Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes

Nasim Mavaddat,[1,*] Kyriaki Michailidou,[1,2] Joe Dennis,[1] Michael Lush,[1] Laura Fachal,[3] Andrew Lee,[1] Jonathan P. Tyrer,[3] Ting-Huei Chen,[4] Qin Wang,[1] Manjeet K. Bolla,[1] Xin Yang,[1] Muriel A. Adank,[5] Thomas Ahearn,[6] Kristiina Aittomäki,[7] Jamie Allen,[1] Irene L. Andrulis,[8,9] Hoda Anton-Culver,[10] Natalia N. Antonenkova,[11] Volker Arndt,[12] Kristan J. Aronson,[13] Paul L. Auer,[14,15] Päivi Auvinen,[16,17,18] Myrto Barrdahl,[19] Laura E. Beane Freeman,[6] Matthias W. Beckmann,[20] Sabine Behrens,[19] Javier Benitez,[21,22] Marina Bermisheva,[23] Leslie Bernstein,[24] Carl Blomqvist,[25,26] Natalia V. Bogdanova,[11,27,28] Stig E. Bojesen,[29,30,31] Bernardo Bonanni,[32] Anne-Lise Børresen-Dale,[33,34] Hiltrud Brauch,[35,36,37] Michael Bremer,[27] Hermann Brenner,[12,37,38] Adam Brentnall,[39] Ian W. Brock,[40] Angela Brooks-Wilson,[41,42] Sara Y. Brucker,[43] Thomas Brüning,[44] Barbara Burwinkel,[45,46] Daniele Campa,[19,47] Brian D. Carter,[48] Jose E. Castelao,[49] Stephen J. Chanock,[6] Rowan Chlebowski,[50] Hans Christiansen,[27] Christine L. Clarke,[51] J. Margriet Collée,[52] Emilie Cordina-Duverger,[53] Sten Cornelissen,[54] Fergus J. Couch,[55] Angela Cox,[40] Simon S. Cross,[56] Kamila Czene,[57]

(Author list continued on next page)

Stratification of women according to their risk of breast cancer based on polygenic risk scores (PRSs) could improve screening and prevention strategies. Our aim was to develop PRSs, optimized for prediction of estrogen receptor (ER)-specific disease, from the largest available genome-wide association dataset and to empirically validate the PRSs in prospective studies. The development dataset comprised 94,075 case subjects and 75,017 control subjects of European ancestry from 69 studies, divided into training and validation sets. Samples were genotyped using genome-wide arrays, and single-nucleotide polymorphisms (SNPs) were selected by stepwise regression or lasso penalized regression. The best performing PRSs were validated in an independent test set comprising 11,428 case subjects and 18,323 control subjects from 10 prospective studies and 190,040 women from UK Biobank (3,215 incident breast cancers). For the best PRSs (313 SNPs), the odds ratio for overall disease per 1 standard deviation in ten prospective studies was 1.61 (95%CI: 1.57–1.65) with area under receiver-operator curve (AUC) = 0.630 (95%CI: 0.628–0.651). The lifetime risk of overall breast cancer in the top centile of the PRSs was 32.6%. Compared with women in the middle quintile, those in the highest 1% of risk had 4.37- and 2.78-fold risks, and those in the lowest 1% of risk had 0.16- and 0.27-fold risks, of developing ER-positive and ER-negative disease, respectively. Goodness-of-fit tests indicated that this PRS was well calibrated and predicts disease risk accurately in the tails of the distribution. This PRS is a powerful and reliable predictor of breast cancer risk that may improve breast cancer prevention programs.

Introduction

Breast cancer is the most common cancer diagnosed among women in Western countries. While rare mutations in genes such as *BRCA1* and *BRCA2* confer high risks of developing breast cancer, these account for only a small proportion of breast cancer cases in the general population. Multiple common breast cancer susceptibility variants discovered through genome-wide association studies (GWASs)[1,2] confer small risk individually, but their combined effect, when

[1]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK; [2]Department of Electron Microscopy/Molecular Pathology, The Cyprus Institute of Neurology and Genetics, 1683 Nicosia, Cyprus; [3]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge CB1 8RN, UK; [4]Department of Mathematics and Statistics, Laval University, Québec City, QC G1V 0A6, Canada; [5]Family Cancer Clinic, the Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital, Amsterdam, 1066 CX, the Netherlands; [6]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD 20850, USA; [7]Department of Clinical Genetics, Helsinki University Hospital, University of Helsinki, Helsinki 00290, Finland; [8]Fred A. Litwin Center for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, ON M5G 1X5, Canada; [9]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada; [10]Department of Epidemiology, Genetic Epidemiology Research Institute, University of California Irvine, Irvine, CA 92617, USA; [11]NN Alexandrov Research Institute of Oncology and Medical Radiology, Minsk 223040, Belarus; [12]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany; [13]Department of Public Health Sciences, and Cancer Research Institute, Queen's University, Kingston, ON K7L 3N6, Canada; [14]Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; [15]Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI 53205, USA; [16]Cancer Center, Kuopio University Hospital, Kuopio 70210, Finland; [17]Institute of Clinical Medicine, Oncology, University of Eastern Finland, Kuopio 70210, Finland; [18]Translational Cancer Research Area, University of Eastern Finland, Kuopio 70210, Finland; [19]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany; [20]Department of Gynecology and Obstetrics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, Erlangen 91054, Germany; [21]Human Cancer Genetics Programme, Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain; [22]Biomedical Network on Rare Diseases (CIBERER), Madrid 28029, Spain; [23]Institute of Biochemistry and Genetics, Ufa Scientific Center of Russian Academy of Sciences, Ufa 450054, Russia; [24]Department of Population Sciences, Beckman Research Institute of City of Hope, Duarte, CA 91010, USA; [25]Department of Oncology, Helsinki University Hospital, University of Helsinki, Helsinki 00290, Finland; [26]Department of Oncology, Örebro University Hospital, Örebro 70185, Sweden; [27]Department of Radiation Oncology, Hannover Medical

(Affiliations continued on next page)

Mary B. Daly,[58] Peter Devilee,[59,60] Thilo Dörk,[28] Isabel dos-Santos-Silva,[61] Martine Dumont,[62] Lorraine Durcan,[63,64] Miriam Dwek,[65] Diana M. Eccles,[64] Arif B. Ekici,[66] A. Heather Eliassen,[67,68] Carolina Ellberg,[69] Christoph Engel,[70,71] Mikael Eriksson,[57] D. Gareth Evans,[72,73] Peter A. Fasching,[20,74] Jonine Figueroa,[6,75,76] Olivia Fletcher,[77] Henrik Flyger,[78] Asta Försti,[79,80] Lin Fritschi,[81] Marike Gabrielson,[57] Manuela Gago-Dominguez,[82,83] Susan M. Gapstur,[48] José A. García-Sáenz,[84] Mia M. Gaudet,[48] Vassilios Georgoulias,[85] Graham G. Giles,[86,87,88] Irina R. Gilyazova,[23,89] Gord Glendon,[8] Mark S. Goldberg,[90,91] David E. Goldgar,[92] Anna González-Neira,[21] Grethe I. Grenaker Alnæs,[33] Mervi Grip,[93] Jacek Gronwald,[94] Anne Grundy,[95] Pascal Guénel,[53] Lothar Haeberle,[20] Eric Hahnen,[96,97] Christopher A. Haiman,[98] Niclas Håkansson,[99] Ute Hamann,[100] Susan E. Hankinson,[101] Elaine F. Harkness,[102,103,104] Steven N. Hart,[105] Wei He,[57] Alexander Hein,[20] Jane Heyworth,[106] Peter Hillemanns,[28] Antoinette Hollestelle,[107] Maartje J. Hooning,[107] Robert N. Hoover,[6] John L. Hopper,[87] Anthony Howell,[108] Guanmengqian Huang,[100] Keith Humphreys,[57] David J. Hunter,[68,109,110]

(Author list continued on next page)

summarized as a polygenic risk score (PRS), can be substantial.[3–5] Such genomic profiles can be used to stratify women according to their risk of developing breast cancer.[6] This in turn holds the promise of improved breast cancer prevention and survival, by targeting screening or other preventative strategies at those women most likely to benefit.

We previously derived a PRS based on 77 established breast cancer susceptibility single-nucleotide polymorphisms (SNPs) and reported levels of risk stratification achieved by this PRS.[7] Based on our findings, several studies have investigated the potential for combining PRSs and other known risk factors for risk stratification and evaluated the impact of risk reduction strategies across risk strata defined by the PRS.[8–10] Preliminary studies investigating the use of the PRS to inform targeted breast cancer screening programs are underway (see CORDIS

School, Hannover 30625, Germany; [28]Gynaecology Research Unit, Hannover Medical School, Hannover 30625, Germany; [29]Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev 2730, Denmark; [30]Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev 2730, Denmark; [31]Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 2200, Denmark; [32]Division of Cancer Prevention and Genetics, IEO, European Institute of Oncology IRCCS, Milan 20141, Italy; [33]Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalet, Oslo 0379, Norway; [34]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo 0450, Norway; [35]Dr Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart 70376, Germany; [36]University of Tübingen, Tübingen 72074, Germany; [37]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg 69120, Germany; [38]Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg 69120, Germany; [39]Centre for Cancer Prevention, Wolfson Institute of Preventive Medicine, Queen Mary University of London, London EC1M 6BQ, UK; [40]Sheffield Institute for Nucleic Acids (SInFoNiA), Department of Oncology and Metabolism, University of Sheffield, Sheffield S10 2TN, UK; [41]Genome Sciences Centre, BC Cancer Agency, Vancouver, BC V5Z 1L3, Canada; [42]Department of Biomedical Physiology and Kinesiology, Simon Fraser University, Burnaby, BC V5A 1S6, Canada; [43]Department of Gynecology and Obstetrics, University of Tübingen, Tübingen 72076, Germany; [44]Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), Bochum 44789, Germany; [45]Department of Obstetrics and Gynecology, University of Heidelberg, Heidelberg 69120, Germany; [46]Molecular Epidemiology Group, C080, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany; [47]Department of Biology, University of Pisa, Pisa 56126, Italy; [48]Epidemiology Research Program, American Cancer Society, Atlanta, GA 30303, USA; [49]Oncology and Genetics Unit, Instituto de Investigacion Sanitaria Galicia Sur (IISGS), Xerencia de Xestion Integrada de Vigo-SERGAS, Vigo 36312, Spain; [50]Division of Medical Oncology and Hematology, University of California at Los Angeles, Los Angeles, CA 90024, USA; [51]Westmead Institute for Medical Research, University of Sydney, Sydney, NSW 2145, Australia; [52]Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam 3015 CN, the Netherlands; [53]Cancer & Environment Group, Center for Research in Epidemiology and Population Health (CESP), INSERM, University Paris-Sud, University Paris-Saclay, Villejuif 94805, France; [54]Division of Molecular Pathology, the Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, Amsterdam 1066 CX, the Netherlands; [55]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905, USA; [56]Academic Unit of Pathology, Department of Neuroscience, University of Sheffield, Sheffield S10 2TN, UK; [57]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm 171 65, Sweden; [58]Department of Clinical Genetics, Fox Chase Cancer Center, Philadelphia, PA 19111, USA; [59]Department of Pathology, Leiden University Medical Center, Leiden 2333 ZA, the Netherlands; [60]Department of Human Genetics, Leiden University Medical Center, Leiden 2333 ZA, the Netherlands; [61]Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK; [62]Genomics Center, Centre Hospitalier Universitaire de Québec - Université Laval Research Center, Université Laval, Québec City, QC G1V 4G2, Canada; [63]Southampton Clinical Trials Unit, Faculty of Medicine, University of Southampton, Southampton SO17 6YD, UK; [64]Cancer Sciences Academic Unit, Faculty of Medicine, University of Southampton, Southampton SO17 6YD, UK; [65]School of Life Sciences, University of Westminster, London W1B 2HW, UK; [66]Institute of Human Genetics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, Erlangen 91054, Germany; [67]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA; [68]Department of Epidemiology, Harvard TH Chan School of Public Health, Boston, MA 02115, USA; [69]Department of Cancer Epidemiology, Clinical Sciences, Lund University, Lund 222 42, Sweden; [70]Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig 04107, Germany; [71]LIFE - Leipzig Research Centre for Civilization Diseases, University of Leipzig, Leipzig 04103, Germany; [72]Division of Evolution and Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester M13 9WL, UK; [73]North West Genomic Laboratory Hub, Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester M13 9WL, UK; [74]David Geffen School of Medicine, Department of Medicine Division of Hematology and Oncology, University of California at Los Angeles, Los Angeles, CA 90095, USA; [75]Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh Medical School, Edinburgh EH16 4TJ, UK; [76]Cancer Research UK Edinburgh Centre, Edinburgh EH4 2XR, UK; [77]The Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research, London SW7 3RP, UK; [78]Department of Breast Surgery, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev 2730, Denmark; [79]Division of Molecular Genetic Epidemiology, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany; [80]Center for Primary Health Care Research, Clinical Research Center, Lund University, Malmö 205 02, Sweden; [81]School of Public Health, Curtin University, Perth, WA 6102,

(Affiliations continued on next page)

Milena Jakimovska,[111] Anna Jakubowska,[94,112] Wolfgang Janni,[113] Esther M. John,[114] Nichola Johnson,[77] Michael E. Jones,[115] Arja Jukkola-Vuorinen,[116] Audrey Jung,[19] Rudolf Kaaks,[19] Katarzyna Kaczmarek,[94] Vesa Kataja,[18,117] Renske Keeman,[54] Michael J. Kerin,[118] Elza Khusnutdinova,[23,89] Johanna I. Kiiski,[119] Julia A. Knight,[120,121] Yon-Dschun Ko,[122] Veli-Matti Kosma,[18,123,124] Stella Koutros,[6] Vessela N. Kristensen,[33,34] Ute Krüger,[69] Tabea Kühl,[125] Diether Lambrechts,[126,127] Loic Le Marchand,[128] Eunjung Lee,[98] Flavio Lejbkowicz,[129] Jenna Lilyquist,[105] Annika Lindblom,[130] Sara Lindström,[131,132] Jolanta Lissowska,[133] Wing-Yee Lo,[35,36] Sibylle Loibl,[134] Jirong Long,[135] Jan Lubiński,[94] Michael P. Lux,[20] Robert J. MacInnis,[86,87] Tom Maishman,[63,64] Enes Makalic,[87] Ivana Maleva Kostovska,[111] Arto Mannermaa,[18,123,124] Siranoush Manoukian,[136] Sara Margolin,[137,138] John W.M. Martens,[107] Maria Elena Martinez,[83,139] Dimitrios Mavroudis,[85] Catriona McLean,[140] Alfons Meindl,[141] Usha Menon,[142] Pooja Middha,[19,143] Nicola Miller,[118] Fernando Moreno,[84] Anna Marie Mulligan,[144,145] Claire Mulot,[146]

(Author list continued on next page)

and GenomeCanada in Web Resources).[11,12] Empirical validation and characterization of the PRS in large-scale epidemiological studies has, however, not been carried out previously. In addition, more informative PRSs would improve the clinical utility of risk prediction. GWASs have now identified ~170 breast cancer susceptibility loci.[1,2] Moreover, genome-wide heritability estimates indicate that these loci explain only ~40% of the heritability explained by all common variants on genome-wide SNP arrays. This suggests that the discrimination provided by the PRS could be improved by incorporating variants associated at more liberal significance thresholds. In addition, many variants confer risks that differ by breast cancer subtype (estrogen-receptor [ER]-positive or -negative), suggesting that subtype-specific PRSs might allow better prediction of subtype-specific disease, including the more aggressive ER-negative breast cancer, and enable selection of women for preventative medication.

