Development and Validation of a Polygenic Risk Score for Breast Cancer

George Busby PhD, Alessandro Bolli PhD, Paolo Di Domenico, Giordano Botta PhD

INTRODUCTION

Advances in genetic research are leading to improved precision medicine in healthcare

An individual's risk of developing disease results from a complex interaction between their environment and their genes. The collection and analysis of large amounts of data is now allowing us to understand these relationships in more detail than ever before and are aiding the development of data-driven approaches to risk prediction that will transform the way that healthcare is provided [1]. From the electronisation of health records to the introduction of wearable devices, developments in digitisation are allowing the collection of biomedical big data at an unprecedented scale.

At the same time, methods involving advanced analytics like **machine learning** are leading to a greater understanding of these data. The integration of these concepts into healthcare systems is paving the way for a new era of **precision medicine**, the goal of which is to use data to identify those at highest risk of developing disease so that interventions and treatments can be targeted at the groups that need them most. This will ensure that finite healthcare resources are used as efficiently as possible and that disease is caught early enough to improve patient outcomes.

Precision medicine is made increasingly possible because we now have the analytical methods and computational power necessary to understand big, complex datasets. These datasets contain clinical outcomes on a large number of people for a variety of diseases and include additional information on physiological, lifestyle and environmental factors such as age, sex and family history of disease. When these data are combined and analysed, sophisticated statistical models can find patterns in the combinations of risk factors that lead to disease, which can provide an estimate of an individual's risk relative to others in their population with similar values across these factors.

As we understand more about the role that **DNA** has to play in understanding disease risk, it is becoming increasingly important to include genetic or genomic information in these datasets. This is because progress in **genomics** means that this novel and potentially powerful type of data can be added into disease risk prediction models. The use of genomic data in risk prediction is now possible thanks to the confluence of three major scientific advances:

1. Bigger sample sizes in genome-wide association studies (GWAS) which lead to greater statistical power to identify genetic variants involved with disease.
2. Better statistical methods to identify the most predictive set of variants to estimate risk for a given disease.

3. The availability of genome-wide data from hundreds of thousands of people linked with thousands of environmental and physiological measurement in the UK Biobank [2]. This magnificent data resource allowed the validation of the algorithms predictive power at an unprecedented scale.

Developing a robust procedure for estimating a breast cancer PRS

In its essence, computing an individual's polygenic risk score (PRS) for a given disease is straightforward. By multiplying the number of risk alleles a person carries by the effect size of each variant and then summing these across all risk loci [3] one can estimate an individual's genetic liability for a given disease. However, identifying the alleles at loci that confer disease risk, estimating the size of the allele's effect on disease and incorporating estimates of uncertainty in a PRS remaining challenging. Furthermore, the accuracy of a PRS depends on several conditions [4].

The first is that the GWAS providing the summary statistics for the score – known as the **discovery** set – should involve an independent set of samples to those on which the scores are being calculated. Secondly, the amount of variation in disease or trait liability that can be accounted for by the genetic variants used in the PRS, known as SNP **heritability**, will influence how predictive a PRS will be, which will also be affected by the **genetic architecture** of the disease. Finally, the sample size of the discovery GWAS will affect how well effect sizes are estimated and therefore in turn affect its accuracy. The best performing PRSs use summary statistics from a discovery GWAS involving hundreds of thousands of independent individuals on a trait with high SNP heritability.

Identifying which SNPs have the best predictive power is a central challenge to developing a robust PRS. There are two main objectives to this effort. The first is to understand at what **threshold** of statistical significance SNPs should be removed from the score generation algorithm. Because there are correlations between the effect sizes of variants that are close to each other in the genome, the second objective is to explore how to combine evidence across multiple variants.

Fortunately, procedures have been developed to select subsets of SNPs that rely on looking only at **GWAS summary statistics** [5]. The simplest approach, known as **clumping and thresholding** (C+T), iterates between two methodological steps [6]. First, genetic variants are filtered, or clumped, so that only the variants with the highest effect size and that are not in **linkage disequilibrium** (LD) are used. In the second thresholding step, genetic variants with a P value larger than a chosen threshold are removed. This process is repeated for different LD windows and P value thresholds.

