. Nevertheless, where data is available, we find that the Health Insights IRTs are well calibrated when the baseline data matches the demographics of the prospective cohort.

For example, in the case of AF, we observe good calibration for individuals of both White (O/E [95% CI] = 1.013 [0.941, 1.085]) and Black race or ethnicity (O/E [95% CI] = 0.937 [0.771, 1.104]) using disease incidence data from 2004 and the ARIC validation cohort where data was collected several decades ago [22] (Figure 5). As AF disease incidence has substantially increased over time [30, 31], we have also generated up-to-date disease incidence baselines, which are used to return AF results to Health Insights users.

© 2023 Genomics plc — All rights reserved

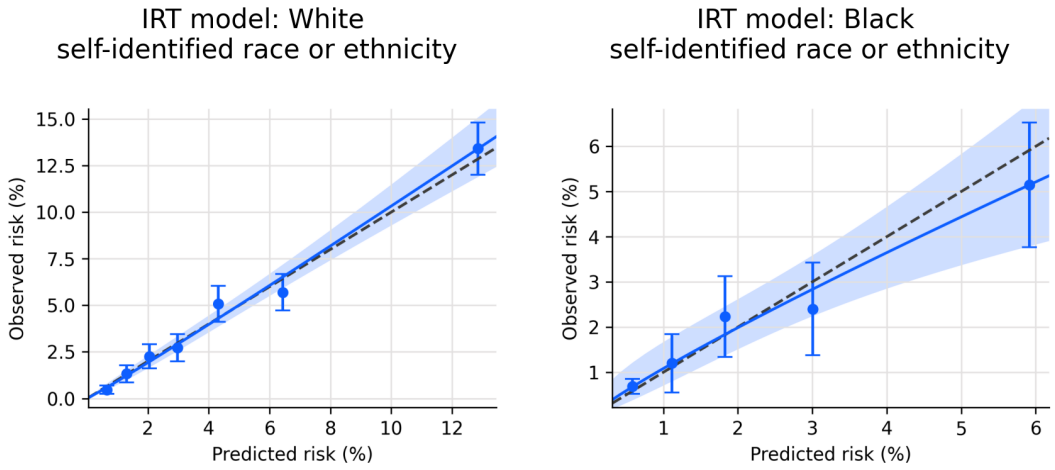

Figure 5: Calibration plots displaying the predicted ten year risk (x-axis) against the observed rate (y-axis) at which individuals develop atrial fibrillation (AF) for individuals of White (left) and Black (right) self-identified race or ethnicity (SIRE) using the ARIC validation set of individuals and 2004 baseline data. For those individuals without a complete ten years of follow up data, we computed the predicted risk of developing disease for the timeframe that we have observational data for the individual (censored predicted risk). Individuals are partitioned into approximately equal sized subgroups (either 5 or 7 subgroups depending on the number of cases) sorted by their censored predicted risk of developing the disease, within which we compare the mean predicted absolute disease risk to the observed disease rate (percentage of individuals in the group who get the disease) over the following ten years; the point estimate and 95% CIs for each group are represented by the vertical blue error bars. The fitted blue line is based on the parametric fit of the data along with 95% CIs. The dashed black line is the $y = x$ line and corresponds to ideal calibration.

Inclusion of the PRS in the risk score improves the predictive performance of IRTs

Calibration alone, whilst critical for providing individuals with accurate estimates of their absolute risk of developing diseases, does not highlight the value of including PRS to generate personalised risk scores. This is reflected in the improved ability to distinguish between cases and controls as measured by Harrell's C [25] for all available SIRE, sex and disease combinations (with $\geq$250 cases and controls; Figure 6). The greatest increase in performance, when considering a ten year risk window, occurs for T2D amongst White males, with Harrell's C increasing from 0.523 with the baseline model to 0.672 for the IRT model. This is also reflected by the increase in the range of the absolute risk scores across all traits (Figure 7). For example, ten year T2D absolute risk scores for White males in ARIC lie between 7.05% and 10.95% for the baseline model compared to 0.5% and 64.11% for the IRT model. We note that the bimodal distribution of the absolute risk scores using the baseline model (Figure 7) results from substantial differences in disease incidence between males and females. These results highlight the value of using an individual's PRS to provide tailored estimates of their absolute disease risk within a given time window.

© 2023 Genomics plc — All rights reserved

Figure 6: Difference in performance (as measured by Harrell's C) between the IRT model that contains the PRS and the baseline model that does not include the PRS on the y-axis for each self-identified race or ethnicity (SIRE) and sex combination (with $\geq$250 cases and controls) on the x-axis for all of the Health Insights traits considering all appropriate risk windows. Harrell's C difference for both sexes combined is displayed in blue, for females in orange and for males in yellow.