Australia; [82]Genomic Medicine Group, Galician Foundation of Genomic Medicine, Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Complejo Hospitalario Universitario de Santiago, SERGAS, Santiago de Compostela 15706, Spain; [83]Moores Cancer Center, University of California San Diego, La Jolla, CA 92093, USA; [84]Medical Oncology Department, Hospital Clínico San Carlos, Instituto de Investigación Sanitaria San Carlos (IdISSC), Centro Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid 28040, Spain; [85]Department of Medical Oncology, University Hospital of Heraklion, Heraklion 711 10, Greece; [86]Cancer Epidemiology & Intelligence Division, Cancer Council Victoria, Melbourne, VIC 3004, Australia; [87]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, VIC 3010, Australia; [88]Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, VIC 3004, Australia; [89]Department of Genetics and Fundamental Medicine, Bashkir State University, Ufa 450076, Russia; [90]Department of Medicine, McGill University, Montréal, QC H4A 3J1, Canada; [91]Division of Clinical Epidemiology, Royal Victoria Hospital, McGill University, Montréal, QC H4A 3J1, Canada; [92]Department of Dermatology and Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, UT 84112, USA; [93]Department of Surgery, Oulu University Hospital, University of Oulu, Oulu 90220, Finland; [94]Department of Genetics and Pathology, Pomeranian Medical University, Szczecin 71-252, Poland; [95]Centre de Recherche du Centre Hospitalier de Université de Montréal (CHUM), Université de Montréal, Montréal, QC H2X 0A9, Canada; [96]Center for Hereditary Breast and Ovarian Cancer, University Hospital of Cologne, Cologne 50937, Germany; [97]Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne 50931, Germany; [98]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA; [99]Institute of Environmental Medicine, Karolinska Institutet, Stockholm 171 77, Sweden; [100]Molecular Genetics of Breast Cancer, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany; [101]Department of Biostatistics & Epidemiology, University of Massachusetts, Amherst, Amherst, MA 1003, USA; [102]Division of Informatics, Imaging and Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester M13 9PT, UK; [103]Nightingale Breast Screening Centre, Wythenshawe Hospital, Manchester University NHS Foundation Trust, Manchester M23 9LT, UK; [104]NIHR Manchester Biomedical Research Centre, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester M13 9WL, UK; [105]Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA; [106]School of Population and Global Health, University of Western Australia, Perth, WA 6009, Australia; [107]Department of Medical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam 3015 CN, the Netherlands; [108]Division of Cancer Sciences, University of Manchester, Manchester M13 9PL, UK; [109]Program in Genetic Epidemiology and Statistical Genetics, Harvard TH Chan School of Public Health, Boston, MA 02115, USA; [110]Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, UK; [111]Research Centre for Genetic Engineering and Biotechnology "Georgi D. Efremov," Macedonian Academy of Sciences and Arts, Skopje 1000, Republic of Macedonia; [112]Independent Laboratory of Molecular Biology and Genetic Diagnostics, Pomeranian Medical University, Szczecin 71-252, Poland; [113]Department of Gynecology and Obstetrics, University Hospital Ulm, Ulm 89075, Germany; [114]Department of Medicine, Division of Oncology, Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA 94304, USA; [115]Division of Genetics and Epidemiology, The Institute of Cancer Research, London SM2 5NG, UK; [116]Department of Oncology, Tampere University Hospital, Tampere, Finland Box 2000, 33521 Tampere, Finland; [117]Central Finland Health Care District, Jyväskylä Central Hospital, Jyväskylä 40620, Finland; [118]Surgery, School of Medicine, National University of Ireland, Galway H91TK33, Ireland; [119]Department of Obstetrics and Gynecology, Helsinki University Hospital, University of Helsinki, Helsinki 00290, Finland; [120]Prosserman Centre for Population Health Research, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON M5T 3L9, Canada; [121]Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, ON M5T 3M7, Canada; [122]Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter Krankenhaus, Bonn 53177, Germany; [123]Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern Finland, Kuopio 70210, Finland; [124]Imaging Center, Department of Clinical Pathology, Kuopio University Hospital, Kuopio 70210, Finland; [125]Cancer Epidemiology Group, University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg 20246, Germany; [126]VIB Center for Cancer Biology, VIB, Leuven 3000, Belgium; [127]Laboratory for Translational Genetics, Department of Human Genetics, University of Leuven, Leuven 3000, Belgium; [128]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI 96813, USA; [129]Clalit National Cancer Control Center, Carmel Medical Center and Technion Faculty of Medicine, Haifa 35254, Israel; [130]Department of Molecular Medicine and Surgery, Karolinska Institutet, and Department of Clinical Genetics, Karolinska University Hospital, Stockholm 171 76, Sweden; [131]Department of Epidemiology, University of Washington School of Public Health, Seattle, WA 98195, USA; [132]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; [133]Department of Cancer Epidemiology and Prevention, M Sklodowska-Curie Cancer Center - Oncology Institute, Warsaw 02-034, Poland; [134]German Breast Group, GmbH, Neu Isenburg 63263, Germany; [135]Division of Epidemiology, Department

(Affiliations continued on next page)

Victor M. Muñoz-Garzon,[147] Susan L. Neuhausen,[24] Heli Nevanlinna,[119] Patrick Neven,[148] William G. Newman,[72,73] Sune F. Nielsen,[29,30] Børge G. Nordestgaard,[29,30,31] Aaron Norman,[105] Kenneth Offit,[149,150] Janet E. Olson,[105] Håkan Olsson,[69] Nick Orr,[151] V. Shane Pankratz,[152] Tjoung-Won Park-Simon,[28] Jose I.A. Perez,[153] Clara Pérez-Barrios,[154] Paolo Peterlongo,[155] Julian Peto,[61] Mila Pinchev,[129] Dijana Plaseska-Karanfilska,[111] Eric C. Polley,[105] Ross Prentice,[14] Nadege Presneau,[65] Darya Prokofyeva,[89] Kristen Purrington,[156] Katri Pylkäs,[157,158] Brigitte Rack,[113] Paolo Radice,[159] Rohini Rau-Murthy,[150] Gad Rennert,[129] Hedy S. Rennert,[129] Valerie Rhenius,[3] Mark Robson,[150] Atocha Romero,[154] Kathryn J. Ruddy,[160] Matthias Ruebner,[20] Emmanouil Saloustros,[161] Dale P. Sandler,[162] Elinor J. Sawyer,[163] Daniel F. Schmidt,[87,164] Rita K. Schmutzler,[96,97] Andreas Schneeweiss,[165] Minouk J. Schoemaker,[115] Fredrick Schumacher,[166] Peter Schürmann,[28] Lukas Schwentner,[113] Christopher Scott,[105] Rodney J. Scott,[167,168,169] Caroline Seynaeve,[107] Mitul Shah,[3] Mark E. Sherman,[170]

(Author list continued on next page)

Here, we used data from 79 studies conducted by the Breast Cancer Association Consortium (BCAC) to optimize PRSs for overall and subtype-specific disease, and we validate their performance in independent datasets.[1,13–15]

Material and Methods

Study Subjects and Genotyping

The dataset used for development of the PRSs comprised 94,075 breast cancer-affected case subjects and 75,017 control subjects of European ancestry from 69 studies in the BCAC (Tables S1 and S2). Data collection for individual studies is described previously.[1] Samples were genotyped using one of two arrays: iCOGS[13,14] and OncoArray.[1,15] The dataset was divided into a training and validation set. The validation set was randomly selected (approximately 10% of case and control subjects) from studies that had been genotyped with the OncoArray, after excluding studies of bilateral breast cancer, studies or sub-studies oversampling for family history, and individuals with *in situ* cancers or case subjects with unknown ER status.

The best PRSs were evaluated in an independent test dataset comprising 11,428 invasive breast cancer-affected case subjects

of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN 37232, USA; [136]Unit of Medical Genetics, Department of Medical Oncology and Hematology, Fondazione IRCCS Istituto Nazionale dei Tumori (INT), Milan 20133, Italy; [137]Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet, Stockholm 118 83, Sweden; [138]Department of Oncology, Södersjukhuset, Stockholm 118 83, Sweden; [139]Department of Family Medicine and Public Health, University of California San Diego, La Jolla, CA 92093, USA; [140]Anatomical Pathology, The Alfred Hospital, Melbourne, VIC 3004, Australia; [141]Department of Gynecology and Obstetrics, Ludwig Maximilian University of Munich, Munich 80336, Germany; [142]MRC Clinical Trials Unit at UCL, Institute of Clinical Trials & Methodology, University College London, London WC1V 6LJ, UK; [143]Faculty of Medicine, University of Heidelberg, Heidelberg 69120, Germany; [144]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON M5S 1A8, Canada; [145]Laboratory Medicine Program, University Health Network, Toronto, ON M5G 2C4, Canada; [146]Université Paris Sorbonne Cité, INSERM UMR-S1147, Paris 75270, France; [147]Radiation Oncology, Hospital Meixoeiro-XXI de Vigo, Vigo 36214, Spain; [148]Leuven Multidisciplinary Breast Center, Department of Oncology, Leuven Cancer Institute, University Hospitals Leuven, Leuven 3000, Belgium; [149]Clinical Genetics Research Lab, Department of Cancer Biology and Genetics, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA; [150]Clinical Genetics Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA; [151]Centre for Cancer Research and Cell Biology, Queen's University Belfast, Belfast BT7 1NN, UK; [152]University of New Mexico Health Sciences Center, University of New Mexico, Albuquerque, NM 87131, USA; [153]Servicio de Cirugía General y Especialidades, Hospital Monte Naranco, Oviedo 33012, Spain; [154]Medical Oncology Department, Hospital Universitario Puerta de Hierro, Madrid 28222, Spain; [155]Genome Diagnostic Program, IFOM the FIRC (Italian Foundation for Cancer Research) Institute of Molecular Oncology, Milan 20139, Italy; [156]Department of Oncology, Wayne State University School of Medicine, Detroit, MI 48201, USA; [157]Laboratory of Cancer Genetics and Tumor Biology, Cancer and Translational Medicine Research Unit, Biocenter Oulu, University of Oulu, Oulu 90220, Finland; [158]Laboratory of Cancer Genetics and Tumor Biology, Northern Finland Laboratory Centre Oulu, Oulu 90220, Finland; [159]Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Research, Fondazione IRCCS Istituto Nazionale dei Tumori (INT), Milan 20133, Italy; [160]Department of Oncology, Mayo Clinic, Rochester, MN 55905, USA; [161]Department of Oncology, University Hospital of Larissa, Larissa 711 10, Greece; [162]Epidemiology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC 27709, USA; [163]Research Oncology, Guy's Hospital, King's College London, London SE1 9RT, UK; [164]Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia; [165]National Center for Tumor Diseases, University Hospital and German Cancer Research Center, Heidelberg 69120, Germany; [166]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH 44106, USA; [167]Division of Molecular Medicine, Pathology North, John Hunter Hospital, Newcastle, NSW 2305, Australia; [168]Discipline of Medical Genetics, School of Biomedical Sciences and Pharmacy, Faculty of Health, University of Newcastle, Callaghan, NSW 2308, Australia; [169]Hunter Medical Research Institute, John Hunter Hospital, Newcastle, NSW 2305, Australia; [170]Department of Health Sciences Research, Mayo Clinic College of Medicine, Jacksonville, FL 32224, USA; [171]Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, VIC 3168, Australia; [172]Department of Clinical Pathology, The University of Melbourne, Melbourne, VIC 3010, Australia; [173]Population Oncology, BC Cancer, Vancouver, BC V5Z 1G1, Canada; [174]School of Population and Public Health, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada; [175]Saarland Cancer Registry, Saarbrücken 66119, Germany; [176]The Curtin UWA Centre for Genetic Origins of Health and Disease, Curtin University and University of Western Australia, Perth, WA 6000, Australia; [177]Division of Breast Cancer Research, The Institute of Cancer Research, London SW7 3RP, UK; [178]Faculty of Medicine, University of Southampton, Southampton SO17 1BJ, UK; [179]Epigenetic and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC 27709, USA; [180]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY 10032, USA; [181]Department of Surgery, Leiden University Medical Center, Leiden 2333 ZA, the Netherlands; [182]Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham B15 2TT, UK; [183]Wellcome Trust Centre for Human Genetics and Oxford NIHR Biomedical Research Centre, University of Oxford, Oxford OX3 7BN, UK; [184]Department of Pathology, University Hospital of Heraklion, Heraklion 711 10, Greece; [185]Frauenklinik der Stadtklinik Baden-Baden, Baden-Baden 76532, Germany; [186]Department of Gynecology and Obstetrics, Helios Clinics Berlin-Buch, Berlin 13125, Germany; [187]Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC 27709, USA; [188]Department of Health Research and Policy - Epidemiology, Stanford

(Affiliations continued on next page)

Martha J. Shrubsole,[135] Xiao-Ou Shu,[135] Susan Slager,[105] Ann Smeets,[148] Christof Sohn,[165] Penny Soucy,[62] Melissa C. Southey,[171,172] John J. Spinelli,[173,174] Christa Stegmaier,[175] Jennifer Stone,[87,176] Anthony J. Swerdlow,[115,177] Rulla M. Tamimi,[67,68,109] William J. Tapper,[178] Jack A. Taylor,[162,179] Mary Beth Terry,[180] Kathrin Thöne,[125] Rob A.E.M. Tollenaar,[181] Ian Tomlinson,[182,183] Thérèse Truong,[53] Maria Tzardi,[184] Hans-Ulrich Ulmer,[185] Michael Untch,[186] Celine M. Vachon,[105] Elke M. van Veen,[72,73] Joseph Vijai,[149,150] Clarice R. Weinberg,[187] Camilla Wendt,[137,138] Alice S. Whittemore,[188,189] Hans Wildiers,[148] Walter Willett,[68,190,191] Robert Winqvist,[157,158] Alicja Wolk,[99,192] Xiaohong R. Yang,[6] Drakoulis Yannoukakos,[193] Yan Zhang,[12] Wei Zheng,[135] Argyrios Ziogas,[10] ABCTB Investigators,[194] kConFab/AOCS Investigators,[195] NBCS Collaborators,[33,34,196,197,198,199,200,201,202,203,204,205] Alison M. Dunning,[3] Deborah J. Thompson,[1] Georgia Chenevix-Trench,[206] Jenny Chang-Claude,[19,125] Marjanka K. Schmidt,[54,207] Per Hall,[57,138] Roger L. Milne,[86,87,171] Paul D.P. Pharoah,[1,3] Antonis C. Antoniou,[1] Nilanjan Chatterjee,[6,208,209] Peter Kraft,[68,109] Montserrat García-Closas,[6] Jacques Simard,[62] and Douglas F. Easton[1,3]

and 18,323 control subjects from ten studies nested within prospective cohorts, all genotyped using the OncoArray (Tables S3 and S4). The overall breast cancer PRS was also evaluated among 190,040 women of European ancestry from the UK Biobank cohort who had not had any cancer diagnosis or mastectomy prior to recruitment. A total of 3,215 incident registry-confirmed invasive breast cancers developed over 1,381,019 person years of prospective follow-up. Follow-up started 6 months after age of baseline questionnaire. The primary endpoint was invasive breast cancer. Follow-up was censored at the earliest of: risk-reducing mastectomy, diagnosis of any type of cancer, death, or January 15, 2017.

Genotype calling, quality control, and imputation for iCOGS and OncoArray were performed as previously described.[1,14] Briefly, imputation was performed for the iCOGS and OncoArray datasets separately using the Phase 3 (October 2014) release of the 1000 Genomes data as reference.[16] We followed a two-stage approach using SHAPEIT for phasing[17] and IMPUTE2 for the imputation.[15] Where samples were genotyped with iCOGS and OncoArray, the OncoArray calling was used. SNPs with MAF > 0.01 and imputation $r^2 > 0.9$ for OncoArray and $r^2 > 0.3$ for iCOGS were included in this analysis (~7 million SNPs); a higher threshold was imposed for OncoArray to ensure accurate determination of the PRS in the validation and test datasets.

UK Biobank samples were genotyped using Affymetrix UK BiLEVE Axiom array and Affymetrix UK Biobank Axiom array and imputed to the combined 1000 Genomes Project v.3 and UK10K reference panels using SHAPEIT3 and IMPUTE3.[18] The lowest imputation info score for the SNPs used in these analyses was 0.86. Samples were included on the basis of female sex (genetic and self-reported) and ethnicity filter (Europeans/White British ancestry subset). Duplicates, individuals with high degree of relatedness (>10 relatives), and one of each related pair of first degree relatives were removed. Samples were also excluded using standard quality control criteria.

Participants provided written informed consent, all studies were approved by the relevant ethics committees, and procedures followed were in accordance with the ethical standards of these committees.

Statistical Analysis

The general aim was to derive a PRS of the form:

$$\mathrm{PRS} = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \ldots + \beta_n x_n$$

where β_k is the per-allele log odds ratio (OR) for breast cancer associated with SNP k, x_k is the allele dosage for SNP k, and n is the total number of SNPs included in the PRS. Previous analyses found no evidence for statistically significant interactions between SNPs[19,20] and little evidence for departures from a log-additive model for individual SNPs. Assuming this is true in general, the PRS summarizes efficiently the combined effects of SNPs on disease risk.

The main challenge is how to determine which SNPs to include and the weighting parameters β_k to assign. Inclusion of only those SNPs reaching a stringent significance threshold ("genome-wide significant," $p < 5 \times 10^{-8}$) threshold ignores information from larger numbers of SNPs that are likely, but not certain, to be associated with the risk of breast cancer. We used two general

University School of Medicine, Stanford, CA 94305, USA; [189]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA 94305, USA; [190]Department of Nutrition, Harvard TH Chan School of Public Health, Boston, MA 02115, USA; [191]Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA; [192]Department of Surgical Sciences, Uppsala University, Uppsala 751 05, Sweden; [193]Molecular Diagnostics Laboratory, INRASTES, National Centre for Scientific Research "Demokritos," Athens 15310, Greece; [194]Australian Breast Cancer Tissue Bank, Westmead Institute for Medical Research, University of Sydney, Sydney, NSW 2145, Australia; [195]Peter MacCallum Cancer Center, Melbourne, VIC 3000, Australia; [196]Department of Research, Vestre Viken Hospital, Drammen 3019, Norway; [197]Department of Cancer Genetics, Vestre Viken Hospital, Drammen 3019, Norway; [198]Section for Breast and Endocrine Surgery, Department of Cancer, Division of Surgery, Cancer and Transplantation Medicine, Oslo University Hospital-Ullevål, Oslo 0450, Norway; [199]Department of Radiology and Nuclear Medicine, Oslo University Hospital, Oslo 0379, Norway; [200]Department of Pathology, Akershus University Hospital, Lørenskog 1478, Norway; [201]Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital, Oslo 0379, Norway; [202]Department of Oncology, Division of Surgery, Cancer and Transplantation Medicine, Oslo University Hospital-Radiumhospitalet, Oslo 0379, Norway; [203]National Advisory Unit on Late Effects after Cancer Treatment, Oslo University Hospital-Radiumhospitalet, Oslo 0379, Norway; [204]Department of Oncology, Akershus University Hospital, Lørenskog 1478, Norway; [205]Breast Cancer Research Consortium, Oslo University Hospital, Oslo 0379, Norway; [206]Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, QLD 4006, Australia; [207]Division of Psychosocial Research and Epidemiology, the Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital, Amsterdam 1066 CX, the Netherlands; [208]Department of Biostatistics, Bloomberg School of Public Health, John Hopkins University, Baltimore, MD 21205, USA; [209]Department of Oncology, School of Medicine, John Hopkins University, Baltimore, MD 21205, USA
*Correspondence: nm274@medschl.cam.ac.uk
https://doi.org/10.1016/j.ajhg.2018.11.002.

approaches for model selection: "hard-thresholding," based on a stepwise regression model that retained SNPs significantly associated with overall or subtype-specific disease at a given threshold, and penalized regression using lasso.[21,22] A schema for the analyses is shown in Figure S1.