A more sophisticated **Bayesian** approach involves modelling LD to shrink each variant effect size to an extent that is proportional to the LD between SNPs [7]. This approach, implemented in the software LDPred, requires the definition of a tuning parameter ρ, which is a statement of the researcher's prior belief on the proportion of genetic variants assumed to be causal.

Recently, a third method involving machine learning that combines C+T and the **LASSO** statistical procedure, called **stacked clumping and thresholding** (SCT) has been developed [8].

In SCT, clumping and thresholding are systematically repeated over a four dimensional grid of parameters (comprising LD squared correlation and p-value thresholds). The algorithm generates over 100,000 alternative C+T variants and combines them through a LASSO-based penalized logistic regression.

Validating and testing PRS is possible with the availability of Biobanks

The algorithms outlined above require a **validation phase** where different PRSs generated with alternative parameter values are validated against an external dataset (Validation dataset). The **test phase** involves computing PRS in a test population (Test dataset) and assessing its **predictive power** in order to confirm its predictive performance and to rule out any possibility of **over-fitting** that may have occurred during the validation step.

Development of PRSs for a number of diseases has been accelerated by the availability of the UK Biobank (UKB) dataset. This is a large **prospective cohort study** that enrolled around 500,000 individuals from across the UK, ranging in age from 40 to 69 years at the time of recruitment and whose genomes have been genotyped and imputed to more than 90 million variants. The astonishing size of the UKB genomic data means that researchers can no build suitably sized Validation and Testing datasets.

The genetic architecture of breast cancer

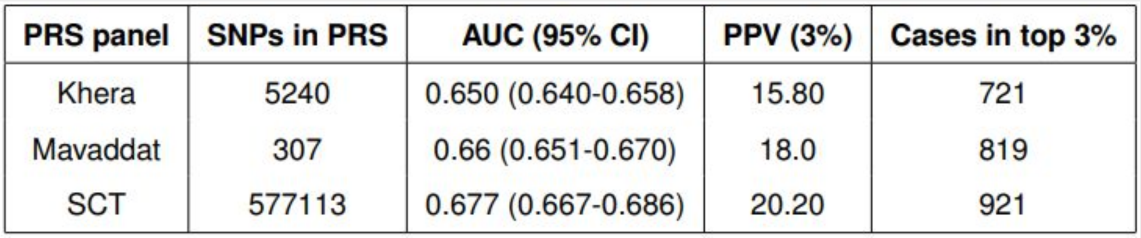
Breast cancer (BC) is the most common cancer diagnosed among women in Western countries. The risk of developing BC is linked to both non-genetic and genetic factors, with the former referring to any circumstance that is not inherited, such as nutrition, environmental toxins, or the use of hormone replacement therapy (HRT).

From a genetic perspective, BC has a complex genetic structure depending on two classes of genetic variants: rare mutations with high **penetrance** such as the BRCA1 and BRCA2 genes, as well as multiple common BC susceptibility loci that have been discovered through GWAS. The polygenic nature of BC makes it a perfect candidate to investigate of how well PRS can predict a common cancer.

DEVELOPMENT OF A BREAST CANCER PRS WITH IMPROVED PREDICTIVE PERFORMANCE

We developed a new BC PRS by applying the SCT algorithm described above in a Discovery dataset comprising a subset of individuals from the UK Biobank, and assessed its predictive performance in an independent Test dataset, also from the UK Biobank. This new SCT PRS displayed higher predictive performance (AUC: 0.766, PPV at 3%: 20.20%) than other published BC PRSs, (e.g. Khera et al[6] (AUC: 0.65, PPV at 3%: 15.8%) and from Mavaddat et al[9] (AUC: 0.66, PPV at 3%: 18.0%) (**Table 1**).

Notably, among the three PRS we compared, the SCT PRS we developed used the largest number of SNPs (**Table 1**). To see whether this affected the predictive performance, we aggregated different PRS into new meta-PRS (e.g. Khera+SCT or Mavaddat+SCT), however this did not result in significant changes in AUC compared to the SCT PRS.



