A Breast Cancer Polygenic Risk Score Validation in 15,490 Brazilians using Exome Sequencing

Flávia Eichemberger Rius[1,2], Rodrigo Guindalini[3], Danilo Viana[1], Júlia Salomão[1], Laila Gallo[1], Renata Freitas[1], Cláudia Bertolacini[1], Lucas Taniguti[1], Danilo Imparato[1], Flávia Antunes[1], Gabriel Sousa[1], Renan Achjian[1], Eric Fukuyama[1], Cleandra Gregório[1], Iuri Ventura[1], Juliana Gomes[1], Nathália Taniguti[1], Simone Maistro[2], José Eduardo Krieger[4], Yonglan Zheng[5], Dezheng Huo[6], Olufunmilayo I. Olopade[5], Maria Aparecida Koike[2], David Schlesinger[1]

1. Mendelics, São Paulo, Brazil.
2. Comprehensive Center for Precision Oncology - C2PO, Centro de Investigação Translacional em Oncologia (CTO), Departamento de Radiologia e Oncologia, Instituto do Cancer do Estado de Sao Paulo (ICESP), Hospital das Clinicas HCFMUSP, Faculdade de Medicina, Universidade de Sao Paulo, Sao Paulo, Brazil.
3. Instituto D'Or de Pesquisa e Ensino (IDOR), São Paulo, Brazil.
4. Instituto do Coração, Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo - FMUSP, São Paulo, Brazil.
5. Medicine and Human Genetics, Center for Clinical Cancer Genetics and Global Health, University of Chicago Medical Center, Chicago, USA.
6. Department of Public Health Sciences, University of Chicago, Chicago, USA.

Abstract

Purpose

Brazil has a highly admixed population. Polygenic Risk Scores (PRS) have been mostly developed from European population studies and applying them to other populations is challenging. To assess the use of PRS for breast cancer (BC) risk in Brazil, we validated four PRSs developed in the Brazilian population.

Patients and Methods

We analyzed 6,362 women with a history of breast cancer and 9,128 unphenotyped adults as controls in a sample obtained from a clinical laboratory. Genomic variants were imputed from exomes and scores were calculated for all samples.

Results

After excluding individuals with known pathogenic or likely pathogenic variants in *BRCA1*, *BRCA2*, *PALB2*, *PTEN*, or *TP53,* and first-degree relatives of the probands, 5,730 cases and 8,847 controls remained. Four PRS models were compared, and PRS 3820 from Mavaddat *et al.* 2019 performed best, with an Odds Ratio (OR) of 1.41 per standard deviation (SD) increase (p-value: < 0.0001) and an OR of 1.94 (p-value: < 0.0001) for the individuals in the top risk decile. PRS 3820 also performed well for different ancestry groups: East Asian majority (Group 1), Non-European majority (Group 2), and European majority (Group 3), showing significant effect sizes for all groups: (Group 1: OR 1.54, p-value 0.006; Group 2: OR 1.44, p-value: <0.001; Group 3 OR:

1.43, p-value: <0.001). PRS 90% compares with monogenic moderate BC risk genes (PRS90 OR: 1.94; CHEK2 OR: 1.89; ATM OR: 1.99).

Conclusion

PRS 3820 can be accurately used in the Brazilian population. This will allow a more precise BC risk assessment of mutation-negative women in Brazil.

Introduction

Breast cancer (BC) is a critical global health concern, representing the most common cancer diagnosed among women[1]. In Brazil, over 70,000 women are diagnosed with BC every year, accounting for 30% of all cancers in the female population[2].

Approximately 10% of all BC cases are attributable to germline pathogenic variants in susceptibility genes[3]. Rare variants in high penetrance genes (*BRCA1, BRCA2, TP53, PTEN*, and *PALB2*) and in moderate penetrance genes (*CHEK2* and *ATM*) are associated with a more than 4-fold and 1.5–4 fold increased risk of BC, respectively[4, 5]. Rare variants in these genes account for approximately 25% of the genetic risk. The remaining genetic risk (~75%) is derived from common, low penetrance variants that individually confer small risk, but which combined effect can be substantial[4–6].