To prioritize SNPs for analysis, single SNP association tests were first conducted in the training set. Per-allele ORs and standard errors were estimated separately in the iCOGS and OncoArray datasets, adjusting for study and nine ancestry informative principal components (PCs) in the iCOGS dataset and by country and ten PCs in the OncoArray dataset, using a purpose-written program.[1] Combined p values were then derived using a fixed-effects meta-analysis with the software METAL.[23] SNPs were sorted by p value and filtered on LD, such that uncorrelated SNPs (correlation $r^2 < 0.9$) with lowest p value for association with overall breast cancer in the training set were retained (more rigorous pruning, for example at $r^2 < 0.2$, would have removed from consideration informative SNPs from regions with multiple correlated signals[24,25]).

In the hard thresholding approach, a series of stepwise forward regression analyses were first carried out in 1 Mb regions centered on SNPs significant at a pre-specified threshold for association with either overall and/or subtype-specific disease in the training set. Only SNPs passing the specified p value thresholds were included in each 1 Mb region. Two analyses were performed in parallel: for overall breast cancer and ER-negative disease. At each stage the SNP with the smallest (conditional) p value for any analysis was added to the model, the threshold for the stepwise regression being the same as that for pre-selection. The process was repeated until no further SNPs could be added at the pre-defined threshold. A second stage of stepwise regressions were then carried out across all regions in each chromosome, to take into account correlated SNPs in different regions. Finally, the effect sizes for the selected SNPs were jointly estimated in a single logistic regression model.

For the best-performing PRSs, SNPs associated with ER-positive at $p < 10^{-6}$ but not with overall breast cancer (at $p < 10^{-5}$) were added at the end of the final SNP list. A third round of stepwise forward regression was then carried out with p value for selection of $p < 10^{-6}$ for ER-positive disease. For completeness we added to this final PRS two rarer variants (*BRCA2* p.Lys3326X and *CHEK2* p.Ile157Tyr) which are established to confer a moderate risk of breast cancer and were genotyped on the OncoArray but did not pass the allele frequency threshold in the PRS development phase.

For the penalized regression using lasso, we used the program *glmnet* [21]. SNPs with $p < 0.001$ in overall BC or ER-negative disease in the training set were pre-selected for inclusion in the lasso, and *BRCA2* p.Lys3326X and *CHEK2* p.Ile157Thr were added. Covariates for 19 PCs (9 for iCOGs and 10 for Oncoarray) and country were included in each model. For overall breast cancer, the penalty parameter (lambda) giving the best overall breast cancer PRS in the validation set was selected.

To construct subtype-specific PRSs, we evaluated four different methods: (1) using effect sizes for overall breast cancer (for each of the subtypes), (2) using effect sizes for subtype-specific (ER-positive or ER-negative) disease, (3) using a hybrid method, in which effect sizes were estimated in the relevant subtype for SNPs passing a certain optimal significance threshold in a case-only logistic regression (ER-positive versus ER-negative disease), and otherwise, using effect sizes estimated for overall breast cancer, or (4) by estimating case-only ORs using lasso and combining these with the overall breast cancer ORs to derive subtype-specific estimates, using the formulae:

$$\beta_{\text{ERpositive}} = \beta_{\text{overall}} + \eta * \beta_{\text{case-only}}$$

$$\beta_{\text{ERnegative}} = \beta_{\text{overall}} - (1 - \eta) * \beta_{\text{case-only}}$$

where $\eta = 0.27$ was the proportion of ER-negative tumors in the validation set.

For the lasso analysis, effect sizes for subtype-specific disease were estimated using method 4 above, combining the estimates from a case-only lasso analysis with the coefficients for overall breast cancer from the lasso analysis. The lambda for the case-only model giving the best subtype-specific PRS in the validation set was selected.

To evaluate the performance of each potential PRS, we standardized the PRSs to have unit standard deviation (SD) in the validation set of control subjects. The association of the standardized PRSs was evaluated in the validation and test (prospective studies) datasets, by logistic regression. We used a Cox proportional hazards regression model to assess the association with risk of breast cancer in UK Biobank. Models were also compared in terms of the area under the receiver operator characteristic curves (AUC), adjusted for study, calculated using the Stata command *comproc*. Meta-analysis of study-specific effects was carried out using the Stata command *metan*.

The goodness of fit of the continuous model (i.e., assuming a linear association between log(OR) and risk) was tested using the Hosmer-Lemeshow (HL) test to compare the observed and predicted risks by quantile and using the tail-based test proposed by Song et al.[26] In addition, we considered specifically the risks in the highest and lowest 1% of the distribution.

Effect modification of the PRS by age and family history of breast cancer in first-degree relatives was evaluated by fitting additional interaction terms in the model. The validation and prospective test datasets were combined for this analysis.

The absolute risks of developing breast cancer (overall and subtype-specific disease) were calculated taking into account the competing risk of dying from causes other than breast cancer, as described previously,[7] with the PRS modeled as a continuous covariate and including a linear "age × PRS" interaction term. The absolute risk of developing subtype-specific disease was obtained constraining to the incidence of overall incidence of ER-negative and ER-positive disease in the UK. Women are at risk of developing both ER-negative and ER-positive disease, so the absolute risks were calculated given that the individual has been free of breast cancer of any subtype.

Analyses were carried out in R v.3.0.2 and Stata v.14.2. All tests of statistical significance were two-sided. Further details are provided in the Supplemental Material and Methods.

Results

Development of the PRS

We tried several approaches to develop PRSs; here we report results for models giving the highest prediction accuracy. Using stepwise forward selection, the best PRS for prediction of overall breast cancer was obtained at a p value threshold for pre-selection and stepwise regression of $p < 10^{-5}$ (Table 1). The OR per unit standard deviation (SD) for this 305-SNP PRS with overall breast cancer in

Table 1. Comparison of Methods for Deriving the PRS: Results for Overall Breast Cancer in the Validation Set

p Value Cutoff[a]	SNPs Entering Model (n)	SNPs Selected (n)	OR[b]	95% CI	AUC
Published PRS[7]					
	77	77	1.49	1.44–1.56	0.612
Hard-Thresholding Stepwise Forward Regression					
$<5 \times 10^{-8}$	1,817	123	1.59	1.52–1.66	0.626
$<10^{-6}$	2,603	197	1.62	1.55–1.68	0.634
$<10^{-5}$	3,818	305	1.65	1.58–1.72	0.637
$<10^{-4}$	6,743	669	1.62	1.56–1.69	0.631
$<10^{-3}$	14,760	1,707	1.55	1.49–1.62	0.623
Penalized Regression					
Lasso	15,032	3,820	1.71	1.64–1.79	0.647

[a]The p value cut off refers to the SNPs considered based on their marginal associations in the training set; the same p value threshold was used in each case in the stepwise regression. Parameter selection and effect size estimation for derivation of the PRS was carried out in the training set as described in the Material and Methods.
[b]OR per 1 SD for the PRS. OR for association with breast cancer in the validation set was derived using logistic regression adjusting for country and ten PCs. AUCs were adjusted for country. The lasso was carried out after pre-selecting SNPs at $p < 10^{-3}$ based on their marginal association in the training set. For the lasso $\lambda = 0.003$ gave the optimal PRS in the validation set.

the validation set was 1.65 (95%CI: 1.58–1.72), compared with 1.59 (95%CI: 1.52–1.66) using a "genome-wide" ($p < 5 \times 10^{-8}$) threshold (123 SNPs).

Using lasso regression, the best PRS (OR = 1.71, 95%CI: 1.64–1.79) was more predictive than the best PRS developed using the stepwise regression model. In the best model ($\lambda = 0.003$), 3,820 SNPs were selected (Table 1).

Optimizing the PRS for Prediction of Subtype-Specific Disease

For evaluation of subtype-specific models following stepwise regression, SNP effect sizes were estimated, in the first instance, in each disease subtype. The best subtype-specific PRSs using this method were also obtained at a p value threshold of $p < 10^{-5}$ (Table S5). The 305-SNP PRS was supplemented with 6 additional SNPs associated with ER-positive at p value $< 10^{-6}$ and, in addition, by two known rare breast cancer susceptibility variants in the *BRCA2* and *CHEK2* genes, bringing the total number of SNPs included to 313 (PRS_{313}).

The optimum subtype-specific PRS was obtained when a subset of these 313 SNPs (196 SNPs with a case-only p value for association with ER-negative versus ER-positive disease of $p < 0.025$) were given subtype-specific weights, while the remaining SNPs were given overall breast cancer weights. For ER-negative disease, the OR improved from OR = 1.45 (95%CI: 1.35–1.56) to OR = 1.47 (95%CI: 1.37–1.58) using the hybrid method compared with using only subtype-specific estimates, while for ER-positive disease the results were similar (OR = 1.74) (Tables S6 and S7).

Subtype-specific prediction using the lasso analysis was optimized using case-only lasso analysis. The OR per 1 SD in the validation set was 1.81 (95%CI: 1.73–1.89) for ER-positive and 1.48 (95%CI: 1.37–1.59) for ER-negative disease (Tables 2 and S8).

Validation of the PRS in the Prospective Test Dataset

The final PRSs were evaluated using data from 11,428 invasive breast cancer-affected case subjects and 18,323 control subjects from ten prospective studies. The ORs for both the overall and subtype-specific PRSs were slightly lower in the prospective test set compared to the validation set (Table 2). The difference between validation and test set may reflect some overfitting due to choosing the optimum p value threshold and for the lasso, the optimum lambda, in the validation set, but could also be due to somewhat different characteristics of the prospective studies. The ORs for overall and ER-positive, but not ER-negative, breast cancer were slightly higher for the 3,820-SNP PRS (PRS_{3820}) compared with PRS_{313}.

The odds ratio (OR) for overall disease per 1 standard deviation (SD) of the PRS_{313} in the prospective studies was 1.61 (95%CI: 1.57–1.65) while for the 77-SNP PRS (PRS_{77}) derived previously OR = 1.46 (95%CI: 1.42–1.49). For ER-negative disease the difference was OR = 1.45 (95%CI: 1.37–1.53) versus 1.35 (95%CI: 1.27–1.43) (Table 2).

The associations between the PRS and overall, ER-positive, and ER-negative breast cancer by percentiles of the PRS_{313} are shown in Figure 1 and Table S9. Compared with women in the middle quintile (40th to 60th percentile), those in the highest 1% of risk for the subtype-specific PRS_{313} had 4.37 (95%CI: 3.59–5.33)- and 2.78 (95%CI: 1.83–4.24)-fold risks, and those in the lowest 1% had 0.16 (95%CI: 0.09–0.30)- and 0.27 (95%CI: 0.09–0.86)-fold risks of developing ER-positive and ER-negative disease, respectively. The ORs by percentile of the PRS_{3820} were similar (Table S10).

Goodness of Fit of the PRS

The remaining analyses concentrated on PRS_{313}. The associations between the PRS and breast cancer risk by

Table 2. Association between PRS and Breast Cancer Risk in the Validation Set and Prospective Test Datasets

	Validation Set			Prospective Test Set		
	OR[a]	95% CI	AUC	OR[a]	95% CI	AUC
77 SNP PRS (PRS_{77})						
Overall BC	1.49	1.44–1.56	0.612	1.46	1.42–1.49	0.603
ER-positive	1.56	1.49–1.63	0.623	1.52	1.48–1.56	0.615
ER-negative	1.40	1.30–1.50	0.596	1.35	1.27–1.43	0.584
313 SNP PRS (PRS_{313})						
Overall BC	1.65	1.59–1.72	0.639	1.61	1.57–1.65	0.630
ER-positive	1.74	1.66–1.82	0.651	1.68	1.63–1.73	0.641
ER-negative	1.47	1.37–1.58	0.611	1.45	1.37–1.53	0.601
3,820 SNP PRS (PRS_{3820})						
Overall BC	1.71	1.64–1.79	0.646	1.66	1.61–1.70	0.636
ER-positive	1.81	1.73–1.89	0.659	1.73	1.68–1.78	0.647
ER-negative	1.48	1.37–1.59	0.611	1.44	1.36–1.53	0.600

Parameter selection and effect size estimation for derivation of the PRS was carried out in the training set as described in the Material and Methods. The optimal subtype-specific PRS was obtained by carrying out case-only logistic regression and estimating effect sizes in the relevant subtype for SNPs passing a p value of 0.025 in case-only ordinary logistic regression (ER-positive versus ER-negative disease). OR for association with breast cancer in the validation set derived using logistic regression adjusting for country and ten PCs. AUCs were adjusted for by country. In the prospective test set, logistic regression models were adjusted for study and 15 PCs. AUCs were adjusted for by study.
[a]OR per 1 SD for the PRS.

percentiles of the risk score were compared with those predicted under a simple polygenic model with the PRS considered as a continuous covariate. The effect sizes did not differ from those predicted, and in particular the estimates for the highest and lowest centile were consistent with the predicted estimates (Table S9). Further tests for goodness of fit and tail-based tests (see Material and Methods) were not statistically significant at $p < 0.05$.

There was no evidence of heterogeneity in the effect sizes among studies (Figure 2). All studies showed a significant association with similar effect sizes for overall and ER-positive breast cancer, and all but one study (FHRISK, based on only six case subjects) showed a significant effect for ER-negative breast cancer.

In the UK Biobank, the estimated hazard ratio (HR) for overall breast cancer per unit PRS (including 306 of the 313 SNPs) was HR = 1.59 (95%CI: 1.54–1.64) (Figure 2).

By way of comparison, we also evaluated a PRS based on 177 previously published susceptibility loci.[1,2] The effect size for this PRS (OR = 1.61, 95%CI: 1.57–1.65) in the ten prospective studies was similar to the PRS_{313}. However, this estimated effect size is biased because the validation and test datasets used here contributed to the GWAS discovery datasets; in the UK Biobank this PRS (based on 174 of 177 available SNPs) performed worse (HR = 1.53, 95%CI: 1.48–1.58).

PRS Effects by Age

A weak decline in the OR with age was observed for ER-positive disease ($p = 0.001$, for the combined validation and test set). There was some evidence that the decline in PRS OR was not linear, driven by a lower estimate below age 40 years (Table S11, Figure S2). There was no evidence of a decline in the OR by age for ER-negative disease ($p = 0.39$).

Combined Effects of PRS and Breast Cancer Family History

The association between PRS and disease risk was observed for women with and without a family history (Table 3). However, there was some evidence that for ER-positive disease, the PRS OR was smaller in women with a family history (interaction OR = 0.91, $p = 0.004$). The log OR for family history was attenuated by 21% (1.59 to 1.44) and 12% (1.66 to 1.56) for ER-positive and ER-negative disease, respectively, after adjusting for the PRS (Tables 3 and S12).

Absolute Risk of Developing Breast Cancer According to the PRS

Estimated lifetime and 10-year absolute risks for UK women in percentiles of the PRS are shown in Figure 3. For ER-positive disease, the estimated lifetime absolute risk by age 80 years ranged from 2% for women in the lowest centile to 31% in the highest centile, while for ER-negative disease, the absolute risks ranged from 0.55% to 4%. The average 10-year absolute risk of breast cancer for a 47-year-old woman (i.e., the age at which women become eligible to enter the UK breast cancer screening program) in the general population is 2.6%. However, the 19% of women with the highest PRSs will attain this level of risk by age 40 years.

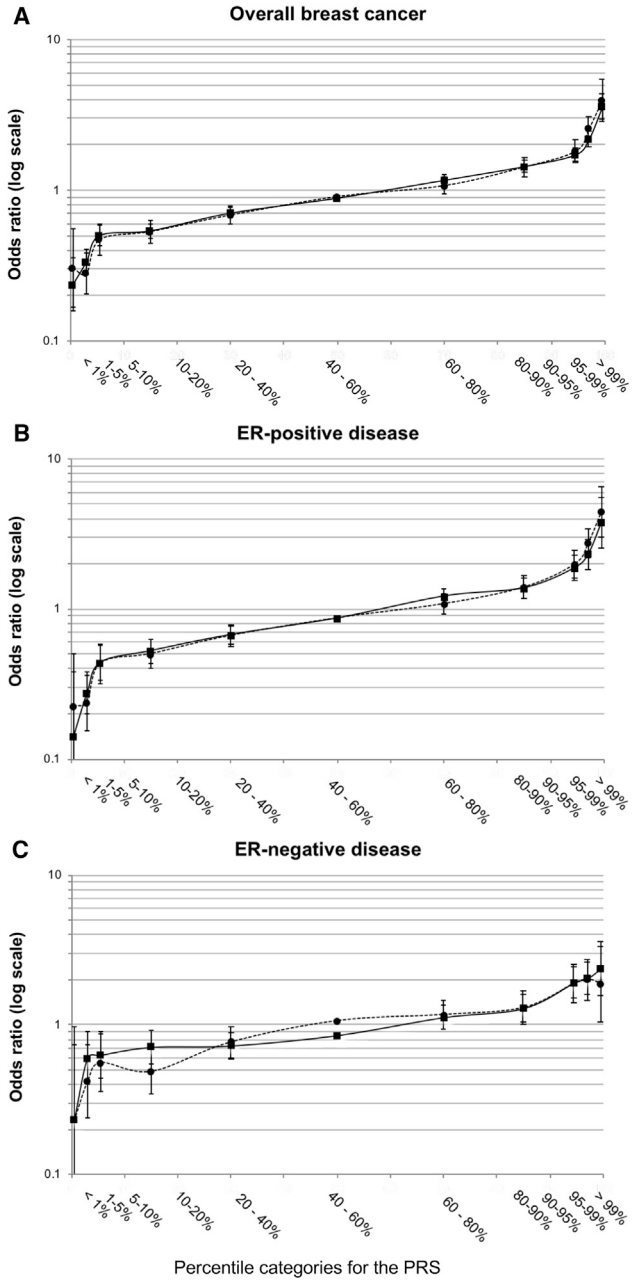

Figure 1. Association between the 313 SNP Polygenic Risk Score and Breast Cancer Risk

Association between the 313 SNP polygenic risk score (PRS) and breast cancer risk in women of European origin for (A) overall breast cancers, (B) estrogen receptor (ER)-positive disease, and (C) ER-negative disease, in the validation (dashed line) and test (solid line) sets. Odds ratios are for different quantiles of the PRS relative to the mean PRS. Odds ratios and 95% confidence intervals are shown.