© 2023 Genomics plc — All rights reserved

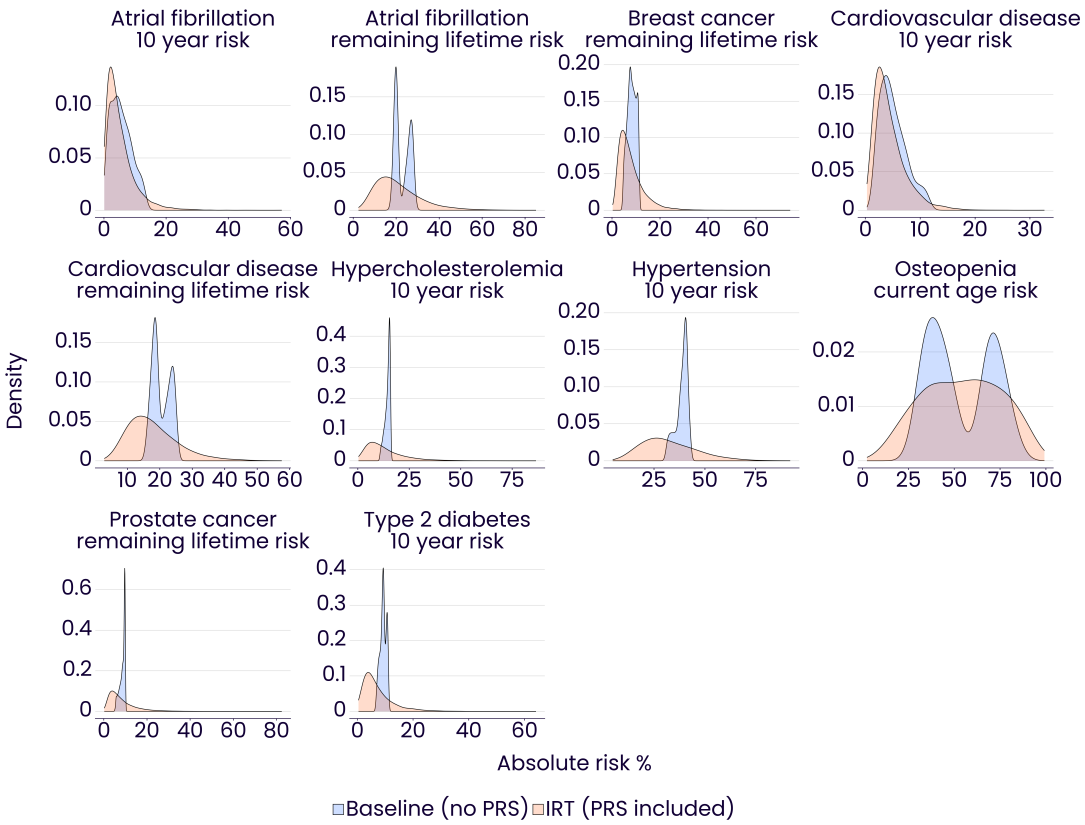

Figure 7: Distribution of the absolute risk scores for the baseline model (blue), which does not include the PRS, and the IRT model, which includes the PRS (orange), for all of the Health Insights traits considering all appropriate risk windows. Atrial fibrillation, cardiovascular disease, hypercholesterolemia, hypertension and type 2 diabetes distributions use absolute risk scores generated for White individuals in the ARIC cohort, breast and prostate cancer distributions use absolute risk scores generated for White non-Hispanic individuals of the appropriate sex in the UKB cohort and the osteopenia distribution uses absolute risk scores generated for White individuals in the SoF and MrOS cohorts. The bimodal distribution of the absolute risk scores using the baseline model results from substantial differences in disease incidence between males and females.

Summary

The Health Insights product includes personalised risk calculations (IRTs) for eight diseases that are common in the US population across racial/ethnic groups, costly to society, and can be mitigated by lifestyle changes and/or accessing readily-available medical pathways and screening programs. The IRTs combine personalised US-appropriate incident rates of disease, based on self-reported age, sex and race/ethnicity, with an individual's

© 2023 Genomics plc — All rights reserved

inherited polygenic disease risk, resulting in well-calibrated, individualised disease risk estimates. The analyses presented in this paper establish the evidence for the accuracy of the risk estimates and their ability to distinguish between those likely and unlikely to develop disease.

Health Insights does not consider medical history or other clinical and environmental risk factors that are known to influence disease risk. Rather, it uses limited, but readily available, self-reported user information, a property that maximizes ease-of-use and enables validation across all groups of users. The information provided should therefore be seen as complementary to that used by specific, disease-focused (non-genetic) disease risk calculators.

Supplementary tables

Trait	Comparator	AUC
Atrial fibrillation	23andMe	0.62
Atrial fibrillation	Color Health	0.57
Breast cancer	Color Health	0.63
Hypertension	23andMe	0.62
Type 2 Diabetes	23andMe	0.652

Supplementary table 1: PRS performance (AUC) in individuals of EUR ancestry of comparator PRS. In all cases the AUC is calculated using models that regress the PRS against disease status in the absence of any covariates.