Genome-wide association studies (GWASs) have been predominantly carried out in European populations[7–10]. Evaluation of PRS across different genetic and environmental backgrounds is essential to enable the implementation of genetic risk stratification strategies for individuals from non-European populations[11].

The Brazilian population exhibits a unique, highly admixed, genetic composition. It is mostly derived from a combination of Native Americans, Southern Europeans (Portuguese, Spanish, and Italian) that immigrated in the period 1500-1900, and Sub-Saharan Africans brought through extensive slave trading until the 1800s. More recently, from 1822 to the first half of the 1900s, other smaller waves of immigration also contributed to Brazil's remarkable diversity, including Japanese, Lebanese, German, and Eastern Europeans[12]. Three in every four Brazilians have multiple genetic ancestries[13,14]. Given Brazil's genetic diversity, any PRS developed in predominantly European populations requires validation before it can be used in clinical settings.

Several laboratory methods are available for genotyping variants directly or indirectly (imputation), including microarrays, whole exome sequencing (WES), and whole genome sequencing (WGS). WES offers an affordable and scalable alternative to arrays and WGS, while allowing for simultaneous rare and common variant genotyping.

In this study, we evaluate four BC PRSs[7,8,15] developed using WES in 15,490 Brazilians.

Methods

Study population

A total of 15,490 individuals were selected for this study, including 6,362 women with breast cancer history, and 9,128 adult unphenotyped controls. Both clinical and genetic data were collected from a database of a College of American Pathology (CAP)–accredited laboratory (Mendelics, São Paulo, SP, Brazil). All BC and control subjects provided Informed Consentment for use of retrospective anonymized data for

research purposes. Samples were anonymized before analysis. Clinical records such as BC histological type and age of diagnosis were obtained from genetic test requisitions. The study was IRB-approved (CAAE: 70112423.3.0000.0068).

Relatedness calculation and data filtering

Relatedness of individuals was obtained from the exomes using somalier software[16], following the standard protocol required for a VCF file (https://github.com/brentp/somalier#readme). Concerning related individuals removal, if two individuals had a first-degree relationship, one of them was randomly selected to be included in the dataset. However, if individuals had two or more first-degree relationships, all related individuals were excluded from the dataset. This process resulted in a total of 211 removals. Furthermore, 73 individuals were removed from the sample due to unavailability of files necessary for genome imputation.

PRS analyses were performed after filtering out cases and controls with pathogenic or likely-pathogenic (P/LP) variants in BC genes *BRCA1*, *BRCA2*, *TP53*, *PALB2*, and *PTEN*.

Exome sequencing and imputation

Exome sequencing data were generated from buccal swab or venous blood samples with standard protocol for Illumina Flex Exome Prep, using a custom probe set from Twist Biosciences. Sequencing was conducted in Illumina sequencers and the bioinformatics pipeline for data analysis followed Broad Institute's GATK best practices

(https://gatk.broadinstitute.org/hc/en-us/sections/360007226651-Best-Practices-Workflows), with alignment to GRCh38.

Imputation of exomes was based on a panel of 2,504 individuals of all ancestries from the 1000 Genomes Project (1KGP)[17] on GRCh38 (2017 release) (https://www.internationalgenome.org/data-portal/data-collection/grch38). All regions captured from the exome sequencing comprehending at least 1x coverage, as well as off-target regions, were considered for the imputation, performed using Glimpse (v1.1.0) software[18].

Polygenic Risk Score calculation

Four BC PRSs with publicly available summary statistics, from three different studies, were evaluated in this work: Khera *et al.* 2018[7], with 5,218 variants; Mavaddat *et al.* 2019[8] PRSs (with 313 and 3,820 variants); and UK Biobank[15] (UKBB) PRS obtained from a variant thresholding (p-value < 10e-5) on summary statistics for phenotype code 20001_1002, with 7,538 variants.