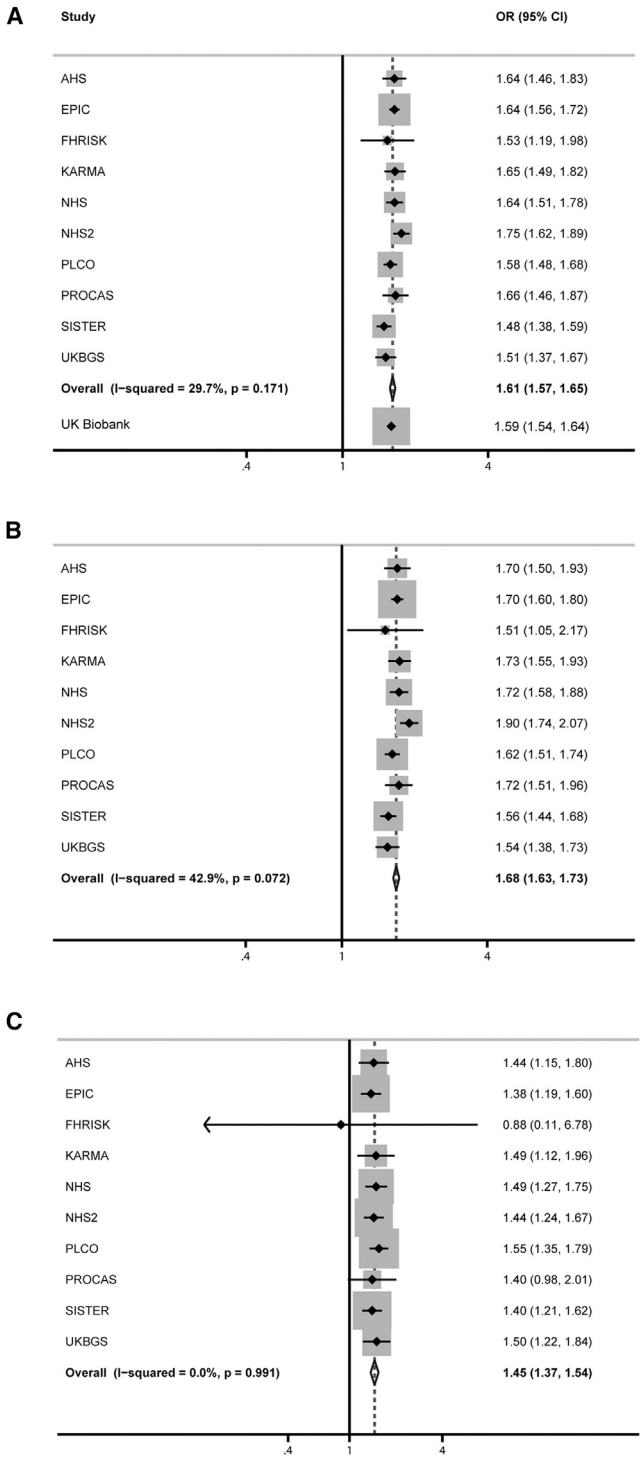

Figure 2. Prospective Validation for the 313 SNP Polygenic Risk Score

Prospective validation for the 313 SNP polygenic risk score (PRS) by study for (A) overall breast cancer, (B) ER-positive disease, and (C) ER-negative disease. Association between the 313 SNP PRS and breast cancer risk in women of European origin. Odds ratios and 95% confidence intervals are shown. I-squared and p value for heterogeneity were calculated using fixed effect meta-analysis.

Discussion

We report development and independent validation of polygenic risk scores for breast cancer, optimized for prediction of subtype-specific disease and based on the largest available GWAS dataset. The best PRS based on a hard thresholding approach included 313 SNPs and was significantly more predictive of risk than the previously reported 77-SNP PRS[7] (OR per 1 SD in the prospective test set: 1.61 versus 1.46; Table 2). The effect sizes were remarkably

Table 3. Associations between the 313-SNP PRS (PRS_{313}) and Breast Cancer Risk by First-Degree Family History of Breast Cancer in the Combined Validation and Prospective Test Dataset

	ER-Positive Disease		ER-Negative Disease	
Model	**OR[a]**	**95% CI**	**OR[a]**	**95% CI**
Association of PRS and Breast Cancer Risk by Family History				
PRS unadjusted	1.67	1.62–1.72	1.44	1.37–1.54
PRS in women without family history	1.71	1.65–1.78	1.45	1.36–1.57
PRS in women with family history	1.55	1.48–1.65	1.40	1.27–1.55
Interaction between PRS and family history	0.91	0.85–0.97 (p = 0.004)	0.96	0.85–1.09 (p = 0.53)
Association between Family History and Breast Cancer Risk (Adjusted and Unadjusted for PRS)				
Family history unadjusted for PRS	1.59	1.46–1.72	1.66	1.41–1.95
Family history adjusted for PRS	1.44	1.33–1.57	1.56	1.32–1.83

Association with breast cancer risk was tested for using logistic regression adjusting for study and ten PCs. For these analyses the validation and test datasets were combined. Analyses were restricted to women with known age and family history information. For ER-negative disease, 4,440 women with and 13,132 women without a family history of breast cancer were included in these analyses. For ER-positive disease, 6,787 women with and 17,351 women without a family history of breast cancer were included in these analyses.
[a]OR per 1 SD for the PRS.

consistent among the 10 cohorts in the prospective test set, and also consistent with that in the UK Biobank cohort (HR = 1.59, 95%CI: 1.54–1.64).

Recently, Khera et al.[27] derived a PRS using our publicly available summary statistics based on analysis of the BCAC data.[1] We were able to construct a PRS based on 5,194 of their 5,218 listed SNPs and compared this to our 313-SNP PRS. In our analysis of this PRS in the prospective UK Biobank data, we obtained a HR of 1.49 (95%CI: 1.44–1.54), substantially lower than that for our PRS_{313}. The corresponding AUCs were 0.613 (95%CI: 0.603–0.623) for their 5,194-SNP PRS versus AUC 0.630 (95%CI: 0.620–0.640) for PRS_{313}. Similarly, PRS_{313} performed better than the Khera et al. PRS in a Biobank dataset consisting of 7,113 case subjects diagnosed before entry and 183,536 control subjects (AUC = 0.642 versus AUC = 0.627). Khera et al. report a much higher AUC (0.68), perhaps reflecting the inclusion of predictors other than SNPs in their model (for example age or principal components).

We specifically aimed to improve prediction for ER-negative breast cancer as to date prediction of this more aggressive disease has been poor. SNP selection was based on association with either ER-negative or overall breast cancer, and the optimum subtype-specific PRSs were derived by weighting a subset of SNPs according to subtype-specific effect sizes, with overall breast cancer weights used for the remaining SNPs. These results are consistent with the observation from genome-wide analyses that the heritability of ER-positive and ER-negative disease are partially correlated.[2] The performance of the PRS_{313} in predicting ER-negative disease was considerably improved over the PRS_{77} reported previously (OR = 1.45 versus 1.35). Nevertheless, the prediction is still better for ER-positive than ER-negative disease, reflecting the fact that ER-negative disease is more infrequent and hence the GWAS data are less powerful. The estimated heritability of ER-negative disease is similar to that of overall breast cancer,[1,2] suggesting that more powerful ER-negative PRSs should be achievable with larger sample sizes.

The best PRS developed using lasso was more predictive for ER-positive disease but slightly less predictive for ER-negative disease in the prospective studies. Given the small differences between the models, we focused on PRS_{313} since this should be more straightforward to implement in diagnostic laboratories using next generation sequencing. However, this will change with developing technology, and the cost effectiveness of using a large marker panel should be further investigated.

From a clinical viewpoint, an important consideration is the performance of the PRS in the tails of the distribution. According to the standard polygenic model, under which the effects of variants combine multiplicatively, the relationship between the PRS and the log-OR should be linear. The PRS was well calibrated at different quantiles. Even in this large study, we observed no deviation from this model, and in particular the observed risks in the highest and lowest centile were consistent with the predicted risk. The sample sizes in the extreme tails, however, were still relatively small, particularly for ER-negative disease.

While the AUC may appear modest, the predicted risk differences in the tails of the distribution are large. For the new PRS_{313}, the women in the top 1% of the distribution have a predicted risk that is approximately 4-fold larger than the risk in the middle quintile. The lifetime risk of overall breast cancer in the top centile of the PRSs, based on UK incidence and mortality data, was 32.6%. Women in the top centile would therefore meet the UK NICE definition of high risk (see Web Resources). In the general population, an estimated 3.6%, 12%, 21%, and 35% of all breast cancers would be expected to occur in women in the highest 1%, 5%, 10%, and 20% of the new

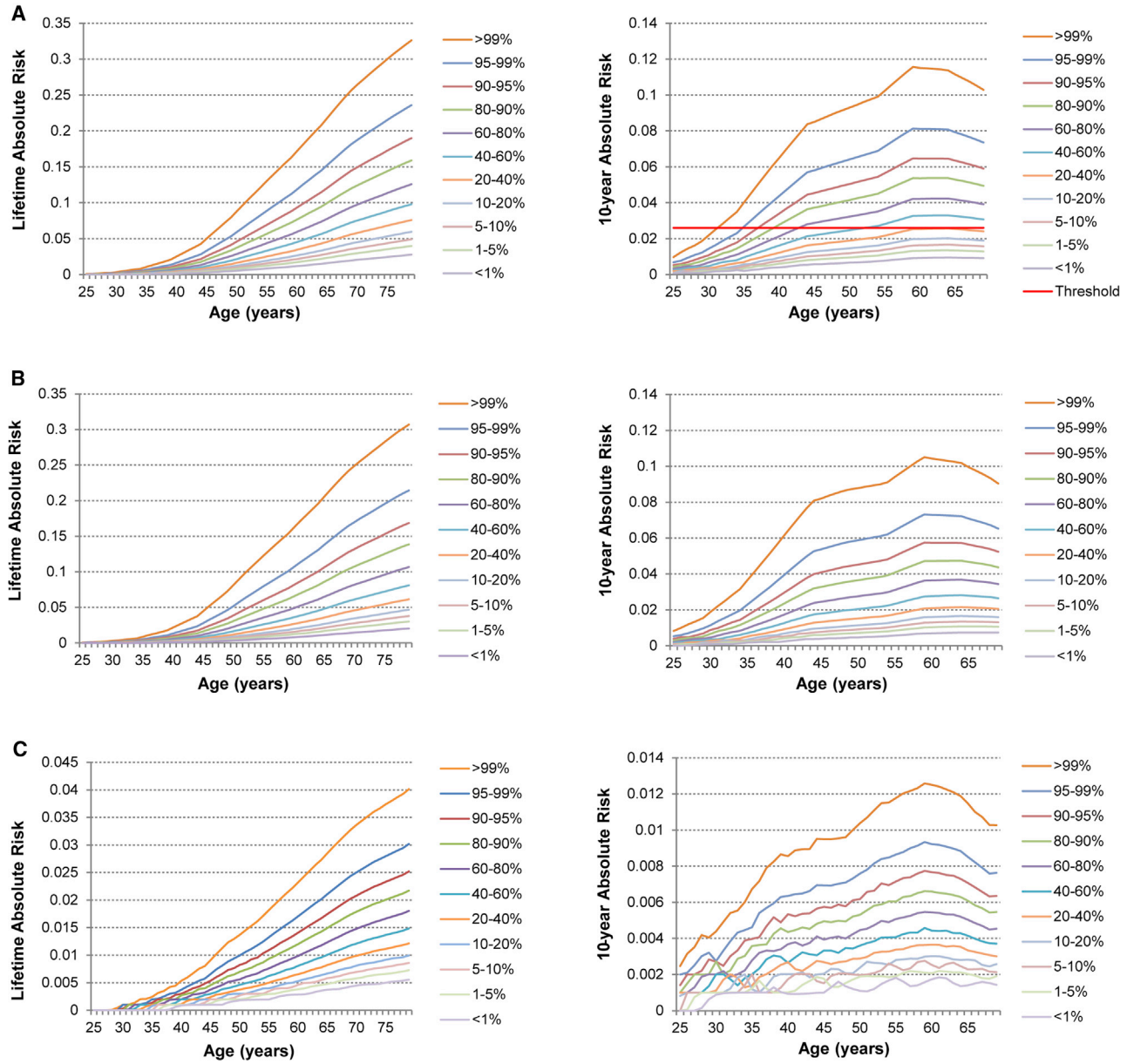

Figure 3. Cumulative and 10-Year Absolute Risk of Developing Breast Cancer
Cumulative and 10-year absolute risk of developing breast cancer for (A) overall breast cancer, (B) ER-positive disease, and (C) ER-negative disease by percentiles of the 313 SNP polygenic risk scores (PRSs). Note different scales and PRS categories in the different panels. The red line shows the 2.6% risk threshold corresponding to the mean risk for women aged 47 years. Absolute risks were calculated based on UK incidence and mortality data and using the PRS relative risks estimated as described in the Material and Methods.

PRS_{313}, respectively, compared to only 9% of breast cancers in women in the lowest 20% of the distribution.

We observed a decline in the relative risk with age for ER-positive disease but not ER-negative disease. Even for ER-positive disease, however, the predicted relative risk, under a linear model, only declined from 1.89 at age 40 to 1.67 at age 70. While there was some indication of a lower relative risk below age 40 (estimated as 1.63 in the test set; Figure S2), these results indicate that PRS_{313} is broadly applicable at all ages. We observed an attenuation of the association between breast cancer family history and breast cancer risk after adjustment for the PRS (~21% for ER-positive, ~12% for ER-negative disease). This finding is broadly in line with the predicted contribution of the PRS to the familial relative risk of breast cancer. The PRS was predictive in women with and without a family history of breast cancer, but the OR was slightly lower in women with a family history, at least for ER-positive disease. This might reflect a weaker relative effect of the PRS in carriers of *BRCA1* or *BRCA2* mutations.[28] We note, however, that the absolute differences in risk by PRS will be larger in women with a family history. These results indicate that the joint effects of family history and PRS need to be considered in risk prediction.

Although we used the largest training dataset available to date for development of the PRS, further improvement should still be possible. We previously estimated using GWAS data that the theoretically best PRS, if the effect sizes of all common SNPs were known with certainty, would explain ~41% of the familial risk of breast cancer, corresponding to a standardized OR~2.1: the PRS_{313} explains ~45% of this "chip" heritability.[1] This implies that larger GWASs, coupled with penalized approaches for subtype-specific disease, should further improve the predictive value of the PRS. Certain genomic features, notably transcription factor binding sites, are enriched among susceptibility loci.[1] Preliminary analyses incorporating these features into the analysis did not improve the predictive value, presumably because the enrichment effect was too small to overcome the increased complexity of the model. Better definition of genomic features to predict causal variants, and more sophisticated methods for integrating external biological information into prediction models, may improve the PRS.[29,30]

The PRS has the potential to improve stratification for screening, while ER-specific PRSs may be informative for prevention with endocrine therapies. Previous studies have suggested that the earlier PRS_{77} was more predictive for screen-detected breast cancers than interval cancers, and that breast cancers arising among women with a low PRS are more aggressive compared with those arising in women with a high PRS, perhaps reflecting the stronger associations with ER-positive disease.[31,32] It will therefore be important to evaluate carefully the associations between the new PRS_{313} and other tumor characteristics. Clinical translational studies are required to assess the risks and benefits of including the PRS in the context of current screening protocols.

While the PRS provides powerful risk discrimination, better risk discrimination will be obtained by combining the PRS with family history and other risk factors.[10] This can be accomplished by incorporating the PRS into risk prediction models, in particular BOADICEA, which can allow for the explicit effects of family history, age, genetic, and other risk factors[33,34] (see Supplemental Material and Methods). However, further studies to validate risk models for individualized risk prediction based on the combined effects of genetic and lifestyle risk factors will be needed. In addition, it is important to note that the PRSs generated in this study were developed and validated in white European populations and need to be validated and potentially adapted for other populations.

Accession Numbers

Requests for access to this dataset should be made to the BCAC coordinator, contact provided in Web Resources.

Supplemental Data

Supplemental Data include 2 figures, 12 tables, Supplemental Acknowledgments, and Supplemental Material and Methods and can be found with this article online at https://doi.org/10.1016/j.ajhg.2018.11.002.

Consortia

ABCTB Investigators are Christine Clarke, Rosemary Balleine, Robert Baxter, Stephen Braye, Jane Carpenter, Jane Dahlstrom, John Forbes, C. Soon Lee, Deborah Marsh, Adrienne Morey, Nirmala Pathmanathan, Rodney Scott, Peter Simpson, Allan Spigelman, Nicholas Wilcken, Desmond Yip, and Nikolajs Zeps.

kConFab/AOCS Investigators are Adrienne Sexton, Alex Dobrovic, Alice Christian, Alison Trainer, Allan Spigelman, Andrew Fellows, Andrew Shelling, Anna De Fazio, Anneke Blackburn, Ashley Crook, Bettina Meiser, Briony Patterson, Christine Clarke, Christobel Saunders, Clare Hunt, Clare Scott, David Amor, David Gallego Ortega, Deb Marsh, Edward Edkins, Elizabeth Salisbury, Eric Haan, Finlay Macrea, Gelareh Farshid, Geoff Lindeman, Georgia Trench, Graham Mann, Graham Giles, Grantley Gill, Heather Thorne, Ian Campbell, Ian Hickie, Liz Caldon, Ingrid Winship, James Cui, James Flanagan, James Kollias, Jane Visvader, Jennifer Stone, Jessica Taylor, Jo Burke, Jodi Saunus, John Forbes, John Hopper, Jonathan Beesley, Judy Kirk, Juliet French, Kathy Tucker, Kathy Wu, Kelly Phillips, Laura Forrest, Lara Lipton, Leslie Andrews, Lizz Lobb, Logan Walker, Maira Kentwell, Mandy Spurdle, Margaret Cummings, Margaret Gleeson, Marion Harris, Mark Jenkins, Mary Anne Young, Martin Delatycki, Mathew Wallis, Matthew Burgess, Melissa Brown, Melissa Southey, Michael Bogwitz, Michael Field, Michael Friedlander, Michael Gattas, Mona Saleh, Morteza Aghmesheh, Nick Hayward, Nick Pachter, Paul Cohen, Pascal Duijf, Paul James, Pete Simpson, Peter Fong, Phyllis Butow, Rachael Williams, Rick Kefford, Rodney Scott, Roger Milne, Rosemary Balleine, Sarah-Jane Dawson, Sheau Lok, Shona O'Connell, Sian Greening, Sophie Nightingale, Stacey Edwards, Stephen Fox, Sue-Anne McLachlan, Sunil Lakhani, Tracy Dudding, and Yoland Antill.