References

[1] Cathryn M. Lewis and Evangelos Vassos. Polygenic risk scores: from research tools to clinical instruments. *Genome Medicine*, 12(1):44, December 2020.

[2] Amit V. Khera, Mark Chaffin, Seyedeh M. Zekavat, Ryan L. Collins, Carolina Roselli, Pradeep Natarajan, Judith H. Lichtman, Gail D'onofrio, Jennifer Mattera, Rachel Dreyer, John A. Spertus, Kent D. Taylor, Bruce M. Psaty, Stephen S. Rich, Wendy Post, Namrata Gupta, Stacey Gabriel, Eric Lander, Yii Der Ida Chen, Michael E. Talkowski, Jerome I. Rotter, Harlan M. Krumholz, and Sekar Kathiresan. Whole genome sequencing to characterize monogenic and polygenic contributions in patients hospitalized with early-onset myocardial infarction. *Circulation*, 139(13):1593–1602, March 2019. Publisher: Lippincott Williams and Wilkins.

[3] Lars G. Fritsche, Ying Ma, Daiwei Zhang, Maxwell Salvatore, Seunggeun Lee, Xiang Zhou, and Bhramar Mukherjee. On cross-ancestry cancer polygenic risk scores. *PLOS Genetics*, 17(9):e1009670, September 2021.

[4] Nasim Mavaddat, Kyriaki Michailidou, Joe Dennis, Michael Lush, Laura Fachal, Andrew Lee, and et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *American Journal of Human Genetics*, 104(1):21–34, January 2019. Publisher: Cell Press.

[5] Guochong Jia, Yingchang Lu, Wanqing Wen, Jirong Long, Ying Liu, Ran Tao, Bingshan Li, Joshua C Denny, Xiao-Ou Shu, and Wei Zheng. Evaluating the Utility of Polygenic Risk Scores in Identifying High-Risk Individuals for Eight Common Cancers. *JNCI Cancer Spectrum*, 4(3):pkaa021, June 2020.

[6] 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. UK Biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits,

© 2023 Genomics plc — All rights reserved

August 2022. ISSN: 2227-6246 Pages: 2022.06.16.22276246.

[7] 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. Integrated Polygenic Tool Substantially Enhances Coronary Artery Disease Prediction. *Circulation: Genomic and Precision Medicine*, 14(2), April 2021.

[8] Amit V. Khera, Mark Chaffin, Krishna G. Aragam, Mary E. Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S. Lander, Steven A. Lubitz, Patrick T. Ellinor, and Sekar Kathiresan. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9):1219–1224, 2018. Publisher: Springer US ISBN: 2,785/120,280.

[9] Nina Mars, Elisabeth Widén, Sini Kerminen, Tuomo Meretoja, Matti Pirinen, and Priit Palta. Polygenic risk, susceptibility genes, and breast cancer over the life course. *Nature Communications*, 11:6383, 2020.

[10] Michael E. Weale, Fernando Riveros-Mckay, Saskia Selzam, Priyanka Seth, Rachel Moore, William A. Tarran, Eva Gradovich, Carla Giner-Delgado, Duncan Palmer, Daniel Wells, Ayden Saffari, R. Michael Sivley, Alexander S. Lachapelle, Hannah Wand, Shoa L. Clarke, Joshua W. Knowles, Jack W. O'Sullivan, Euan A. Ashley, Gil McVean, Vincent Plagnol, and Peter Donnelly. Validation of an Integrated Risk Tool, Including Polygenic Risk Score, for Atherosclerotic Cardiovascular Disease in Multiple Ethnicities and Ancestries. *The American Journal of Cardiology*, 148:157–164, June 2021.

[11] 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. 2013 ACC/AHA Guideline on the assessment of cardiovascular risk. *Circulation*, 129(25 suppl 2):S49–S73, June 2014. Publisher: Lippincott Williams & Wilkins Hagerstown, MD.

[12] Donna K. Arnett, Roger S. Blumenthal, Michelle A. Albert, Andrew B. Buroker, Zachary D. Goldberger, Ellen J. Hahn, Cheryl Dennison Himmelfarb, Amit Khera, Donald Lloyd-Jones, J. William McEvoy, Erin D. Michos, Michael D. Miedema, Daniel Muñoz, Sidney C. Smith, Salim S. Virani, Kim A. Williams, Joseph Yeboah, and Boback Ziaeian. 2019 ACC/AHA Guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Journal of the American College of Cardiology*, 74(10):e177–e232, September 2019. Publisher: Elsevier USA.