The PRS variants were selected based on exome bed kit distance and minor allele frequency (MAF). Additionally, the PRS from Mavaddat study, originally with 3,820 variants, had a pathogenic variant of moderate-penetrance in *CHEK2* gene (*CHEK2* p.Ile157Thr - Clinvar: RCV000144596) that was removed to avoid conflation with monogenic risk.

PRS calculation was performed using a software developed by Mendelics, evaluating the weighted sum of beta values, in which weights are based on the number of the

129 individual's alleles containing the variant of the PRS file. The sum is normalized by all
130 beta positive and negative values so the final value can be between zero and one.
131

132 **Genetic Principal Component Analyses (PCA) Assessment**

133 PCA was calculated for exomes from a projection in 1KGP[17] and Human Genome
134 Diversity Project (HGDP)[19] samples. Only variants with MAF > 1% and that could have
135 been directly genotyped using WES were included for the PCA analysis in 1KGP and
136 HGDP samples using plink2[20]. Exomes were converted to plink bfile format (bed, bim,
137 and fam files) and had duplicated variants removed. PCA projection for 10 PCs was
138 calculated using plink2 –score method, with allele frequencies from the breast cancer
139 case-control sample.
140

141 **Ancestry evaluation**

142 Admixture[21] was used to extract continental ancestries from all non-related and data
143 completed exomes. The analysis was supervised by the 1KGP samples, after removal
144 of South Asian, Oceania, and admixed Americans from the GRCh38 1KGP release of
145 2017. South Asian and Oceania ancestries were removed because they are not a
146 significant part of Brazilian ancestral composition. Latin American admixed populations
147 (Colombian, Peruvian, Puerto Rican, and Mexican) were removed to avoid confounding
148 with the native americans belonging to the same population label. Continents evaluated
149 were: Africa - AFR, America - AMR, East Asia - EAS, and Europe - EUR. Ancestry

results were further used for splitting individuals into groups according to their ancestry composition, to further analyze the effect size of PRS on each group.

Paired imputed and sequenced genomes analysis

Exome-imputed variants and directly sequenced variants from WGS were compared using 1001 samples from an independent Brazilian population dataset (http://elsabrasil.org/) that had both WES and WGS available. The WES were sequenced and imputed also using the same method previously described. BC PRS-3820 from Mavaddat *et al.* study was calculated for both imputed and sequenced genomes, and their Spearman correlation was calculated using R software base function *cor.test*.

Statistical analyses

PRS values were standardized according to the control values prior to all statistical analyses. PCs were Z-scored prior to analyses. To assess the effect size of PRS on breast cancer status (0 = control, 1 = case) corrected for PCs, Odds Ratio (OR) per standard deviation of PRS was calculated by performing a logistic regression of BC status with PRS and PCs 1 to 10 as predictors. AUC for the full dataset evaluation was obtained using the yardstick R package (yardstick.tidymodels.org/) roc_auc function, in the testing data split (25%). In order to find segmentation effect-sizes, individuals were classified into deciles or percentiles following the left-open and right-closed intervals. OR for deciles was calculated by first selecting only the decile analyzed and the interval

from 40-60% individuals as the control section, and binarizing it (0 = belongs to the control interval 40-60%, 1 = belongs to the decile analyzed, for example, 10%); and performing a logistic regression analysis on the binarized decile information with correction for PCs 1 to 10. A similar approach was conducted for calculating the OR on percentiles for comparison with Mavaddat's[8] PRS validation. For each ancestry proportion group, AUC was estimated using 10-fold cross-validation with the R package *caret*[22]. All PRS 95% confidence intervals (CI) were obtained from the logistic regression output from the R function *glm* (stats package[23]). OR and CI for genes *BRCA1*, *BRCA2*, *PALB2*, *TP53*, *ATM* and *CHEK2* were obtained using *epitools* R package[24]. All statistical tests performed were two-tailed.

Results

Case-control sample selection and characteristics

After removal of 211 subjects with a first-degree relationship and 73 with missing files necessary for imputation, a total of 15,206 subjects remained (**Supplementary Table 1**). Four percent of all cases and controls were removed from the analysis due to their presence of pathogenic or likely-pathogenic (P/LP) variants in high penetrance genes with OR > 5 for breast cancer: *BRCA1*, *BRCA2*, *TP53*, *PALB2*, and *PTEN* (n = 629). Therefore, the sample used for PRS evaluation consisted of 5,730 women with a BC history, and 8,847 unphenotyped controls, both with known sex and age (**Table 1**).