PRS panel	SNPs in PRS	AUC (95% CI)	PPV (3%)	Cases in top 3%
Khera	5240	0.650 (0.640-0.658)	15.80	721
Mavaddat	307	0.66 (0.651-0.670)	18.0	819
SCT	577113	0.677 (0.667-0.686)	20.20	921

Table 1: List of the PRS compared to Allelica's SCT PRS for breast cancer Khera refers to the PRS for BC developed by Khera et al[6]. Mavaddat refers to the PRS for BC developed by Mavaddat et al[9]. SCT refers to the PRS for BC developed in this paper with the SCT algorithm. For each PRS, we show the number of genetic variants composing the PRS (**SNPs in PRS)**, the predictive performances quantified as AUC values and 95% confidence intervals (**AUC (95% CI)**), the positive predictive values in the top 3% of PRS distributions (**PPV (3%)**), and the number of BC cases in the top 3% of PRS distributions (**Cases in top 3%**).

Predictive performance of the newly developed BC PRS

We computed the PRS of the individuals in the Test dataset and plotted the distributions of the scores for BC cases and controls (**Figure 1A**). The distributions are both gaussian, with BC cases showing a greater median PRS than controls (median: 0.51 and -0.04, respectively) and an AUC of 0.677.

To assess the ability of the SCT BC PRS to predict BC risk in a Test population we compared the predicted with observed prevalence values. For each individual in the Test dataset, the probability of having the disease was calculated using a logistic regression model with the per-individual PRS score as predictive variable. The predicted prevalence of BC within each percentile of the PRS distribution was calculated as the average probability in each percentile. For all percentiles, predicted BC prevalence was plotted against the corresponding values of observed prevalence (Figure 1B). The values of observed and predicted BC prevalence are in excellent agreement as demonstrated by the localization of the points of the bisector of the graph. The non-significant P-value generated by the HL test (Figure 1B) is further confirmation of the good statistical agreement between predicted and observed prevalence values.

We next evaluated the ability of the SCT BC PRS to stratify BC risk for the female population in the testing dataset. We divided the PRS distribution into percentiles and computed the prevalence of BC in each percentile. Here we use disease prevalence in the test dataset as a measure of the risk of developing BC. Risk stratification for women in the test dataset is shown in Figure 1C. BC risk rises sharply as PRS percentile increases, ranging from 1.64% to 24.6%, for the lowest and highest percentiles respectively.

We also estimated the relative increased risk, which is the ratio between the prevalence at the top 5% of the PRS distribution and the prevalence in the average of the distribution (defined as between the 40% and the 60% percentiles, dashed line in Figure 1C). For the female test population, the relative risk in the top 5% is 2.9 times higher than the average.

PRS is more effective at predicting BC risk than family history

Family history is a well-recognized risk factor of BC and prospective studies have demonstrated a consistent association with the disease [10]. Current prevention guidelines recommend that family history to be incorporated into the risk estimation process that guides treatment decisions for BC [11]. We therefore considered the relationship between family history and genetic risk factors for BC. The aim of this was to answer the following questions:

1. Can SCT BC PRS stratify risk in people with family history?
2. Is SCT BC PRS a better predictor than family history?
3. Does prediction performance increase if a combination of family history and PRS is used?