[13] Sadiya S. Khan, Richard Cooper, and Philip Greenland. Do polygenic risk scores improve patient selection for prevention of coronary artery disease? *Journal of the American Medical Association*, 323(7):614–615, February 2020. Publisher: American Medical Association ISBN: 2018314211.

[14] Huaying Fang, Qin Hui, Julie Lynch, Jacqueline Honerlaw, Themistocles L. Assimes, Jie Huang, Marijana Vujkovic, Scott M. Damrauer, Saiju Pyarajan, J. Michael Gaziano, Scott L. DuVall, Christopher J. O'Donnell, Kelly Cho,

© 2023 Genomics plc — All rights reserved

Kyong-Mi Chang, Peter W.F. Wilson, Philip S. Tsao, Yan V. Sun, Hua Tang, J. Michael Gaziano, Rachel Ramoni, Jim Breeling, Kyong-Mi Chang, Grant Huang, Sumitra Muralidhar, Christopher J. O'Donnell, Philip S. Tsao, Sumitra Muralidhar, Jennifer Moser, Stacey B. Whitbourne, Jessica V. Brewer, John Concato, Stuart Warren, Dean P. Argyres, Brady Stephens, Mary T. Brophy, Donald E. Humphries, Nhan Do, Shahpoor Shayan, Xuan-Mai T. Nguyen, Saiju Pyarajan, Kelly Cho, Elizabeth Hauser, Yan Sun, Hongyu Zhao, Peter Wilson, Rachel McArdle, Louis Dellitalia, John Harley, Jeffrey Whittle, Jean Beckham, John Wells, Salvador Gutierrez, Gretchen Gibson, Laurence Kaminsky, Gerardo Villareal, Scott Kinlay, Junzhe Xu, Mark Hamner, Kathlyn Sue Haddock, Sujata Bhushan, Pran Iruvanti, Michael Godschalk, Zuhair Ballas, Malcolm Buford, Stephen Mastorides, Jon Klein, Nora Ratcliffe, Hermes Florez, Alan Swann, Maureen Murdoch, Peruvemba Sriram, Shing Shing Yeh, Ronald Washburn, Darshana Jhala, Samuel Aguayo, David Cohen, Satish Sharma, John Callaghan, Kris Ann Oursler, Mary Whooley, Sunil Ahuja, Amparo Gutierrez, Ronald Schifman, Jennifer Greco, Michael Rauchman, Richard Servatius, Mary Oehlert, Agnes Wallbom, Ronald Fernando, Timothy Morgan, Todd Stapley, Scott Sherman, Gwenevere Anderson, Elif Sonel, Edward Boyko, Laurence Meyer, Samir Gupta, Joseph Fayad, Adriana Hung, Jack Lichy, Robin Hurley, Brooks Robey, and Robert Striker. Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. *The American Journal of Human Genetics*, 105(4):763–772, October 2019.

[15] Dan Yu Lin and Patrick F. Sullivan. Meta-analysis of genome-wide association studies with overlapping subjects. *American Journal of Human Genetics*, 85(6):862–872, December 2009.

[16] Bjarni J. J. Vilhjálmsson, Jian Yang, Hilary K. K. Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, and Giulio et al. Genovese. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human Genetics*, 97(4):576–592, October 2015. Publisher: Elsevier ISBN: 1537-6605\r0002-9297.

[17] Charles E. Metz. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4):283–298, October 1978.

[18] Mitchell H. Gail, Louise A. Brinton, David P. Byar, Donald K. Corle, Sylvan B. Green, Catherine Schairer, and John J. Mulvihill. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*, 81(24):1879–1886, 1989.

[19] 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. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, October 2018. Publisher: Nature Publishing Group.

[20] Steven R. Cummings. Appendicular Bone Density and Age Predict Hip Fracture in Women. *JAMA: The Journal of the American Medical Association*, 263(5):665, February 1990.

[21] Janet Babich Blank, Peggy Mannen Cawthon, Mary Lou Carrion-Petersen, Loretta Harper, J. Phillip Johnson, Eileen Mitson, and Romelia Ramírez Delay. Overview of recruitment for the osteoporotic fractures in men study (MrOS).

© 2023 Genomics plc — All rights reserved

Contemporary Clinical Trials, 26(5):557–568, October 2005.

[22] The ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. *Am J Epidemiol.*, 129(4):687–702, 1989.

[23] Adam R. Brentnall and Jack Cuzick. Risk Models for Breast Cancer and Their Validation. *Statistical Science*, 35(1):14–30, 2020. 32226220.

[24] Richard J. Stevens and Katrina K. Poppe. Validation of clinical prediction models: what does the "calibration slope" really measure? *Journal of Clinical Epidemiology*, 118:93–99, February 2020.

[25] Frank E. Harrell, Robert M. Califf, David B. Pryor, Kerry L. Lee, Robert A. Rosati, Frank E Harrell Jr, Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the Yield of Medical Tests. *The Journal of the American Medical Association*, 247(18):2543–2546, May 1982.