Table 1. Demographics of cases and controls in BC dataset used for PRS evaluation

		Case	Control	Total	p value
	Total	5,730	8,847	14,577	-
Sex	F	4,225	5,730	9,955	-
	M	-	4,662	4,662	-
Age	Total	49.5 (11.7)	41.6 (13.3)	44.8 (13.3)	0.000
	F	49.5 (11.7)	42 (13.7)	46.4 (13.1)	0.000
	M	-	41.3 (12.9)	41.3 (12.9)	-

p-values obtained from two-tailed t-tests.

Ancestry composition of our sample was obtained using ADMIXTURE analysis[21], supervised by EUR, EAS, AFR and non-admixed AMR populations of 1KGP and HGDP. The results show that the majority of individuals have EUR as their greatest ancestry proportion (median 84%, SD 18%). Besides that, a significant portion of AFR (median 6%, std. dev. 12%) and AMR (median 8%, SD 7%) ancestries are present, complemented with a variety of EUR proportions. A small quota of EAS is also observed (median < 1%, SD 12%), composed by 214 individuals with over 70% of this ancestry.

Figure 1. Ancestry composition of our Brazilian cohort. Estimated ancestries are shown as proportions per individual. Each thin bar represents one individual and their ancestry proportion. Europe (EUR) in purple, Africa (AFR) in blue, East Asia (EAS) in green and America (AMR) in yellow.

Effect sizes of four different PRSs in the Brazilian population

Four PRS files from three studies were selected for initial effect size investigation in our cohort (**Supplementary Table 2**). All four PRS files had their variants further filtered to address only variants covered by the imputation of our exomes.

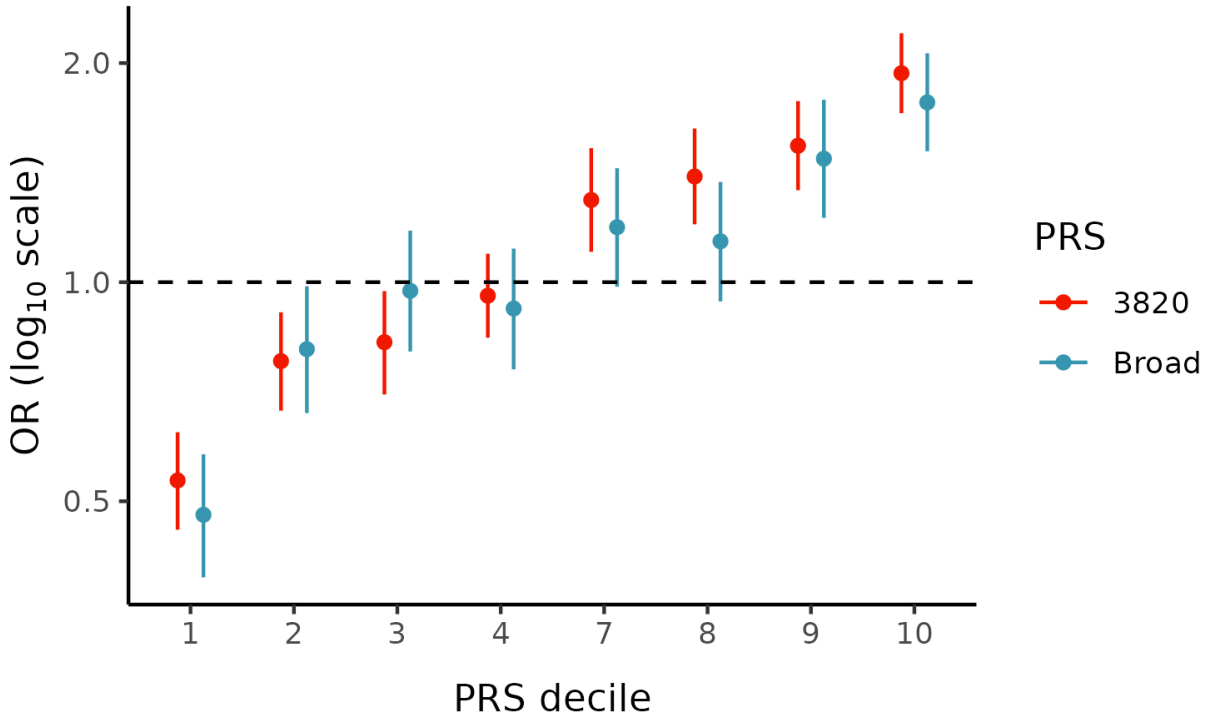
PCA was performed on the exomes to capture the population genetic structure. PRSs were calculated for the imputed genomes (details described in the **Methods**) and standardized by z-score to improve interpretability. To avoid confounding from P/LP variants on PRS effect, we have evaluated only individuals without those rare variants (n = 14,577). Effects were corrected for the ten first PCs, and results are all reported in **Supplementary Table 3**.