NBCS collaborators are Kristine K. Sahlberg, Lars Ottestad, Rolf Kåresen, Ellen Schlichting, Marit Muri Holmen, Toril Sauer, Vilde Haakensen, Olav Engebråten, Bjørn Naume, Alexander Fosså, Cecile E. Kiserud, Kristin V. Reinertsen, Åslaug Helland, Margit Riis, Jürgen Geisler, and OSBREAC.

Acknowledgments

BCAC was funded by Cancer Research UK (C1287/A16563) and by the European Community's Seventh Framework Programme under grant agreement no. 223175 (HEALTH-F2-2009-223175) (COGS) and by the European Union's Horizon 2020 Research and Innovation Programme under grant agreements 633784 (B-CAST) and 634935 (BRIDGES). Genotyping of the OncoArray was principally funded by Government of Canada through Genome Canada and the Canadian Institutes of Health Research (grant GPH-129344), the Ministère de l'Économie, de la Science et de l'Innovation du Québec through Genome Québec, the Quebec Breast Cancer Foundation; NIH grants U19 CA148065 and X01HG007492; and Cancer Research UK (C1287/A10118 and C1287/A16563). Genotyping of the iCOGS array was funded by the European Union (HEALTH-F2-2009-223175), Cancer Research UK (C1287/A10710), the Canadian Institutes of Health Research for the "CIHR Team in Familial Risks of Breast Cancer" program, and the Ministry of Economic Development, Innovation and Export Trade of Quebec (grant # PSR-SIIRI-701). Combining the GWAS data was supported in part by the National Institutes

of Health (NIH) Cancer Post-Cancer GWAS initiative grant: No. 1 U19 CA 148065 (DRIVE, part of the GAME-ON initiative). We thank all the individuals who took part in these studies and all researchers, clinicians, technicians, and administrative staff who enabled this work to be carried out. For other acknowledgments and sources of funding, see Supplemental Acknowledgments.

Declaration of Interests

D.G.E. reports grants from AstraZeneca and AmGen, outside the submitted work; U.M. has stock ownership and has received research funding from Abcodia Pvt Ltd.; A. Smeets reports other from MSD, outside of the submitted work; P.A.F. reports grants and personal fees from Novartis and personal fees from Pfizer, Roche, Teva, and Celgene, outside the submitted work; R.C. declares personal fees from Novartis, AstraZeneca, and Genentech, outside the submitted work. B.R. reports funding for the conduct of the clinical Success trial paid to her institution from AstraZeneca, Chugai, Lilly, Novartis, Veridex (now Janssen Diagnostics), and Sanofi Aventis. M. Robson reports grants, personal fees, and non-financial support from AstraZeneca, personal fees from McKesson, grants and personal fees from Pfizer, non-financial support from Myriad, non-financial support from Invitae, and grants from AbbVie, Tesaro, and Medivation, outside the submitted work; and M.P.L. reports personal fees from Novartis, Pfizer, Roche, Teva, AstraZeneca, Lilly, and Eisai, outside the submitted work.

Received: August 9, 2018
Accepted: November 3, 2018
Published: December 13, 2018

Web Resources

BCAC data access, http://bcac.ccge.medschl.cam.ac.uk
BCAC Summary statistics, http://bcac.ccge.medschl.cam.ac.uk/bcacdata/oncoarray/gwas-icogs-and-oncoarray-summary-results/
CORDIS, https://cordis.europa.eu/project/rcn/212694_en.html
GenomeCanada 2018 projects, https://www.genomecanada.ca/sites/default/files/2017lsarp_backgrounder_en.pdf
NICE, familial breast cancer clinical guidelines (accessed June 4, 2018), http://guidance.nice.org.uk/CG164
Nomis (26 March 2018), https://www.nomisweb.co.uk/
Office of National Statistics, https://www.ons.gov.uk/
West Midlands Cancer Intelligence Unit, http://www.wmciu.nhs.uk/

References

1. Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A., et al.; NBCS Collaborators; ABCTB Investigators; and ConFab/AOCS Investigators (2017). Association analysis identifies 65 new breast cancer risk loci. Nature *551*, 92–94.
2. Milne, R.L., Kuchenbaecker, K.B., Michailidou, K., Beesley, J., Kar, S., Lindström, S., Hui, S., Lemaçon, A., Soucy, P., Dennis, J., et al.; ABCTB Investigators; EMBRACE; GEMO Study Collaborators; HEBON; kConFab/AOCS Investigators; and NBSC Collaborators (2017). Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nat. Genet. *49*, 1767–1778.
3. Pashayan, N., Duffy, S.W., Chowdhury, S., Dent, T., Burton, H., Neal, D.E., Easton, D.F., Eeles, R., and Pharoah, P. (2011). Polygenic susceptibility to prostate and breast cancer: implications for personalised screening. Br. J. Cancer *104*, 1656–1663.
4. Hall, P., and Easton, D. (2013). Breast cancer screening: time to target women at risk. Br. J. Cancer *108*, 2202–2204.
5. Burton, H., Chowdhury, S., Dent, T., Hall, A., Pashayan, N., and Pharoah, P. (2013). Public health implications from COGS and potential for risk stratification and screening. Nat. Genet. *45*, 349–351.
6. Torkamani, A., Wineinger, N.E., and Topol, E.J. (2018). The personal and clinical utility of polygenic risk scores. Nat. Rev. Genet. *19*, 581–590.
7. Mavaddat, N., Pharoah, P.D., Michailidou, K., Tyrer, J., Brook, M.N., Bolla, M.K., Wang, Q., Dennis, J., Dunning, A.M., Shah, M., et al. (2015). Prediction of breast cancer risk based on profiling with common genetic variants. J. Natl. Cancer Inst. *107*, djv036. https://doi.org/10.1093/jnci/djv036.
8. Maas, P., Barrdahl, M., Joshi, A.D., Auer, P.L., Gaudet, M.M., Milne, R.L., Schumacher, F.R., Anderson, W.F., Check, D., Chattopadhyay, S., et al. (2016). Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States. JAMA Oncol. *2*, 1295–1302.
9. Garcia-Closas, M., Gunsoy, N.B., and Chatterjee, N. (2014). Combined associations of genetic and environmental risk factors: implications for prevention of breast cancer. J. Natl. Cancer Inst. *106*, dju305. https://doi.org/10.1093/jnci/dju305.
10. Rudolph, A., Song, M., Brook, M.N., Milne, R.L., Mavaddat, N., Michailidou, K., Bolla, M.K., Wang, Q., Dennis, J., Wilcox, A.N., et al. (2018). Joint associations of a polygenic risk score and environmental risk factors for breast cancer in the Breast Cancer Association Consortium. Int. J. Epidemiol. *47*, 526–536.
11. Evans, D.G., Astley, S., Stavrinos, P., Harkness, E., Donnelly, L.S., Dawe, S., Jacob, I., Harvie, M., Cuzick, J., Brentnall, A., et al. (2016). Improvement in risk prediction, early detection and prevention of breast cancer in the NHS Breast Screening Programme and family history clinics: a dual cohort study. Southampton UK: NIHR Journals Library (Programme Grants for Applied Research, No. 4.11.), https://www.ncbi.nlm.nih.gov/books/NBK379488/doi:10.3310/pgfar04110.
12. Shieh, Y., Eklund, M., Madlensky, L., Sawyer, S.D., Thompson, C.K., Stover Fiscalini, A., Ziv, E., Van't Veer, L.J., Esserman, L.J., Tice, J.A.; and Athena Breast Health Network Investigators (2017). Breast cancer screening in the precision medicine era: risk-based screening in a population-based trial. J. Natl. Cancer Inst. *109*. https://doi.org/10.1093/jnci/djw290.
13. Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M.J., Maranian, M.J., Bolla, M.K., Wang, Q., Shah, M., et al.; BOCS; kConFab Investigators; AOCS Group; NBCS; and GENICA Network (2015). Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat. Genet. *47*, 373–380.
14. Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R.L., Schmidt, M.K., Chang-Claude, J., Bojesen, S.E., Bolla, M.K., et al.; Breast and Ovarian Cancer Susceptibility Collaboration; Hereditary Breast and Ovarian Cancer Research Group Netherlands (HEBON); kConFab Investigators; Australian Ovarian Cancer Study Group; and GENICA (Gene Environment Interaction and Breast Cancer in Germany) Network (2013). Large-scale genotyping identifies

41 new loci associated with breast cancer risk. Nat. Genet. *45*, 353–361, e1–e2.

15. Amos, C.I., Dennis, J., Wang, Z., Byun, J., Schumacher, F.R., Gayther, S.A., Casey, G., Hunter, D.J., Sellers, T.A., Gruber, S.B., et al. (2017). The OncoArray Consortium: a network for understanding the genetic architecture of common cancers. Cancer Epidemiol. Biomarkers Prev. *26*, 126–135.
16. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
17. O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J.E., Rudan, I., et al. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. *10*, e1004234.
18. O'Connell, J., Sharp, K., Shrine, N., Wain, L., Hall, I., Tobin, M., Zagury, J.F., Delaneau, O., and Marchini, J. (2016). Haplotype estimation for biobank-scale data sets. Nat. Genet. *48*, 817–820.
19. Milne, R.L., Herranz, J., Michailidou, K., Dennis, J., Tyrer, J.P., Zamora, M.P., Arias-Perez, J.I., González-Neira, A., Pita, G., Alonso, M.R., et al.; kConFab Investigators; Australian Ovarian Cancer Study Group; GENICA Network; and TNBCC (2014). A large-scale assessment of two-way SNP interactions in breast cancer susceptibility using 46,450 cases and 42,461 controls from the breast cancer association consortium. Hum. Mol. Genet. *23*, 1934–1946.
20. Joshi, A.D., Lindström, S., Hüsing, A., Barrdahl, M., VanderWeele, T.J., Campa, D., Canzian, F., Gaudet, M.M., Figueroa, J.D., Baglietto, L., et al.; Breast and Prostate Cancer Cohort Consortium (BPC3) (2014). Additive interactions between susceptibility single-nucleotide polymorphisms identified in genome-wide association studies and breast cancer risk factors in the Breast and Prostate Cancer Cohort Consortium. Am. J. Epidemiol. *180*, 1018–1027.
21. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. *33*, 1–22.
22. Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B *58*, 267–288.
23. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190–2191.
24. French, J.D., Ghoussaini, M., Edwards, S.L., Meyer, K.B., Michailidou, K., Ahmed, S., Khan, S., Maranian, M.J., O'Reilly, M., Hillman, K.M., et al.; GENICA Network; and kConFab Investigators (2013). Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. Am. J. Hum. Genet. *92*, 489–503.
25. Meyer, K.B., O'Reilly, M., Michailidou, K., Carlebur, S., Edwards, S.L., French, J.D., Prathalingham, R., Dennis, J., Bolla, M.K., Wang, Q., et al.; GENICA Network; kConFab Investigators; and Australian Ovarian Cancer Study Group (2013). Fine-scale mapping of the FGFR2 breast cancer risk locus: putative functional variants differentially bind FOXA1 and E2F1. Am. J. Hum. Genet. *93*, 1046–1060.
26. Song, M., Kraft, P., Joshi, A.D., Barrdahl, M., and Chatterjee, N. (2015). Testing calibration of risk models at extremes of disease risk. Biostatistics *16*, 143–154.
27. Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. *50*, 1219–1224.
28. Kuchenbaecker, K.B., McGuffog, L., Barrowdale, D., Lee, A., Soucy, P., Dennis, J., Domchek, S.M., Robson, M., Spurdle, A.B., Ramus, S.J., et al. (2017). Evaluation of polygenic risk scores for breast and ovarian cancer risk prediction in BRCA1 and BRCA2 mutation carriers. J. Natl. Cancer Inst. *109*. https://doi.org/10.1093/jnci/djw302.
29. Shi, J., Park, J.H., Duan, J., Berndt, S.T., Moy, W., Yu, K., Song, L., Wheeler, W., Hua, X., Silverman, D., et al.; MGS (Molecular Genetics of Schizophrenia) GWAS Consortium; GECCO (The Genetics and Epidemiology of Colorectal Cancer Consortium); GAME-ON/TRICL (Transdisciplinary Research in Cancer of the Lung) GWAS Consortium; PRACTICAL (PRostate cancer AssoCiation group To Investigate Cancer Associated aLterations) Consortium; PanScan Consortium; and GAME-ON/ELLIPSE Consortium (2016). Winner's curse correction and variable thresholding improve performance of polygenic risk modeling based on genome-wide association study summary-level data. PLoS Genet. *12*, e1006493.
30. Pereira, M., Thompson, J.R., Weichenberger, C.X., Thomas, D.C., and Minelli, C. (2017). Inclusion of biological knowledge in a Bayesian shrinkage model for joint estimation of SNP effects. Genet. Epidemiol. *41*, 320–331.
31. Holm, J., Li, J., Darabi, H., Eklund, M., Eriksson, M., Humphreys, K., Hall, P., and Czene, K. (2016). Associations of breast cancer risk prediction tools with tumor characteristics and metastasis. J. Clin. Oncol. *34*, 251–258.
32. Li, J., Holm, J., Bergh, J., Eriksson, M., Darabi, H., Lindström, L.S., Törnberg, S., Hall, P., and Czene, K. (2015). Breast cancer genetic risk profile is differentially associated with interval and screen-detected breast cancers. Ann. Oncol. *26*, 517–522.
33. Lee, A.J., Cunningham, A.P., Kuchenbaecker, K.B., Mavaddat, N., Easton, D.F., Antoniou, A.C.; Consortium of Investigators of Modifiers of BRCA1/2; and Breast Cancer Association Consortium (2014). BOADICEA breast cancer risk prediction model: updates to cancer incidences, tumour pathology and web interface. Br. J. Cancer *110*, 535–545.
34. Macinnis, R.J., Antoniou, A.C., Eeles, R.A., Severi, G., Al Olama, A.A., McGuffog, L., Kote-Jarai, Z., Guy, M., O'Brien, L.T., Hall, A.L., et al. (2011). A risk prediction algorithm based on family history and common genetic variants: application to prostate cancer with potential clinical impact. Genet. Epidemiol. *35*, 549–556.

The American Journal of Human Genetics, Volume *104*

Supplemental Data

Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes

Nasim Mavaddat, Kyriaki Michailidou, Joe Dennis, Michael Lush, Laura Fachal, Andrew Lee, Jonathan P. Tyrer, Ting-Huei Chen, Qin Wang, Manjeet K. Bolla, Xin Yang, Muriel A. Adank, Thomas Ahearn, Kristiina Aittomäki, Jamie Allen, Irene L. Andrulis, Hoda Anton-Culver, Natalia N. Antonenkova, Volker Arndt, Kristan J. Aronson, Paul L. Auer, Päivi Auvinen, Myrto Barrdahl, Laura E. Beane Freeman, Matthias W. Beckmann, Sabine Behrens, Javier Benitez, Marina Bermisheva, Leslie Bernstein, Carl Blomqvist, Natalia V. Bogdanova, Stig E. Bojesen, Bernardo Bonanni, Anne-Lise Børresen-Dale, Hiltrud Brauch, Michael Bremer, Hermann Brenner, Adam Brentnall, Ian W. Brock, Angela Brooks-Wilson, Sara Y. Brucker, Thomas Brüning, Barbara Burwinkel, Daniele Campa, Brian D. Carter, Jose E. Castelao, Stephen J. Chanock, Rowan Chlebowski, Hans Christiansen, Christine L. Clarke, J. Margriet Collée, Emilie Cordina-Duverger, Sten Cornelissen, Fergus J. Couch, Angela Cox, Simon S. Cross, Kamila Czene, Mary B. Daly, Peter Devilee, Thilo Dörk, Isabel dos-Santos-Silva, Martine Dumont, Lorraine Durcan, Miriam Dwek, Diana M. Eccles, Arif B. Ekici, A. Heather Eliassen, Carolina Ellberg, Christoph Engel, Mikael Eriksson, D. Gareth Evans, Peter A. Fasching, Jonine Figueroa, Olivia Fletcher, Henrik Flyger, Asta Försti, Lin Fritschi, Marike Gabrielson, Manuela Gago-Dominguez, Susan M. Gapstur, José A. García-Sáenz, Mia M. Gaudet, Vassilios Georgoulias, Graham G. Giles, Irina R. Gilyazova, Gord Glendon, Mark S. Goldberg, David E. Goldgar, Anna González-Neira, Grethe I. Grenaker Alnæs, Mervi Grip, Jacek Gronwald, Anne Grundy, Pascal Guénel, Lothar Haeberle, Eric Hahnen, Christopher A. Haiman, Niclas Håkansson, Ute Hamann, Susan E. Hankinson, Elaine F. Harkness, Steven N. Hart, Wei He, Alexander Hein, Jane Heyworth, Peter Hillemanns, Antoinette Hollestelle, Maartje J. Hooning, Robert N. Hoover, John L. Hopper, Anthony Howell, Guanmengqian Huang, Keith Humphreys, David J. Hunter, Milena Jakimovska, Anna Jakubowska, Wolfgang Janni, Esther M. John, Nichola Johnson, Michael E. Jones, Arja Jukkola-Vuorinen, Audrey Jung, Rudolf Kaaks, Katarzyna Kaczmarek, Vesa Kataja, Renske Keeman, Michael J. Kerin, Elza Khusnutdinova, Johanna I. Kiiski, Julia A. Knight, Yon-Dschun Ko, Veli-Matti Kosma, Stella Koutros, Vessela N. Kristensen, Ute Krüger, Tabea Kühl, Diether Lambrechts, Loic Le Marchand, Eunjung Lee, Flavio Lejbkowicz, Jenna Lilyquist, Annika Lindblom, Sara Lindström, Jolanta Lissowska, Wing-Yee Lo, Sibylle Loibl, Jirong Long, Jan Lubiński, Michael P. Lux, Robert J. MacInnis, Tom Maishman, Enes Makalic, Ivana Maleva Kostovska, Arto Mannermaa, Siranoush Manoukian, Sara Margolin, John W.M. Martens, Maria Elena Martinez, Dimitrios Mavroudis, Catriona McLean, Alfons Meindl, Usha Menon, Pooja Middha, Nicola Miller, Fernando Moreno, Anna Marie Mulligan, Claire Mulot, Victor M. Muñoz-Garzon, Susan L. Neuhausen, Heli Nevanlinna, Patrick Neven, William G. Newman, Sune F. Nielsen, Børge G. Nordestgaard, Aaron Norman, Kenneth Offit, Janet E. Olson, Håkan Olsson, Nick Orr, V. Shane Pankratz, Tjoung-Won Park-Simon, Jose I.A. Perez, Clara Pérez-Barrios, Paolo