[26] Sjw Willems, A Schat, Ms Van Noorden, and M Fiocco. Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator. *Statistical Methods in Medical Research*, 27(2):323–335, February 2018.

[27] Thomas A. Gerds, Michael W. Kattan, Martin Schumacher, and Changhong Yu. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine*, 32(13):2173–2184, June 2013.

[28] Yi Ding, Kangcheng Hou, Ziqi Xu, Aditya Pimplaskar, Ella Petter, Kristin Boulier, Florian Privé, Bjarni J. Vilhjálmsson, Loes M. Olde Loohuis, and Bogdan Pasaniuc. Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature*, 618(7966):774–781, June 2023.

[29] Florian Privé, Hugues Aschard, Shai Carmi, Lasse Folkersen, Clive Hoggart, Paul F. O'Reilly, and Bjarni J. Vilhjálmsson. Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *The American Journal of Human Genetics*, 109(1):12–23, January 2022.

[30] Brent A. Williams, Ashley M. Honushefsky, and Peter B. Berger. Temporal Trends in the Incidence, Prevalence, and Survival of Patients With Atrial Fibrillation From 2004 to 2016. *The American Journal of Cardiology*, 120(11):1961–1965, December 2017.

[31] Renate B Schnabel, Xiaoyan Yin, Philimon Gona, Martin G Larson, Alexa S Beiser, David D McManus, Christopher Newton-Cheh, Steven A Lubitz, Jared W Magnani, Patrick T Ellinor, Sudha Seshadri, Philip A Wolf, Ramachandran S Vasan, Emelia J Benjamin, and Daniel Levy. 50 year trends in atrial fibrillation prevalence, incidence, risk factors, and mortality in the Framingham Heart Study: a cohort study. *The Lancet*, 386(9989):154–162, July 2015.

[32] National Cancer Institute Surveillance Research Program. SEER*Explorer: An interactive website for SEER cancer statistics [Internet].

[33] Christine Delon, Katrina F. Brown, Nick W. S. Payne, Yannis Kotrotsios, Sally Vernon, and Jon Shelton. Differences in cancer incidence by broad ethnic group in England, 2013–2017. *British Journal of Cancer*, March 2022.

[34] Michael Blaise Cook, Zhaoming Wang, Edward D. Yeboah, Yao Tettey, Richard B. Biritwum, Andrew A. Adjei, Evelyn Tay, Ann Truelove, Shelley Niwa, Charles C. Chung, Annand P. Chokkalingam, Lisa W. Chu, Meredith Yeager, Amy Hutchinson, Kai Yu, Kristin A. Rand, Christopher A. Haiman, African Ancestry Prostate Cancer GWAS

© 2023 Genomics plc — All rights reserved

Consortium, Robert N. Hoover, Ann W. Hsing, and Stephen J. Chanock. A genome-wide association study of prostate cancer in West African men. *Human Genetics*, 133(5):509–521, May 2014.

Acknowledgements and data sources

Atherosclerosis Risk in Communities (ARIC)

The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institute of Health, Department of Health and Human Services, under contract numbers (HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I, and HHSN268201700005I). The authors thank the staff and participants of the ARIC study for their important contributions. Funding for GENEVA was provided by National Human Genome Research Institute grant U01HG004402 (E. Boerwinkle). The datasets used for the analyses described in this paper were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000090.v7.p1.

BCAC

The breast cancer genome-wide association analyses for BCAC and CIMBA were supported by Cancer Research UK (C1287/A10118, C1287/A16563, C1287/A10710, C12292/A20861, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692, C8197/A16565), The National Institutes of Health (CA128978, X01HG007492- the DRIVE consortium), the PERSPECTIVE project supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research (grant GPH-129344) and the Ministère de l'Économie, Science et Innovation du Québec through Genome Québec and the PSRSIIRI-701 grant, the Quebec Breast Cancer Foundation, the European Community's Seventh Framework Programme under grant agreement n° 223175 (HEALTH-F2-2009-223175) (COGS), the European Union's Horizon 2020 Research and Innovation Programme (634935 and 633784), the Post-Cancer GWAS initiative (U19 CA148537, CA148065 and CA148112 – the GAME-ON initiative), the Department of Defence (W81XWH-10-1-0341), the Canadian Institutes of Health Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer (CRN-87521), the Komen Foundation for the Cure, the Breast Cancer Research Foundation and the Ovarian Cancer Research Fund. All studies and funders are listed in Zhang H et al (Nat Genet, 2020).