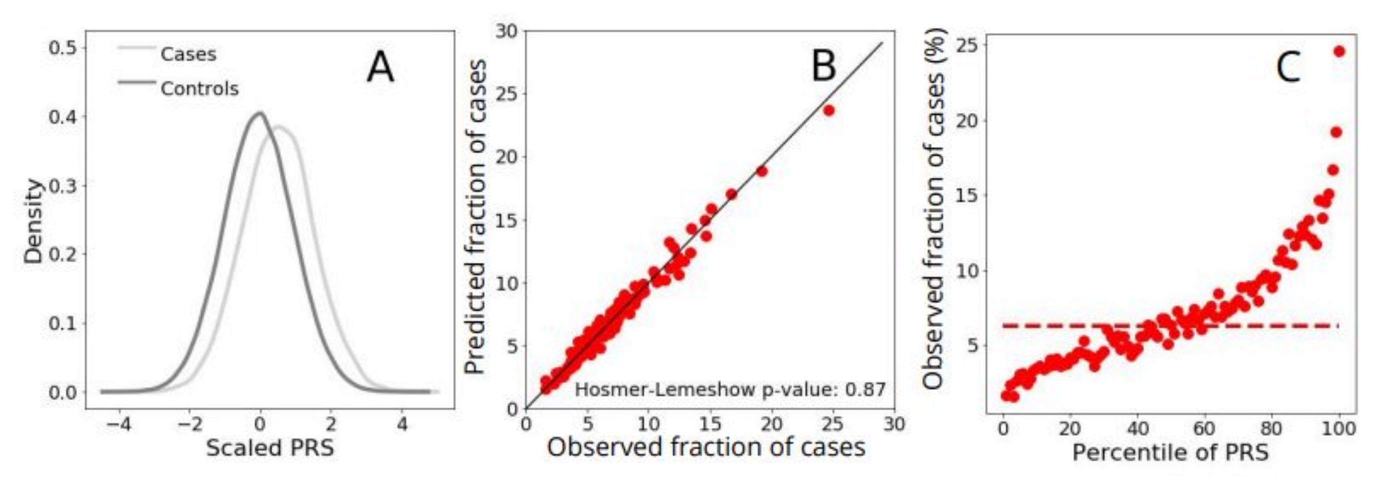


Figure 1: Risk for BC according to the SCT BC PRS panel. A) Distributions (scaled to a mean of 0 and a standard deviation of 1) of the per-individual PRS in female testing populations. **B)** Comparison of fraction of cases between observed and predicted BC. Observed fraction of cases has been calculated as the per-percentile fraction of cases of BC in the per-individual PRS distribution.

(Figure 1 cont.)Predicted BC fraction of cases was calculated for each individual using a logistic regression model with per-individual PRS as predictive variables. Within each percentile of the PRS distribution, BC probability was averaged and this returned the predicted fraction of cases of BC. C) Fraction of cases of BC per percentile of the per-individual PRS distribution calculated in the women from the testing population. Dashed horizontal line: fraction of cases of the average of the PRS distribution (defined as between the 40% and the 60% percentiles)

We computed scaled per-individual distributions for BC cases and controls for those individuals in the test dataset with at least one first-degree relative with a history of BC. Risk distributions for cases and controls are shown in **Figure 2A.** As before, both distributions are gaussian with cases having a higher median value than controls (median: 0.70 and 0.17, respectively). This suggests that the good discriminatory ability of the SCT BC PRS is maintained even in the smaller number of individuals already considered at BC risk based on family history.