Supplemental Figures and Legends

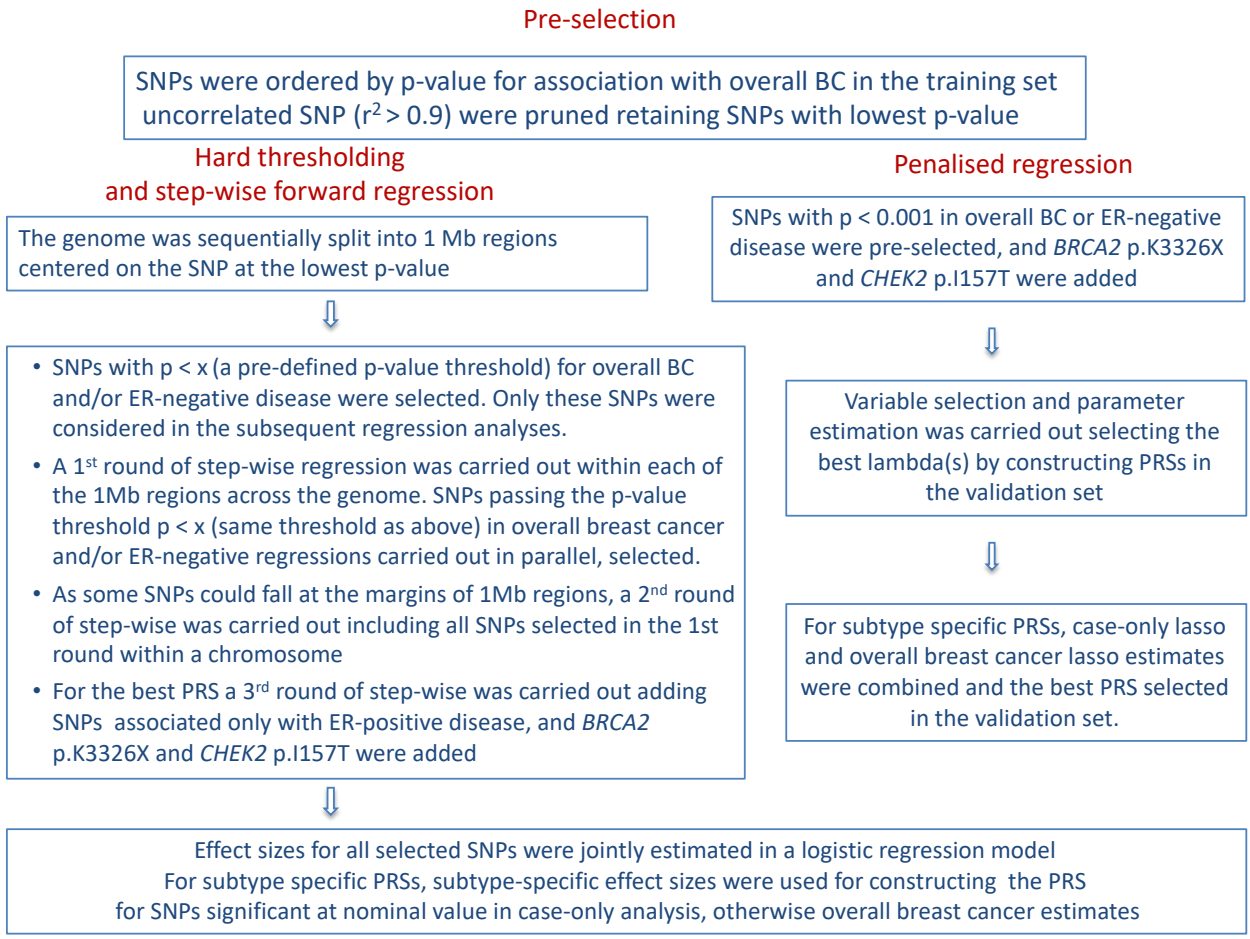

Figure S1. Schema for development of polygenic risk scores

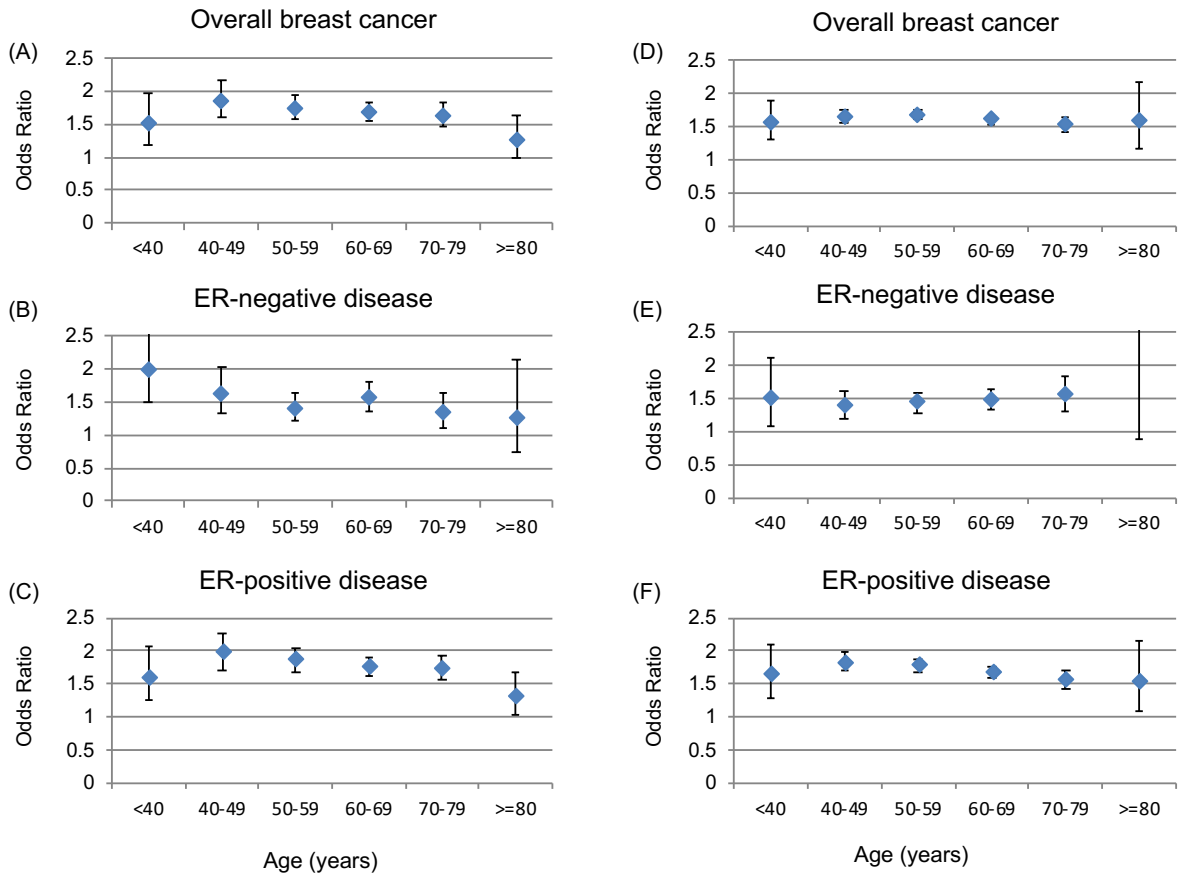

Figure S2. Association between the PRS and breast cancer risk in different age categories. Association between the PRS as a continuous variable and breast cancer risk in different age categories (age at diagnosis/age at interview) of the validation (A-C) and prospective test (D-F) datasets. (A) and (D) overall breast cancer; (B) and (E), ER-negative disease; (C) and (F) ER-positive disease. Odds ratios and 95% confidence intervals are shown

Supplemental Tables

Table S1. Studies and samples in the training set

Table S2. Studies and samples in the validation set

Study	Controls	Invasive cases	ER-positive	ER-negative	Total
ABCTB	74	188	110	78	262
BBCC	49	68	62	6	117
BCEES	166	133	107	26	299
BCINIS	144	262	213	49	406
BREOGAN	145	242	188	54	387
CBCS	163	110	90	20	273
CCGP	66	132	97	35	198
CGPS	142	230	199	31	372
CPSII	605	375	360	15	980
CTS	115	220	196	24	335
GENICA	56	91	75	16	147
HABCS	173	147	128	19	320
LMBC	87	156	132	24	243
MCBCS	35	99	86	13	134
MCCS	142	86	74	12	228
MISS	304	83	65	18	387
MMHS	320	52	44	8	372
NBHS	122	79	44	35	201
ORIGO	132	149	110	39	281
PBCS	331	217	110	107	548
PKARMA	602	195	157	38	797
SEARCH	197	628	550	78	825
SMC	141	244	205	39	385
UCIBCS	51	68	53	15	119
WHI	923	905	778	127	1,828
Total	5,285	5,159	4,233	926	10,444

Table S3. Studies and samples in the prospective test set

Study	Controls	Invasive Cases	ER-positive	ER-negative	Unknown ER status
AHS	1,137	513	377	91	45
EPIC	3,644	3,435	2,004	181	1,250
FHRISK	296	102	43	6	53
KARMA	3,019	451	391	49	11
NHS	1,804	1,103	827	167	109
NHS2	1,905	1,112	868	190	54
PLCO	2,595	1,820	1,371	220	229
PROCAS	1,656	342	304	31	7
SISTER	1,562	1,502	1,205	214	83
UKBGS	705	1,048	602	110	336
Total	18,323	11,428	7,992	1,259	2,177

Table S4. Study design for studies in the prospective test set

Study	Country	Cohort and case control definition	Participation rates	Age (cases)	Selected familial cases
AHS [14,15]	USA	Pesticide (57,310) applicators and their spouses (n=32,345) enrolled during 1993-1997 in Iowa and North Carolina. Cases are women with incident breast cancer diagnosed 1993-2012 in North Carolina and 2013 in Iowa with no previous history of any cancer. Controls are frequency matched to cases on age (5-year age groups), race, state of residence, participant type (applicator or spouse) with no personal history of any cancer.	75% of married spouses completed enrolment questionnaire, 60% of female participants completed the Female and Family Health questionnaire. 46% of Spouses provided buccal cells, ~40% of applicators provided buccal cells	31-89	No
EPIC [16]	Denmark, France, Germany, Greece, Italy, Norway, Spain, Sweden, The Netherlands, UK	Recruitment via 23 research centres in 10 European countries 1992-2000. Cases are women diagnosed with invasive breast cancer after baseline. Controls matched to cases were free from disease up to matching date; incidence density sampling.	Participation rates varied across countries	29- 88	No
FHRISK [17]	UK	Women attending the Family History Clinic in Manchester for increased risk of breast cancer 2009-2012. Cases are women diagnosed with breast cancer and attending the clinic for increased risk of breast cancer. Controls are women attending the same clinic as the cases but without a BC diagnosis. Recruitment period is the same as for cases.	80% for cancers ~70% for controls	29-69	Yes
KARMA	Sweden	Women attending breast screening in Stockholm area. 70,877 women recruited 2010-2013. Cases are incident breast cancers in cohort. Controls are non-BC cases in cohort	All incident cases are included	34-77	No
NHS [18]	USA	Sub-cohort of NHS (32,826) who gave a blood sample 1989-1990 Cases diagnosed with breast cancer prior to July 1, 2000. Controls were women in this sub-cohort who were not diagnosed with breast cancer. Controls were matched to cases on age, postmenopausal status and postmenopausal hormone use.	All incident cases and selected controls are included. All cases and controls had provided blood sample.	34-89	No
NHS2 [19]	USA	Subcohort of NHS2 (n=29,611) who gave a blood sample 1993-1995. Cases are incident cancers arising in the sub-cohort Controls are women in this sub-cohort who were not diagnosed with breast cancer. Controls are matched to cases on age, postmenopausal status and postmenopausal hormone use.	All incident cases and selected controls are included. All cases and controls provided blood sample.	26-63	No
PLCO [20]	USA	Recruitment via mutiple screening centers across the US. Sub-cohort of 78,232 women who gave a blood specimen in 1993-2001. Cases are incident cancers arising in the sub-cohort. Controls are women in this sub-cohort not diagnosed with breast cancer. Controls were matched to cases on age at randomization (4 categories) and fiscal year of randomization (2 categories).	All incident cases and selected controls are included. All cases and controls had provided a blood sample.	55-88	No
PROCAS [17]	UK	Women attending the Breast Screening Programme (NHSBSP) in Greater Manchester. Oct 2009-May 2014.Cases are incident cancers arising in the cohort. Controls are women attending routine NHS breast screening without a breast cancer diagnosis, recruited during the same period as for the cases.	37% uptake to the study. 100% completed questionnaire on risk factors including family history, hormonal factors, lifestyle. 17% provided saliva sample	46-76	No
SISTER [21,22]	USA Puerto Rica	SISTER participants are US/Puerto Rican women (35-74 years) never diagnosed with breast cancer themselves but with a sister diagnosed with breast cancer prior to study start. Recruitment from 2003 to 2009. Cases are participants with incident breast cancer diagnosis after enrolment. Controls are a random selection of SISTER participants.	Questionnaire: 92.2% at last scheduled health follow-up, Blood: 99.4% with blood; additional 0.4% provided saliva sample	36-83	All have a sister with breast cancer
UKBGS [23]	UK	Women from the Breakthrough Generations Study recruited from the UK during 2003-2011). Cases are women who developed breast cancer during follow-up. Controls are women who had not had breast cancer, matched to cases on: age at entry to study (5 year group), year of entry into the study (2005, 2006, 2007, 2008), source of recruitment, blood sample availability and ethnicity.	All selected subjects were recruited from within the cohort study Questionnaire completed for 100%. Questionnaire plus blood sample provided by 92% of the cohort	24-87	No

Table S5: Prediction of subtype-specific disease at different *P*-value thresholds for step-wise regression (validation set)

		ER-positive disease			ER-negative disease		
P-value cutoff [a]	SNPs selected (n)	OR[b]	95% CI	AUC	OR	95% CI	AUC
$< 5 \times 10^{-8}$	123	1.67	1.60 - 1.75	0.640	1.41	1.32 - 1.52	0.596
$< 10^{-6}$	197	1.71	1.64 - 1.79	0.647	1.43	1.33 - 1.53	0.598
$< 10^{-5}$	305	1.74	1.67 - 1.82	0.650	1.45	1.35 - 1.56	0.605
$< 10^{-4}$	669	1.71	1.64 - 1.80	0.645	1.43	1.33 - 1.53	0.598

[a] The *P*-value cutoff refers to the SNPs considered based on their marginal associations in the training set and the same *P*-value threshold for step-wise regression. SNPs were selected using step-wise regression and effect sizes estimated in the relevant subtype. Association with breast cancer risk was tested for using logistic regression adjusting for country and ten PCs. The AUC was adjusted for country. [b] OR per 1 SD for the PRS.

Table S6: Prediction of subtype-specific disease for the 313 SNP PRS (validation set)

	ER-positive disease		ER-negative disease	
Method used for constructing PRS	OR[a]	95% CI	OR	95% CI
Overall Breast cancer SNP effects	1.73	1.65 - 1.81	1.37	1.27 - 1.47
Subtype-specific SNP effects	1.74	1.67 - 1.82	1.45	1.35 - 1.56
Combined overall and subtype-SNP specific effects	1.74	1.67 - 1.82	1.47	1.37 - 1.58
Combined overall and lasso-derived case-only effects	1.75	1.67 – 1.83	1.47	1.37 – 1.58

The 313 SNP PRS was constructed using effect sizes estimated in overall breast cancer, sub-type specific disease, or a combination of overall breast cancer or subtype-specific estimates as described in the Methods. Association with breast cancer risk was tested for using logistic regression adjusting for country and ten PCs. [a]OR per 1 SD for the PRS.