Funding for BCAC and iCOGS came from: Cancer Research UK [grant numbers C1287/A16563, C1287/A10118, C1287/A10710, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692, C8197/A16565], the European Union's Horizon 2020 Research and Innovation Programme (grant numbers 634935 and 633784 for BRIDGES and B-CAST respectively), the European Community's Seventh Framework Programme under grant agreement n° 223175 [HEALTHF2-2009-223175] (COGS), the National Institutes of Health [CA128978] and Post-Cancer GWAS initiative [1U19 CA148537, 1U19 CA148065-01 (DRIVE) and 1U19 CA148112 - the GAME-ON initiative], the Department of Defence [W81XWH-10-1-0341], and the Canadian Institutes of Health Research CIHR) for the CIHR Team in Familial Risks of Breast Cancer [grant PSR-SIIRI-701]. All studies and funders as listed in Michailidou K et al (2013 and 2015) and in Guo Q et al (2015) are acknowledged for their contributions.

Biobank Japan

The datasets used for the analyses described in this paper were obtained from JENGER at http://jenger.riken.jp/en/result.

© 2023 Genomics plc — All rights reserved

CIMBA

The CIMBA data management and data analysis were supported by Cancer Research UK grants C12292/A20861, C12292/A11174. iCOGS: the European Community's Seventh Framework Programme under grant agreement no. 223175 (HEALTH-F2-2009-223175) (COGS), Cancer Research UK (C1287/A10118, C1287/A 10710, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692, C8197/A16565), the National Institutes of Health (CA128978) and Post-Cancer GWAS initiative (1U19 CA148537, 1U19 CA148065 and 1U19 CA148112 – the GAME-ON initiative), the Department of Defence (W81XWH-10-1-0341), the Canadian Institutes of Health Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer (CRN-87521), and the Ministry of Economic Development, Innovation and Export Trade (PSR-SIIRI-701), Komen Foundation for the Cure, the Breast Cancer Research Foundation, and the Ovarian Cancer Research Fund. The PERSPECTIVE project was supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, the Ministry of Economy, Science and Innovation through Genome Québec, and The Quebec Breast Cancer Foundation. All studies and funders are listed in Milne et al (Nat Genet, 2017) and Phelan et al (Nat Genet, 2017).

Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE)

OncoArray genotyping and phenotype data harmonization for the Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) breast-cancer case control samples was supported by X01 HG007491 and U19 CA148065 and by Cancer Research UK (C1287/A16563). Genotyping was conducted by the Center for Inherited Disease Research (CIDR), Centre for Cancer Genetic Epidemiology, University of Cambridge, and the National Cancer Institute. The following studies contributed germline DNA from breast cancer cases and controls: the Two Sister Study (2SISTER), Breast Oncology Galicia Network (BREOGAN), Copenhagen General Population Study (CGPS), Cancer Prevention Study 2 (CPSII), The European Prospective Investigation into Cancer and Nutrition (EPIC), Melbourne Collaborative Cohort Study (MCCS), Multiethnic Cohort (MEC), Nashville Breast Health Study (NBHS), Nurses Health Study (NHS), Nurses Health Study 2 (NHS2), Polish Breast Cancer Study (PBCS), Prostate Lung Colorectal and Ovarian Cancer Screening Trial (PLCO), Studies of Epidemiology and Risk Factors in Cancer Heredity (SEARCH), The Sister Study (SISTER), Swedish Mammographic Cohort (SMC), Women of African Ancestry Breast Cancer Study (WAABCS), Women's Health Initiative (WHI).

ELucidating Loci Involved in Prostate cancer SuscEptibility (ELLIPSE)

Genotyping of the OncoArray was funded by the US National Institutes of Health (NIH) [U19 CA 148537 for ELucidating Loci Involved in Prostate cancer SuscEptibility (ELLIPSE) project and X01HG007492 to the Center for Inherited Disease Research (CIDR) under contract number HHSN268201200008I]

The datasets used for the analyses described in this paper were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001391.v1.p1.

FinnGen

We want to acknowledge the participants and investigators of the FinnGen study.

Genes & Health

Genes & Health is/has recently been core-funded by Wellcome (WT102627, WT210561), the Medical Research Council (UK) (M009017), Higher Education Funding Council for England Catalyst, Barts Charity (845/1796), Health Data Research UK (for London substantive site), and research delivery support from the NHS National Institute for Health Research

© 2023 Genomics plc — All rights reserved

Clinical Research Network (North Thames). We thank Social Action for Health, Centre of The Cell, members of our Community Advisory Group, and staff who have recruited and collected data from volunteers. We thank the NIHR National Biosample Centre (UK Biocentre), the Social Genetic & Developmental Psychiatry Centre (King's College London), Wellcome Sanger Institute, and Broad Institute for sample processing, genotyping, sequencing and variant annotation. We thank: Barts Health NHS Trust, NHS Clinical Commissioning Groups (City and Hackney, Waltham Forest, Tower Hamlets, Newham, Redbridge, Havering, Barking and Dagenham), East London NHS Foundation Trust, Bradford Teaching Hospitals NHS Foundation Trust, Public Health England (especially David Wyllie), Discovery Data Service/Endeavour Health Charitable Trust (especially David Stables) - for GDPR-compliant data sharing backed by individual written informed consent. Most of all we thank all of the volunteers participating in Genes & Health.