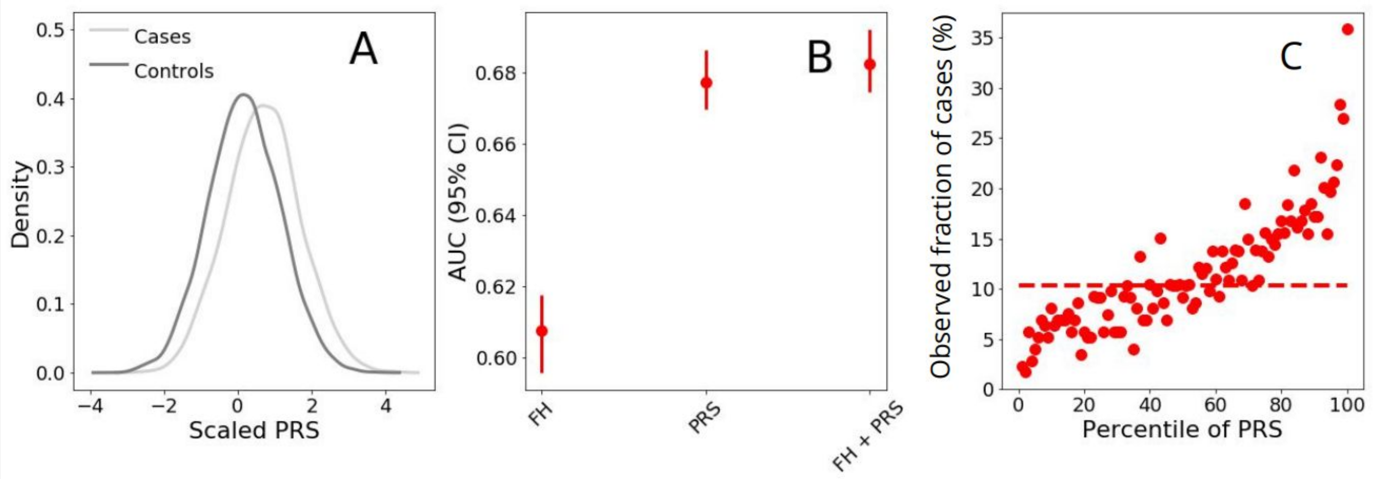


Figure 2: **Risk for BC according to the SCT BC PRS panel in presence of Family history of BC**. **A)** Distributions (scaled to a mean of 0 and a standard deviation of 1) of the per-individual PRS score for BC cases and control individuals with at least one first-degree relative with a history of BC. **B)** AUC values on the testing set of UKBB calculated with logistic regression models having family history of BC (FH), per-individual PRS calculated with the SCT BC PRS panel (PRS), or both (FH + PRS) as explanatory variables. The response variable of the logistic regression model was absence/presence of BC. The logistic regression model comprised additional covariates as control variables such as: age, genotyping array, and the first 4 principal components (PCs) of ancestry. **C)** Observed fraction of cases of BC per percentile of the per-individual PRS score distribution calculated in the testing population for women with at least one first-degree relative with history of BC. Dashed horizontal line: fraction of cases of BC for the average of the PRS distribution (defined as between the 40% and the 60% percentiles).

We then evaluated the ability of the SCT BC PRS to stratify BC risk in the sub-population of women with at least one first-degree relative with a history of BC. As before, we calculated the per-individual PRS distributions and BC prevalence for each percentile of the PRS distributions. BC risk stratification for women with family history of BC are shown in **Figure 2B.** Even with individuals considered at higher risk based on family history, the SCT BC PRS is able to further stratify BC risk over a range of values comprised between 2.3% and 35.8%, for the lowest and highest percentiles, respectively. The observed prevalence is higher in women with family history than in the general population for any percentile considered. For women with family history of BC the relative risk in the top 5% is 2.6 folds higher than the average.

Lastly, we assessed the predictive performance of family history, BC PRS, and the combination of the two, by computing AUC. Figure 2B shows that the BC PRS displays a higher AUC value than family history and it is therefore a better predictor of BC disease. When both risk factors are combined, the predictive performances further improves slightly.

These findings demonstrate that family history and PRS capture different components of the risk of BC and family history should not be considered in isolation without further PRS risk stratification. This is in line with complex genetic structure of the BC disease that can be caused by either rare mutations with high penetrance or by the combination of multiple common BC susceptibility loci.

Modeling BC risk at 75 years of age

The total Test dataset was divided in two sub-populations: women with and without family history of BC. These two subpopulations were divided into percentiles and average percentile values were used as predictive variable in Cox-proportional hazard models with age of event (BC or follow-up) as timescale. The Hazard Ratios estimated in the two Cox models were used to calculate the dependence of the probability (cumulative incidence) of BC by age 75 from BC SCT PRS percentiles, conditioned on the mean values of control covariates: Genotyping Array and first four principal components of ancestry. Figure 3 show that lifetime risk of BC increases with increasing PRS percentiles for both populations (without and with family history of BC).

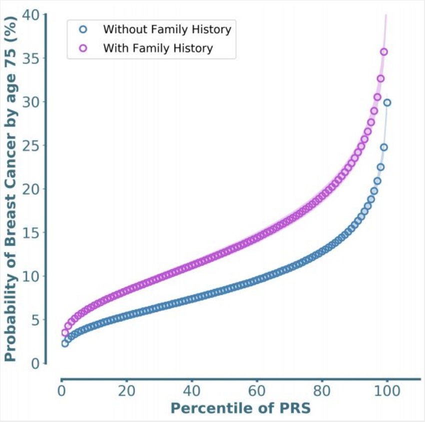


Figure 3: Probability of BC by age 75 for women with and without family history of BC.