Table S7: SNPs and effect sizes for 313 SNPs used in the construction of overall breast cancer and subtype-specific PRSs

Table S8: SNPs and effect sizes for 3820 SNPs used in the construction of overall breast cancer and subtype-specific PRSs

Table S9. Association between the 313 SNP PRS and overall breast cancer risk in the test set: theoretical and observed odds ratios

	Overall breast cancer			ER-positive disease			ER-negative disease		
	Percentile categories (estimated)		predicted	Percentile categories (estimated)		predicted	Percentile categories (estimated)		predicted
Percentile (%)	OR[a]	95% CI	OR	OR	95% CI	OR	OR	95% CI	OR
<1	0.27	0.18 - 0.40	0.30	0.16	0.09 -0.30	0.27	0.27	0.09 - 0.86	0.39
1-5	0.38	0.31 - 0.46	0.40	0.32	0.25 - 0.40	0.37	0.70	0.47 - 1.05	0.48
5-10	0.56	0.49 - 0.65	0.49	0.50	0.42 - 0.60	0.46	0.74	0.52 - 1.05	0.58
10-20	0.61	0.54 - 0.68	0.59	0.61	0.53 - 0.69	0.56	0.83	0.64 - 1.08	0.66
20-40	0.79	0.73 - 0.86	0.77	0.77	0.70 - 0.85	0.74	0.85	0.69 - 1.04	0.83
40-60	1.00	-	1.00	1.00	-	1.00	1.00	-	1.00
60-80	1.32	1.22 - 1.42	1.29	1.40	1.28 -1.52	1.31	1.32	1.09 - 1.59	1.22
80-90	1.62	1.48 - 1.78	1.65	1.59	1.44 - 1.76	1.73	1.51	1.22 - 1.87	1.48
90-95	1.94	1.74 - 2.17	2.03	2.17	1.93 - 2.44	2.10	2.24	1.76 - 2.85	1.72
95-99	2.47	2.20 - 2.77	2.52	2.68	2.37 - 3.03	2.74	2.39	1.86 - 3.07	2.14
>99	4.04	3.34 - 4.89	3.60	4.37	3.59 - 5.33	4.22	2.78	1.83 - 4.24	2.78

Estimates for women in different percentiles of the PRS in the prospective test dataset were compared with those predicted under a model with the PRS considered as a continuous covariate using the fitted probabilities from the Hosmer-Lomeshaw test. [a] OR per 1 SD for the PRS.

Table S10. Association between the 3,820 SNP PRS and overall breast cancer risk in the test set

	Overall breast cancer		ER-positive disease		ER-negative disease	
	Percentile categories (estimated)		Percentile categories (estimated)		Percentile categories (estimated)	
Percentile (%)	OR[a]	95% CI	OR	95% CI	OR	95% CI
<1	0.18	0.11 - 0.30	0.08	0.04 - 0.19	0.61	0.27 - 1.41
1-5	0.40	0.33 - 0.48	0.30	0.24 - 0.38	0.72	0.48 - 1.08
5-10	0.46	0.39 - 0.54	0.39	0.32 - 0.47	0.81	0.57 - 1.15
10-20	0.67	0.60 - 0.75	0.61	0.54 - 0.69	0.81	0.62 - 1.06
20-40	0.77	0.71 - 0.84	0.68	0.62 - 0.74	1.01	0.83 - 1.24
40-60	1.00	-	1.00	-	1.00	-
60-80	1.40	1.29 - 1.51	1.30	1.20 - 1.42	1.30	1.07 - 1.57
80-90	1.70	1.56 - 1.86	1.60	1.45 - 1.77	1.76	1.43 - 2.18
90-95	2.11	1.89 - 2.35	2.10	1.87 - 2.36	2.18	1.71 - 2.78
95-99	2.68	2.40 - 3.01	2.57	2.27 - 2.90	2.57	2.00 - 3.32
>99	3.95	3.27 - 4.78	4.32	3.55 - 5.26	3.02	1.98 - 4.60

Estimates for women in different percentiles of the PRS in the prospective test dataset. [a] OR per 1 SD for the PRS.

Table S11. The effect of age on the association between the PRS and breast cancer

	ER-positive disease			ER-negative disease		
	OR[a]	95% CI	P	OR[a]	95% CI	P
Combined validation and test set						
*PRS+ age + PRS*age*						
PRS	2.22	1.92 - 2.56	1.02×10^{-26}	1.63	1.28 - 2.06	0.00058
PRS*age	0.996	0.993 - 0.998	0.001	0.998	0.994 - 1.002	0.388
*PRS+age+age²+PRS*age +PRS*age²*						
PRS	1.06	0.62 - 1.81	0.833	1.79	0.83 - 3.85	0.13
PRS*age	1.02	1.004 - 1.04	0.017	0.99	0.97 - 1.09	0.71
PRS*age²	0.9998	0.9996 - 0.9999	0.003	1.00	1.00 - 1.00	0.82
Validation set						
*PRS+ age + PRS*age*						
PRS	2.20	1.74 - 2.78	6×10^{-12}	2.04	1.47 - 2.82	0.000016
PRS*age	0.996	0.99 - 1.00	0.055	0.995	0.989 - 1.00	0.052
*PRS+age+age²+PRS*age +PRS*age²*						
PRS	0.85	0.41 - 1.78	0.667	1.76	0.66 - 4.65	0.26
PRS*age	1.03	1.01 - 1.06	0.013	1.002	0.97 - 1.04	0.93
PRS*age²	0.9997	0.9995 - 0.9999	0.004	1.00	1.00 - 1.00	0.66
Prospective test set						
*PRS+ age + PRS*age*						
PRS	2.25	1.87 - 2.71	9.49×10^{-18}	1.27	0.90 - 1.80	0.18
PRS*age	0.995	0.99 - 0.998	0.003	1.00	0.996 - 1.01	0.44
*PRS+age+age²+PRS*age +PRS*age²*						
PRS	0.97	0.42 - 2.24	0.950	1.13	0.29 - 4.41	0.86
PRS*age	1.03	0.996 - 1.05	0.095	1.01	0.96 - 1.06	0.79
PRS*age²	0.9998	0.9995 - 1.00	0.047	1.00	1.00 - 1.00	0.87

Association with breast cancer risk was tested for using logistic regression adjusting for country and ten PCs (validation set) and study and 15 PCs (test set). For the combined validation and test dataset, analyses were adjusted for study and ten PCs. Analyses were restricted to women with known age at diagnosis/interview. Age (years) was coded as a continuous variable. [a] OR per 1 SD for the PRS.

Table S12. Associations between PRS and breast cancer risk by first-degree family history of breast cancer (validation and test sets separately)

	ER-positive disease		ER-negative disease	
	OR[a]	95% CI	OR[a]	95% CI
Validation set				
Association of PRS and breast cancer risk by family history				
PRS unadjusted	1.76	1.65 - 1.87	1.49	1.35 - 1.65
PRS in women without family history	1.78	1.65 - 1.91	1.55	1.38 - 1.74
PRS in women with family history	1.56	1.36 - 1.78	1.22	0.98 - 1.52
Interaction between PRS and family history	0.89	0.77 - 1.03 (P = 0.109)	0.79	0.62 - 1.00 (P = 0.052)
Association between family history and breast cancer risk (adjusted and unadjusted for PRS)				
Family history unadjusted	1.80	1.55 - 2.09	1.88	1.46 - 2.41
Family history adjusted by the PRS	1.57	1.34 - 1.83	1.74	1.33 - 2.24
Prospective test set				
Association of PRS and breast cancer risk by family history				
PRS unadjusted	1.63	1.58 - 1.69	1.43	1.33 - 1.53
PRS in women without family history	1.67	1.60 - 1.75	1.40	1.27 - 1.53
PRS in women with family history	1.55	1.47 - 1.64	1.45	1.30 - 1.62
Interaction between PRS and family history	0.93	0.87 - 0.997 (P = 0.041)	1.04	0.90 - 1.20 (P = 0.575)
Association between family history and breast cancer risk (adjusted and unadjusted for PRS)				
Family history unadjusted	1.55	1.41 - 1.71	1.56	1.27 - 1.90
Family history adjusted by the PRS	1.43	1.30 - 1.57	1.48	1.21 - 1.81

Association with breast cancer risk was tested for using logistic regression adjusting for country and ten PCs (validation set) and study and 15 PCs (prospective test set). Analyses were restricted to women with known family history information. There was no information on ER-subtype in the family. [a] OR per 1 SD for the PRS.

Supplemental Acknowledgements

We thank in particular: Katerina Kubelka, Christian Baisch, Hans-Peter Fischer, Beate Pesch, Sylvia Rabstein, Anne Lotz, and Volker Harth, Sara Miranda Ponte, Carmen Redondo Marey, José Antúnez, Máximo Fraga and staff of the Department of Pathology and Biobank of the University Hospital Complex of Santiago-CHUS, Instituto de Investigación Sanitaria de Santiago, IDIS, Xerencia de Xestion Integrada de Santiago-SERGAS; Department of Pathology and Biobank of University Hospital Complex of Vigo, Instituto de Investigacion Biomedica (IISGS) Galicia-Sur, Vigo-SERGAS, Spain. Kristine K. Sahlberg, Lars Ottestad, Rolf Kåresen, Dr. Ellen Schlichting, Marit Muri Holmen, Toril Sauer, Vilde Haakensen, Olav Engebråten, Bjørn Naume, Alexander Fosså, Cecile E. Kiserud, Kristin V. Reinertsen, Åslaug Helland, Margit Riis, Jürgen Geisler and OSBREAC, the UKOPS team. The content of this manuscript does not reflect views or policies of the National Cancer Institute, CDC, Mississippi Cancer Registry nor any of the collaborating centers in the Breast Cancer Family Registry (BCFR). Mention of trade names, commercial products, or organizations imply endorsement by the USA Government or the BCFR. We thank participants and staff of the Nurses' Health Study and Nurses' Health Study II and the following state cancer registries: AL, AZ, AR, CA, CO, CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WY. The authors assume full responsibility for analyses and interpretation of these data. This research was conducted using UK Biobank Resource under Application Number 28126. The EU Horizon 2020 Research and Innovation Programme funding source had no role in study design, data collection, data analysis, data interpretation or writing of the report.

Other sources of funding: European Research Council [ERC-2011-294576], European Commission (DG-SANCO, MSCA-IF-2014-EF-656144)), International Agency for Research on Cancer (IARC), National Health and Medical Research Council of Australia (NHMRC) [400413, 400281, 199600, 209057, 251553 and 504711], Victorian Health Promotion Foundation (Australia), Victorian Breast Cancer Research Consortium, National Breast Cancer Foundation (Australia), Queensland Cancer Fund, Cancer Councils of New South Wales, Victoria, Tasmania, South Australia, and Western Australia, Cancer Foundation of Western Australia, Cancer Institute NSW, VicHealth, Australian Institute of Health and Welfare (AIHW), Stichting tegen Kanker (Belgium), FWO (Belgium), Chief Physician Johan Boserup and Lise Boserup Fund (Denmark), Danish Medical Research Council, Herlev and Gentofte Hospital (Copenhagen), Helsinki University Central Hospital Research Fund, Finnish Cancer Society, Sigrid Juselius Foundation (Finland), special Government Funding (EVO) [Kuopio University Hospital, Oulu University Hospital], Cancer Fund of North Savo (Finland), Finnish Cancer Organizations, University of Eastern

Finland, Finnish Cancer Foundation, Academy of Finland [250083, 122715, 251314, 266528], University of Oulu, University of Oulu Support Foundation, Fondation de France, Institut National du Cancer (INCa), Ligue Nationale contre le Cancer, Agence Nationale de Sécurité Sanitaire, de l'Alimentation, de l'Environnement et du Travail (ANSES), Agence Nationale de la Recherche (ANR), Institut Gustave Roussy, Mutuelle Générale de l'Education Nationale (France), Institut National de la Santé et de la Recherche Médicale (INSERM) (France), ELAN-Fond (University Hospital of Erlangen, Germany), Dietmar-Hopp Foundation, Helmholtz Society (Germany), German Cancer Research Center, Heidelberg, Alexander von Humboldt Foundation (Germany), German Cancer Aid [70492, 110837 70-2892-BR I, 106332, 108253, 108419, 110826, 110828], Baden Württemberg Ministry of Science, Research and Arts, Federal Ministry of Education and Research (BMBF) (Germany) [01KW9975/5, 01KW9976/8, 01KW9977/0, 01KW0114, 01KH0402, RUS08/017], Robert Bosch Foundation (Germany), Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), Evangelische Kliniken Bonn gGmbH (Germany), Johanniter Krankenhaus (Germany), Claudia von Schilling Foundation for Breast Cancer Research, Lower Saxonian Cancer Society (Germany), Rudolf Bartling Foundation (Germany), Friends of Hannover Medical School (Germany), German Academic Exchange Program, DAAD, Hamburg Cancer Society, University of Crete, Hellenic Health Foundation (Greece), Stavros Niarchos Foundation (Greece), Associazione Italiana per la Ricerca sul Cancro-AIRC (Italy), Italian citizens "5x1000" to Fondazione IRCCS Istituto Nazionale Tumori, Research Centre for Genetic Engineering and Biotechnology "Georgi D. Efremov" (Macedonia), PBZ_KBN_122/P05/2004 (Poland), Dutch Cancer Society [NKI 2007-3839; 2009 4363, RUL 1997-1505, DDHK 2004-3124, DDHK 2009-4318], Statistics Netherlands (The Netherlands), Biomolecular Resources Research Infrastructure [BBMRI-NL CP16], the K.G. Jebsen Centre for Breast Cancer Research; the Research Council of Norway (grant 193387/V50, 193387/H10), South Eastern Norway Health Authority (39346), the Norwegian Cancer Society. Russian Foundation for Basic Research [14-04-97088,17-29-06014,17-44-020498], Instituto de Salud Carlos III (ISCIII) [JR14/00017], Desarrollo e Innovación Tecnológica de la Consellería de Industria de la Xunta de Galicia, Exp 10CSA012E; Fomento de la Investigación Clínica Independiente, Ministerio de Sanidad, Servicios Sociales e Igualdad (Spain), Exp EC11-192; FEDER-Innterconecta, Ministerio de Economia y Competitividad, Xunta de Galicia (Spain), Exp 00064940; Desarrollo e Innovación Tecnológica de la Consellería de Industria de la Xunta de Galicia, Exp 10CSA012E; Fomento de la Investigación Clínica Independiente, Ministerio de Sanidad, Servicios Sociales e Igualdad (Spain), Exp EC11-192; FEDER-Interconecta, Ministerio de Economia y Competitividad, Xunta de Galicia (Spain), Exp 00064940; Fondo de Investigación Sanitario (FIS)

(Spain) [PI12/02125, PI17/00918, PI16/00440], Regional Governments of Andalucía, Asturias, Basque Country, Murcia and Navarra (Spain), Stockholm County Council, Karolinska Institutet, Swedish Cancer Society, Gustav V Jubilee foundation, Bert von Kantzows foundation, Märit and Hans Rausings Initiative Against Breast Cancer (Sweden), Swedish Research Council, Berta Kamprad Foundation, Gunnar Nilsson, Swedish Cancer Foundation, Cancer Research UK [14136, C490/A16561, C490/A10124, C570/A16491, C1275/A11699, C1275/C22524, C1275/A19187, C1275/A15956, C8221/A19170, C8620/A8372, C12292/A11174, C12292/A20861], Breast Cancer Research Foundation, Breast Cancer Research Trust (UK), Breast Cancer Campaign [2010PR62, 2013PR044], Breast Cancer Now [PR515], Breast Cancer Now Tissue Bank (UK), NHS funding to Institute of Cancer Research NIHR Biomedical Research Centre and NIHR Comprehensive Biomedical Research Centre, Guy's & St. Thomas' NHS Foundation Trust in partnership with King's College London, UK, the Oxford Biomedical Research Centre, Institute of Cancer Research (UK), National Cancer Research Network (NCRN), Against Breast Cancer (UK), Medical Research Council (UK) [1000143 , MR/M012190/1], Sheffield Experimental Cancer Medicine Centre, Health Research Biomedical Research Centre (Cambridge), Eve Appeal (The Oak Foundation), NIHR University College London Hospitals Biomedical Research Centre, the Manchester NIHR Biomedical Research Centre [IS-BRC-1215-20007], NIHR Applied Research programme [RP-PG-0707-10031], Prevent Breast Cancer [GA13-006, GA15-002], US National Cancer Institute (NCI) [CA54281, CA58860, CA63464, CA92044, CA97396, CA098758, CA116167, CA128931, CA132839, CA140286, CA164973, CA176785, CA177150, CA192393, CA194393, D43 TW009112, K24 CA169004, N01CN25403, P01 CA87969, P30 CA68485, P30 CA008748, P41-GM1035, P50 CA125183, R01CA64277, R01 CA77398, R01 CA89085, R01 CA092447, R01 CA100374, R01 CA120120, R01-CA121941, R01 CA148667, R01 CA159868, K05 CA136967, R37CA70867, U01 CA164973,U01 CA199277, U19 CA148065, U41-HG006623, UM1 CA164917, UM1 CA164920, UM1 CA186107, UM1 CA176726, UMCA182910, Z01-CP010119, a Specialized Program of Research Excellence (SPORE) in Breast Cancer P50 CA116201, Intramural Research Funds], Institute of Environmental Health Sciences [Intramural Program, Z01-ES044005, Z01-ES049033 Z01-ES044005, Z01-ES102245, Z01-ES049030], National Heart, Lung, and Blood Institute, US Department of Health and Human Services [HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C and HHSN271201100004C], United States Army Medical Research and Materiel Command [DAMD17-01-1-0729], American Cancer Society, California Breast Cancer Act of 1993, California Breast Cancer Research Fund [97-10500], California Department of Public Health, Lon V Smith Foundation [LVS39420], California Breast Cancer Research Program [1RB-0287, 3PB-0102,

5PB-0018, 10PB-0098], Breast Cancer Research Foundation, David F. and Margaret T. Grohne Family Foundation, Robert and Kate Niehaus Clinical Cancer Genetics Initiative, Susan G. Komen Breast Cancer Foundation [FAS0703856, SAC110026], Dr. Ralph and Marian Falk Medical Research Trust, Avon Foundation for Women, Lon V Smith Foundation [LVS39420]. The US NHLBI supporting the Women's health Initiative program under contract. The Mayo Clinic Cancer Center, the International Agency for Research on Cancer, Ligue Contre le Cancer, Institut Gustave Roussy, Mutuelle Générale de l'Education Nationale, Institut National de la Santé et de la Recherche Médicale (INSERM) (France); German Cancer Aid [70-2892-BR I, 106332, 108253, 108419, 110826, 110828], German Cancer Research Center (DKFZ), Dutch Ministry of Public Health, Welfare and Sports (VWS), Netherlands Cancer Registry (NKR), LK Research Funds, Dutch Prevention Funds, Dutch ZON (Zorg Onderzoek Nederland), World Cancer Research Fund (WCRF), Statistics Netherlands (The Netherlands); Health Research Fund (FIS) [PI13/0006, PI13/01162], Regional Governments of Andalucía, Asturias, Basque Country, Murcia and Navarra, ISCIII RETIC (RD06/0020) (Spain), Medical Research Council (1000143, MR/M012190/1) (United Kingdom), NIHR [PGfAR 0707-10031, PGfAR 0707-10031], NIHR Biomedical Research Centre (University of Cambridge), the NHS in the East of England through the Clinical Academic Reserve, KL2TR002379 from the National Center for Advancing Translational Sciences (NCATS). The contents of this manuscript are solely the responsibility of the authors and do not necessarily represent the official view of NIH. The Canadian Cancer Society (# 313404), the Canadian Institutes of Health Research, Against Breast Cancer Registered Charity No. 1121258, UCLH and the NCRN.