Genomics England

This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support.

Ghana Prostate Study

The genome-wide association study of prostate cancer in West African men project was supported by the Intramural Research Program of the National Cancer Institute, National Institutes of Health, Department of Health and Human Services including Contract No. HHSN261200800001E. The datasets have been accessed through the NIH database for Genotypes and Phenotypes (dbGaP). A full list of acknowledgements can be found in the Supplementary Note [34] The datasets used for the analyses described in this paper were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000838.v1.p1.

Hispanic Community Health Study /Study of Latinos (HCHS/SOL)

The Hispanic Community Health Study/Study of Latinos is a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina (HHSN268201300001I / N01-HC-65233), University of Miami (HHSN268201300004I / N01-HC-65234), Albert Einstein College of Medicine (HHSN268201300002I / N01-HC-65235), University of Illinois at Chicago – HHSN268201300003I / N01-HC-65236 Northwestern Univ), and San Diego State University (HHSN268201300005I / N01-HC-65237). The following Institutes/ Centers/ Offices have contributed to the HCHS/SOL through a transfer of funds to the NHLBI: National Institute on Minority Health and Health Disparities, National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research, National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Neurological Disorders and Stroke, NIH Institution-Office of Dietary Supplements. The Genetic Analysis Center at the University of Washington was supported by NHLBI and NIDCR contracts (HHSN268201300005C, AM03 and MOD03). The datasets used for the analyses described in this paper were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000810.v1.p1 and http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000880.

© 2023 Genomics plc — All rights reserved

v1.p1.

Charles R. Bronfman Institute for Personalized Medicine (IPM)

The datasets used for the analyses described in this paper were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000388.v1.p1.

Million Veteran Program

The authors thank Million Veteran Program (MVP) staff, researchers, and volunteers, who have contributed to MVP, and especially participants who previously served their country in the military and now generously agreed to enroll in the study. (See https://www.research.va.gov/mvp/ for more details). The citation for MVP is Gaziano, J.M. et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. J Clin Epidemiol 70, 214-23 (2016). This research is based on data from the Million Veteran Program, Office of Research and Development, Veterans Health Administration, and was supported by the Veterans Administration (VA) Cooperative Studies Program (CSP) award #G002. The datasets used for the analyses described in this paper were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001672.v11.p1.

Multiethnic Cohort (MEC)

The Multiethnic Cohort and the genotyping in this study were funded by grants from the National Institute of Health (CA63464, CA54281, CA098758, CA132839 and HG005922) and the Department of Defense Breast Cancer Research Program (W81XWH-08-1-0383). The datasets used for the analyses described in this paper were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000517.v3.p1. Funding support for the GENEVA Prostate Cancer study was provided through the National Cancer Institute (R37CA54281, R01CA6364, P01CA33619, U01CA136792, and U01CA98758) and the National Human Genome Research Institute (U01HG004726). Assistance with phenotype harmonization, SNP selection, data cleaning, meta-analyses, data management and dissemination, and general study coordination, was provided by the GENEVA Coordinating Center (U01HG004789-01). The datasets used for the analyses described in this paper were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000306.v4.p1.

Funding support for the PAGE Multiethnic Cohort study was provided through the National Cancer Institute (R37CA54281, R01CA6364, P01CA33619, U01CA136792, and U01CA98758) and the National Human Genome Research Institute (U01HG004802). Assistance with phenotype harmonization, SNP selection, data cleaning, meta-analyses, data management and dissemination, and general study coordination, was provided by the PAGE Coordinating Center (U01HG004801-01). The datasets used for the analyses described in this paper were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000220.v1.p1.

Osteoporotic Fractures in Men (MrOS) Study

The data used for the analyses described in this document were obtained from the database of Genotypes and Phenotypes (dbGaP), at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000373.v1.p1. Genotype and phenotype data for the Osteoporotic Fractures in Men (MrOS) Study were provided by the Osteoporotic Fractures in Men Research Group. Funding support for the original study was provided by the National Institutes of Health, including the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS); the National Institute on Aging (NIA); the National Center for Research Resources (NCRR); and the National Heart, Lung, and Blood Institute (NHLBI).

© 2023 Genomics plc — All rights reserved

Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. Schumacher, Al Olama, Berndt, et al. Nature Genetics 2018, acknowledge "The PRACTICAL consortium, CRUK, BPC3, CAPS, PEGASUS"

The Prostate cancer genome-wide association analyses are supported by the Canadian Institutes of Health Research, European Commission's Seventh Framework Programme grant agreement no. 223175 (HEALTH-F2-2009-223175), Cancer Research UK Grants C5047/A7357, C1287/A10118, C1287/A16563, C5047/A3354, C5047/A10692, C16913/A6135, and The National Institute of Health (NIH) Cancer Post-Cancer GWAS initiative grant: No. 1 U19 CA 148537-01 (the GAME-ON initiative).