CONCLUSIONS

In this article we have described the development of a PRS for Breast cancer. We tested its prediction and risk stratification performances in the UK Biobank, which is the largest population dataset currently available. When compared with previously published PRS, the PRS showed highest predictive performance (Table 1). Additionally, we have demonstrated that PRS for BC is able to identify a notable fraction of the UK Biobank population (5%) with a 3 fold or higher increased risk of developing BC compared to the population average. The risk stratification ability of PRS is maintained even in individuals already considered at risk based on family history.

We have developed an easy to use, intuitive software suite to perform PRS analysis, which is now available to clinical laboratories and research groups as a fully automated, HIPPA and GDPR compliant and certified as a CE marked medical device. The software calculates Polygenic Risk Scores for a number of complex diseases and can analyse thousands of samples in parallel having the potential to improve health care prevention through its large scale implementation into public health practice.

REFERENCES

[1] T. A. o. M. Sciences, “Our data-driven future in healthcare”, no. November, 2018. [Online]. Available: https://acmedsci.ac.uk/file-download/74634438.

[2] C. Bycroft, L. T. Elliott, A. Young, et al., “The UK Biobank resource with deep phenotyping and genomic data”, Nature, vol. 562, no. 7726, pp. 203–209, 2018, ISSN: 0028-0836. DOI: 10.1038/s41586-018-0579-z. [Online]. Available: https://www.nature.com/articles/s41586- 018- 0579-zhttp://www.nature.com/articles/s41586- 018- 0579-z.

[3] N. R. Wray, S. H. Lee, D. Mehta, et al., “Research Review: Polygenic methods and their application to psychiatric traits”, en, Journal of Child Psychology and Psychiatry and Allied Disciplines, vol. 55, no. 10, pp. 1068–1087, 2014, ISSN: 14697610. DOI: 10.1111/jcpp.12295. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/jcpp.12295.

[4] A. R. Martin, M. J. Daly, E. B. Robinson, S. E. Hyman, and B. M. Neale, Predicting Polygenic Risk ofù Psychiatric Disorders, Jul. 2019. DOI: 10.1016/j.biopsych.2018.12.015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S000632231832119X.

[5] S. M. Purcell, N. R. Wray, J. L. Stone, et al., “Common polygenic variation contributes to risk of schizophrenia and bipolar disorder”, Nature, vol. 460, no. 7256, pp. 748–752, 2009, ISSN: 00280836. DOI: 10.1038/nature08185.

[6] A. V. Khera, M. Chaffin, K. G. Aragam, et al., “Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations”, eng, Nature Genetics, vol. 50, no. 9, pp. 1219–1224, 2018, ISSN: 15461718. DOI: 10.1038/ s41588-018-0183-z.

[7] B. J. Vilhjálmsson, J. Yang, H. K. Finucane, et al., “Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores”, eng, American Journal of Human Genetics, vol. 97, no. 4, pp. 576–592, Oct. 2015, ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2015.09.001.

[8] F. Privé, B. J. Vilhjálmsson, H. Aschard, and M. G. B. Blum, “Making the most of Clumping and Thresholding for polygenic scores”, bioRxiv, p. 653 204, Jan. 2019. DOI: 10.1101/653204. [Online]. Available: http://biorxiv.org/content/early/2019/06/28/653204.abstract.

[9] N. Mavaddat, K. Michailidou, J. Dennis, et al., “Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes”, Am. J. Hum. Genet., vol. 104, no. 1, pp. 21–34, 2019.

[10] H. R. Brewer, M. E. Jones, M. J. Schoemaker, A. Ashworth, and A. J. Swerdlow, “Family history and risk of breast cancer: an analysis accounting for family structure”, Breast Cancer Res. Treat., vol. 165, no. 1, pp. 193–200, 2017.

[11] D. G. Evans, J. Graham, S. O’Connell, S. Arnold, and D. Fitzsimmons, “Familial breast cancer: summary of updated NICE guidance”, BMJ, vol. 346, f3829, 2013.