Supplemental Methods

The dataset used for development of the PRS comprised 94,075 cases (women diagnosed with invasive breast cancer as well as breast cancer of unknown invasiveness) and 75,017 controls of European ancestry from 69 studies in the Breast Cancer Association Consortium (BCAC). Samples were genotyped using one of two arrays: iCOGS[1,2] and OncoArray[3,4]. The dataset was divided into a training and validation set. The validation set was randomly selected (approximately 10% of cases and controls genotyped on the OncoArray), from studies that had been genotyped with the OncoArray, after excluding studies of bilateral breast cancer, studies or sub-studies oversampling for family history, and individuals with in-situ cancers or cases with unknown ER-status. The remaining samples were designated as the 'training set'.

Further validation was conducted using a test dataset comprising 11,428 invasive breast cancer cases and 18,323 controls from ten studies nested within prospective cohorts.

The overall breast cancer PRS was also evaluated among 190,040 women of European ancestry from the UK Biobank cohort, who had not had any cancer diagnosis or mastectomy prior to recruitment. 3,215 registry-confirmed invasive breast cancers developed over 1,381,019 person years of follow-up. A Cox proportional hazards regression model was used to assess the association with risk of breast cancer in the UK Biobank. Follow-up started 6 months after age of baseline questionnaire. The primary endpoint was invasive breast cancer. Follow-up was censored at the earliest of: risk-reducing mastectomy, diagnosis of any type of cancer, death, or January 15 2017.

Genotyping calling, quality control and imputation was performed as previously described.[1,2,4] Where samples were genotyped with iCOGS and OncoArray, the OncoArray calling was used. Imputation was performed for the iCOGS and OncoArray datasets separately using the Phase 3 (October 2014) release of the 1000 genomes data as reference [5]. We followed a two-stage approach using SHAPEIT for phasing [6] and IMPUTE2 for the imputation[3]. UK Biobank samples were genotyped using Affymetrix UK BiLEVE Axiom array and Affymetrix UK Biobank Axiom® array and imputed to the combined 1000 Genome Project v3 and UK10K reference panels using SHAPEIT3 and IMPUTE3 [7]. The lowest imputation info score for the SNPs used in these analyses was 0.86. Samples were included for this analysis of the UK BIOBANK study on the basis of female sex (genetic and self-reported) and ethnicity filter (Europeans/White British ancestry subset). Duplicates, individuals with high degree of relatedness (>10 relatives), and one of each related pair first degree relatives were removed. Samples were also excluded on standard quality control criteria.

We used two general approaches for model selection: “hard-thresholding”, based on a stepwise penalized regression model and retaining SNPs significant at a given threshold, and step-wise forward regression and penalized regression using the lasso penalty with an L1 penalty corresponding to a double-exponential prior placed on the regression coefficients [8,9]. Alternative penalized approaches including ridge regression and minimax concave penalty [10] were also evaluated but did not improve performance [data not shown]. To prioritise SNPs for analysis, single SNP association tests were first conducted in the training set. Per-allele ORs and standard errors were estimated separately in the iCOGS and OncoArray datasets, adjusting for study and nine ancestry informative principal components (PCs) in the iCOGS dataset and by country and 10 PCs

in the OncoArray dataset, using a purpose-written program [1]. Combined p-values were then derived using a fixed-effects meta-analysis with the software METAL[11]. SNPs were sorted by *P*-value and filtered on LD, such that uncorrelated SNPs (correlation $r^2 < 0.9$) with lowest *P*-value for association with overall breast cancer in the training set were retained.

In the hard thresholding approach, a series of step-wise forward regression analyses were first carried out in 1 Mb regions centered on SNPs significant at a pre-specified threshold for association with either overall and/or subtype-specific disease in the training-set. Only SNPs passing the specified *P*-value thresholds were included in each 1Mb region. Two analyses were performed in parallel: for overall breast cancer and ER-negative disease. At each stage the SNP with the smallest (conditional) *P*-value for any analysis was added to the model, the threshold for the step-wise regression being the same as that for pre-selection. The process was repeated until no further SNPs could be added at the pre-defined threshold. A second stage of stepwise regressions were then carried out across all regions in each chromosome, to take into account correlated SNPs in different regions. Finally, the effect sizes for the selected SNPs were jointly estimated in a single logistic regression model.

For the best-performing PRS, SNPs associated with ER-positive at *P*-value $< 10^{-6}$ but not with overall breast-cancer (at *P*-value $< 10^{-5}$) were added at the end of the final SNP list. A third round of step-wise forward regression was then carried out with *P*-value for selection of $P < 10^{-6}$ for ER-positive disease. For completeness we added to this final PRS two rarer variants (*BRCA2* p.K3326X (MIM: 600185; NM_000059.3:c.9976A>T; NP_000050.2:p.Lys3326Ter) and *CHEK2* p.I157T (MIM: 604373; NM_007194.3:c.470T>C; NP_009125.1:p.Ile157Ser)) which are established to confer a moderate risk of breast cancer and were genotyped on the OncoArray but did not pass the allele frequency threshold in the PRS development phase. An additional PRS based on SNPs reported as associated in the literature at "genome-wide" significance ($P<5x10^{-8}$) was also constructed. 177 of 178 reported SNPs were included (1 SNP was not present on the 1000 genomes reference panel). Effect sizes for overall breast cancer were taken from publicly available BCAC summary statistics (see Web Resources). In UK Biobank, imputed genotypes were available for 306 SNPs from the 313 SNP PRS (excluding: 22_38583315_AAAAG_AAAAGAAAG, 3_63887449_T_TTG, 4_126752992_A_AAT, 4_187503758_A_T, 4_84370124_TAA_TA, 5_52679539_C_CA, and 7_91459189_A_ATT) and 174 SNPs from the 177 SNP PRS (excluding:

4_84370124_TAA_TA, 9_136151579_TGGTGCAGGCGCAGGAAAAAATTGTGGCAATTCCTCA_T, and 22_39359355_C_.CN0). The PRS tested in UK Biobank used the same weights as in the other prospective studies but with 7 and 3 SNPs fewer, respectively.

We evaluated whether prioritizing SNPs with certain genomic features might improve risk prediction. The set of credible set of causal variants at breast cancer risk loci are enriched for binding sites determined by ChIP-seq, for certain transcription factors (ESR1, FOXA1, GATA3, HA-E2F1, EP300, MYC) in ER-positive breast cancer cell lines (T-47D, ZR-75-1, and MCF-7) [4]. SNPs were assigned as positive for the biofeature if correlated at $r^2 > 0.9$ with another SNP overlapping any of these TFBS. Within each region, the stepwise program was run first on SNPs with the bio-feature. SNPs significant at a certain p-value threshold were selected. Subsequently the step-wise program was run on SNPs without the biofeature and a more stringent *P*-value threshold was used for selection from the remaining SNPs.

For the penalized regression using lasso we used the program *glmnet* [8]. SNPs with $P < 0.001$ in overall BC or ER-negative disease were pre-selected for inclusion in the lasso, and *BRCA2* p.K3326X and *CHEK2* p.I157T were added. Covariates for 19 PCs (9 for iCOGs and 10 for Oncoarray) and country were include in each model. For overall breast cancer, the penalty parameter (lambda) giving the best overall breast cancer PRS in the validation set was selected.

A schema for the analyses is shown in Figure S1.

To evaluate the performance of each potential PRS, we standardized the PRS to have unit standard deviation in the validation set controls. The association of the standardized PRS was evaluated by logistic regression adjusted for country (validation set) or study (test set) and PCs. The standard deviation in the validation set controls was 0.61, 0.65 and 0.59 for the overall breast cancer, ER-positive, and ER-negative 313 SNP PRS respectively. Models were also compared in terms of the area under the receiver operator characteristic curves (AUC), adjusted for study, calculated using the Stata command *comproc.* Bootstrap standard errors and confidence intervals for the AUC were calculated.

For the best performing models, ORs were calculated for quantiles, relative the median (40-60th percentile) quantile. To convert the latter into ORs relative to the population mean, these ORs were divided by the weighted sum of the odds ratios:

$$\Psi = P_j OR_j$$

where P_j is the proportion of the population in bin j.

The modification of the PRS by age or by family history (FH) of breast cancer in a first-degree relative was evaluated by fitting additional interaction terms in the model. Validation and prospective test datasets were combined in order to obtain larger sample size.

Attenuation of the association between family history and breast cancer risk after adjustment for the PRS was calculated as

(logOR(FHunadj)-logOR(FHadj))/logOR(FHunadj)

The absolute risk of breast cancer for individuals in each risk category was calculated as described previously [12], but with effect size for the PRS modeled as a continuous covariate. Age interactions were included assuming a linear effect of age on the PRS breast cancer association. Absolute risks were calculated taking into account the competing risk of dying from other causes apart from breast cancer and using UK based on 2016 mid-year incidence and mortality rates (Office of National Statistics and Nomis, see Web Resources). The absolute risk of developing subtype specific disease was obtained constraining to the incidence of overall incidence of ER-negative and ER-positive disease in the UK (derived from the overall incidence of breast cancer from the UK population data, and the age-specific proportions of ER-negative and ER-positive tumours). Estimates of the age-specific proportions of breast cancer by tumour subtype in the UK population were obtained from the West Midlands Cancer Intelligence Unit (see Web Resources). Women are at risk of developing both ER-negative and ER-positive disease, therefore the absolute risks were calculated given that the individual has been free of breast cancer of any subtype.

For overall breast cancer and for each disease subtype, mean absolute risk for women in different categories of the PRS were calculated. The probability of a woman developing breast cancer by any age t_2, given she is alive and free of breast cancer at age t_1, calculated as: $(AR(t_2)-AR(t_1))/(S_g(t_1)*S_m(t_1))$

Where $S_g(t_1)$ is the probability of being free of breast cancer to age t and $S_m(t_1))$ the probability of surviving to age t, i.e. not dying from a cause other than from breast cancer.

Implementation of the PRS in the risk prediction algorithm BOADICEA

The risk prediction model BOADICEA assumes genetic susceptibility to breast cancer is conferred by a combination of rare major genes and a polygenic components. The PRS can be implemented into

BOADICEA by assuming that a proportion of the polygenic component is known. The overall polygenic variance varies with age, but the proportion explained by the PRS is assumed to be constant with age. This assumption induces a specific form to the age-specific relative risks conferred by the PRS. The proportion can be estimated from this dataset.

The polygenic variance in BOADICEA is assumed to be a linear function of age: $\sigma^2 = \alpha + \beta t$ where t = age (years). The parameters α and β have been previously estimated, using complex segregation analysis, as 4.86 and -0.06 respectively[13]. The variance due to known component is therefore of the form: $\sigma^2_K = \gamma^2 (\alpha + \beta t)$.

σ_K is, alternatively, the hazard ratio per unit SD of the PRS (strictly, the hazard ratio conditional on the unknown polygenic component rather than the marginal hazard ratio). We can therefore estimate the proportionality constant γ by logistic regression, with the PRS covariate replaced by

$$S' = \sqrt{4.86 + (-0.06 * t)} * S_{standardised}$$

Where $S_{standardised}$ is the standardised (unit standard deviation) PRS used in the main analysis.

Using this approach, we estimate γ =0.481 (95%CI:0.442-0.519) using the validation dataset, 0.434 (95%CI:0.411-0.456) using the test dataset and 0.445 (95%CI:0.425-0.464) using the combined dataset. The current PRS_{313} can therefore be implemented in BOADICEA by assuming that the polygenic variance explained by the PRS is γ^2=0.20.

Supplemental References

1. Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M.J., Maranian, M.J., Bolla, M.K., Wang, Q., Shah, M., et al. (2015). Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat. Genet. *47*, 373-380.

2. Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R.L., Schmidt, M.K., Chang-Claude, J., Bojesen, S.E., Bolla, M.K., et al. (2013). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat. Genet. *45*, 353-352.

3. Amos, C.I., Dennis, J., Wang, Z., Byun, J., Schumacher, F.R., Gayther, S.A., Casey, G., Hunter, D.J., Sellers, T.A., Gruber, S.B., et al. (2017). The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. Cancer Epidemiol. Biomarkers Prev. *26*, 126-135.

4. Michailidou, K., Lindstrom, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemacon, A., Soucy, P., Glubb, D., Rostamianfar, A., et al. (2017). Association analysis identifies 65 new breast cancer risk loci. Nature *551*, 92-94.

5. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68-74.

6. O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J.E., Rudan, I., et al. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. 10, e1004234.

7. O'Connell, J., Sharp, K., Shrine, N., Wain, L., Hall, I., Tobin, M., Zagury, J.F., Delaneau, O., and Marchini, J. (2016). Haplotype estimation for biobank-scale data sets. Nat. Genet. *48*, 817-820.

8. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. *33*, 1-22.

9. Tibshirani, R. (1996). Regression Shrinkage and selection via the Lasso. J. R. Stat. Soc. B. *58*, 267-288.

10. Breheny P, Huang J. Coordinate Descent Algorithms for nonconvex penalized regression, with applications to biological feature selection. (2011) Ann. Appl. Stat. *5*, 232-253.

11. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190-2191.

12. Mavaddat, N., Pharoah, P.D., Michailidou, K., Tyrer, J., Brook, M.N., Bolla, M.K., Wang, Q., Dennis, J., Dunning, A.M., Shah, M., et al. (2015). Prediction of breast cancer risk based on profiling with common genetic variants. J. Natl. Cancer Inst. *107*.

13. Antoniou, A.C., Cunningham, A.P., Peto, J., Evans, D.G., Lallo, F., Narod, S.A., Risch, H.A., Eyfjord, J.E., Hopper, J.L., Southey, M.C. et al. (2008) The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. Br. J. Cancer *98*, 1457-1466.

14. Lerro, C.C., Koutros, S., Andreotti, G., Friesen, M.C., Alavanja, M.C., Blair, A., Hoppin, J.A., Sandler, D.P., Lubin, J.H., Ma, X., et al. (2015). Organophosphate insecticide use and cancer incidence among spouses of pesticide applicators in the Agricultural Health Study. Occup. Environ. Med. *72*, 736-744.

15. Koutros S, A.M., Lubin JH, et al. (2010). An Update of Cancer Incidence in the Agricultural Health Study. J. Occup. Environ. Medicine *52*, 1098-1105.

16. Riboli, E., Hunt, K.J., Slimani, N., Ferrari, P., Norat, T., Fahey, M., Charrondiere, U.R., Hemon, B., Casagrande, C., Vignat, J., et al. (2002). European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutr *5*, 1113-1124.

17. Evans, D.G., Astley, S., Stavrinos, P., Harkness, E., Donnelly, L.S., Dawe, S., Jacob, I., Harvie, M., Cuzick, J., Brentnall, A., et al. (2016). Improvement in risk prediction, early detection and prevention of breast cancer in the NHS Breast Screening Programme and family history clinics: a dual cohort study. Southampton UK: NIHR Journals Library (Programme Grants for Applied Research, No. 4.11.) https://www.ncbi.nlm.nih.gov/books/NBK379488/doi: 10.3310/pgfar04110.

18. Hankinson, S.E., Willett, W.C., Manson, J.E., Colditz, G.A., Hunter, D.J., Spiegelman, D., Barbieri, R.L., and Speizer, F.E. (1998). Plasma sex steroid hormone levels and risk of breast cancer in postmenopausal women. J. Natl. Cancer Inst. *90*, 1292-1299.

19. Tworoger, S.S., Missmer, S.A., Eliassen, A.H., Spiegelman, D., Folkerd, E., Dowsett, M., Barbieri, R.L., and Hankinson, S.E. (2006). The association of plasma DHEA and DHEA sulfate with breast cancer risk in predominantly premenopausal women. Cancer Epidemiol. Biomarkers Prev. *15*, 967-971.

20. Pfeiffer, R.M., Park, Y., Kreimer, A.R., Lacey, J.V., Jr., Pee, D., Greenlee, R.T., Buys, S.S., Hollenbeck, A., Rosner, B., Gail, M.H., et al. (2013). Risk prediction for breast, endometrial, and ovarian cancer in white women aged 50 y or older: derivation and validation from population-based cohort studies. PLoS Med. *10*, e1001492.

21. Nichols, H.B., Baird, D.D., DeRoo, L.A., Kissling, G.E., and Sandler, D.P. (2013). Tubal ligation in relation to menopausal symptoms and breast cancer risk. British journal of cancer *109*, 1291-1295.

22. Xu, Z., Bolick, S.C., DeRoo, L.A., Weinberg, C.R., Sandler, D.P., and Taylor, J.A. (2013). Epigenome-wide association study of breast cancer using prospectively collected sister study samples. J. Natl. Cancer Inst. *105*, 694-700.

23. Swerdlow, A.J., Jones, M.E., Schoemaker, M.J., Hemming, J., Thomas, D., Williamson, J., and Ashworth, A. (2011). The Breakthrough Generations Study: design of a long-term UK cohort study to investigate breast cancer aetiology. Br. J. Cancer *105*, 911-917