We would also like to thank the following for funding support: The Institute of Cancer Research and The Everyman Campaign, The Prostate Cancer Research Foundation, Prostate Research Campaign UK (now PCUK), The Orchid Cancer Appeal, Rosetrees Trust, The National Cancer Research Network UK, The National Cancer Research Institute (NCRI) UK. We are grateful for support of NIHR funding to the NIHR Biomedical Research Centre at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust.

The Prostate Cancer Program of Cancer Council Victoria also acknowledge grant support from The National Health and Medical Research Council, Australia (126402, 209057, 251533, , 396414, 450104, 504700, 504702, 504715, 623204, 940394, 614296,), VicHealth, Cancer Council Victoria, The Prostate Cancer Foundation of Australia, The Whitten Foundation, PricewaterhouseCoopers, and Tattersall's. EAO, DMK, and EMK acknowledge the Intramural Program of the National Human Genome Research Institute for their support.

Genotyping of the OncoArray was funded by the US National Institutes of Health (NIH) [U19 CA 148537 for ELucidating Loci Involved in Prostate cancer SuscEptibility (ELLIPSE) project and X01HG007492 to the Center for Inherited Disease Research (CIDR) under contract number HHSN268201200008I] and by Cancer Research UK grant A8197/A16565. Additional analytic support was provided by NIH NCI U01 CA188392 (PI: Schumacher).

Funding for the iCOGS infrastructure came from: the European Community's Seventh Framework Programme under grant agreement n° 223175 (HEALTH-F2-2009-223175) (COGS), Cancer Research UK (C1287/A10118, C1287/A 10710, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692, C8197/A16565), the National Institutes of Health (CA128978) and Post-Cancer GWAS initiative (1U19 CA148537, 1U19 CA148065 and 1U19 CA148112 – the GAME-ON initiative), the Department of Defence (W81XWH-10-1-0341), the Canadian Institutes of Health Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer, Komen Foundation for the Cure, the Breast Cancer Research Foundation, and the Ovarian Cancer Research Fund.

The BPC3 was supported by the U.S. National Institutes of Health, National Cancer Institute (cooperative agreements U01-CA98233 to D.J.H., U01-CA98710 to S.M.G., U01-CA98216 toE.R., and U01-CA98758 to B.E.H., and Intramural Research Program of NIH/National Cancer Institute, Division of Cancer Epidemiology and Genetics).

CAPS GWAS study was supported by the Swedish Cancer Foundation (grant no 09-0677, 11-484, 12-823), the Cancer Risk Prediction Center (CRisP; www.crispcenter.org), a Linneus Centre (Contract ID 70867902) financed by the Swedish Research Council, Swedish Research Council (grant no K2010-70X-20430-04-3, 2014-2269)

PEGASUS was supported by the Intramural Research Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health.

ROOT

The GWAS of Breast Cancer in the African Diaspora is conducted by the University of Chicago and supported by the National Cancer

© 2023 Genomics plc — All rights reserved

Institute (R01 CA142996-02). This manuscript was not prepared in collaboration with investigators of the GWAS of Breast Cancer in the African Diaspora and does not necessarily reflect the opinions or views of University of Chicago, or NCI. The datasets used for the analyses described in this paper were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000383.v1.p1.

Study of Osteoporotic Fractures (SOF)

The data used for the analyses described in this document were obtained from the database of Genotypes and Phenotypes (dbGaP), at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000510.v1.p1. Genotype and phenotype data for the Study of Osteoporotic Fractures (SOF) were provided by the Study of Osteoporotic Fractures Research Group. Funding support for the original study was provided by the National Institutes of Health, including the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) and the National Institute on Aging (NIA).

UK Biobank (UKB)

This research has been conducted using the UK Biobank Resource under Application Number 9659.

Women's Health Initiative (WHI)

The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C, and HHSN268201600004C. This document was not prepared in collaboration with investigators of the WHI, has not been reviewed and/or approved by the Women's Health Initiative (WHI), and does not necessarily reflect the opinions of the WHI investigators or the NHLBI. WHI PAGE is funded through the NHGRI Population Architecture Using Genomics and Epidemiology (PAGE) network (Grant Number U01 HG004790). Assistance with phenotype harmonization, SNP selection, data cleaning, meta-analyses, data management and dissemination, and general study coordination, was provided by the PAGE Coordinating Center (U01HG004801-01). Funding for WHI SHARe genotyping was provided by NHLBI Contract N02-HL-64278. The data used for the analyses described in this document were obtained from the database of Genotypes and Phenotypes (dbGaP), at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000200.v12.p3.

© 2023 Genomics plc — All rights reserved