Both PRS_{Broad} and PRS_{3820} performed well, with very significant effect sizes (both p-values < 0.0001) following the direction of risk rise as the PRS increases (OR_{Broad}: 1.52; OR_{3820}: 1.41). PRS_{313} and PRS_{UKBB} have not reached significance level for their OR results ($p\text{-value}_{313}$: 0.315 and $p\text{-value}_{UKBB}$: 0.985). Goodness of fit of the model is also greater for PRS_{3820} (Nagelkerke pseudo-R^2: 0.061) and PRS_{Broad} (Nagelkerke pseudo-R^2: 0.051). Note that pseudo-R^2 values should not be interpreted as a linear regression R^2 value, but as a metric of improvement from null model to fitted model, which has its value mainly by being compared between different PRS models in which a greater pseudo-R^2 indicates a better goodness of fit to the data.

Since PRS_{Broad} and PRS_{3820} showed significant results per standard deviation, they were used to split the data into deciles to evaluate BC risk conferred by PRS in each strata. These analyses were also corrected for the first ten PCs. Interestingly, shorter confidence intervals and a better “staircase” shape can be seen for PRS_{3820} plot in comparison to PRS_{Broad} (**Figure 2**). Moreover, especially the top 10% (90-100% interval) present a much greater effect for PRS_{3820} ($OR_{90\text{-}100}$: 1.94; CI: 1.71 - 2.20) compared to

237 PRS_{Broad} ($OR_{90\text{-}100}$: 1.77; CI: 1.51 - 2.10) (**Supplementary Table 4**), indicating a better
238 performance of the former in identifying women with increased risk of BC. Therefore we
239 decided to focus our next analyses on PRS_{3820}, which was the best PRS to identify BC
240 risk in our Brazilian population.

241



242

243 **Figure 2**. **Effect sizes by decile of PRS_{3820} and PRS_{Broad}.** Odds Ratios (OR) and
244 Confidence Intervals (CI) for PRS_{3820} (red) and PRS_{Broad} (blue). ORs for both PRS
245 deciles were corrected for the first ten PCs.

246

247 **PRS_{3820} performance compared with the original study**

248 As seen in the previous results, the PRS_{3820} showed a positive association with
249 increased risk of BC (OR per standard deviation: 1.41; CI: 1.36 - 1.47) after correction

for the first 10 principal components (PCs). This association was slightly lower when compared to the original study test set, composed of only Europeans (OR: 1.66; CI: 1.61 - 1.70). Besides that, performance of our model with PRS_{3820} in identifying BC cases was very similar to the original study ($AUC_{Brazilians}$: 0.610 vs. $AUC_{Europeans}$: 0.636). After calculating OR per percentiles, we observed that the PRS_{3820} exhibited an expressive risk increase for our admixed population, although the increase was smaller than the original study, which applied the PRS_{3820} to a population with the same ancestry it was originated from ($OR_{Brazilians}$ >99: 2.72; $OR_{Europeans}$ >99: 3.95).

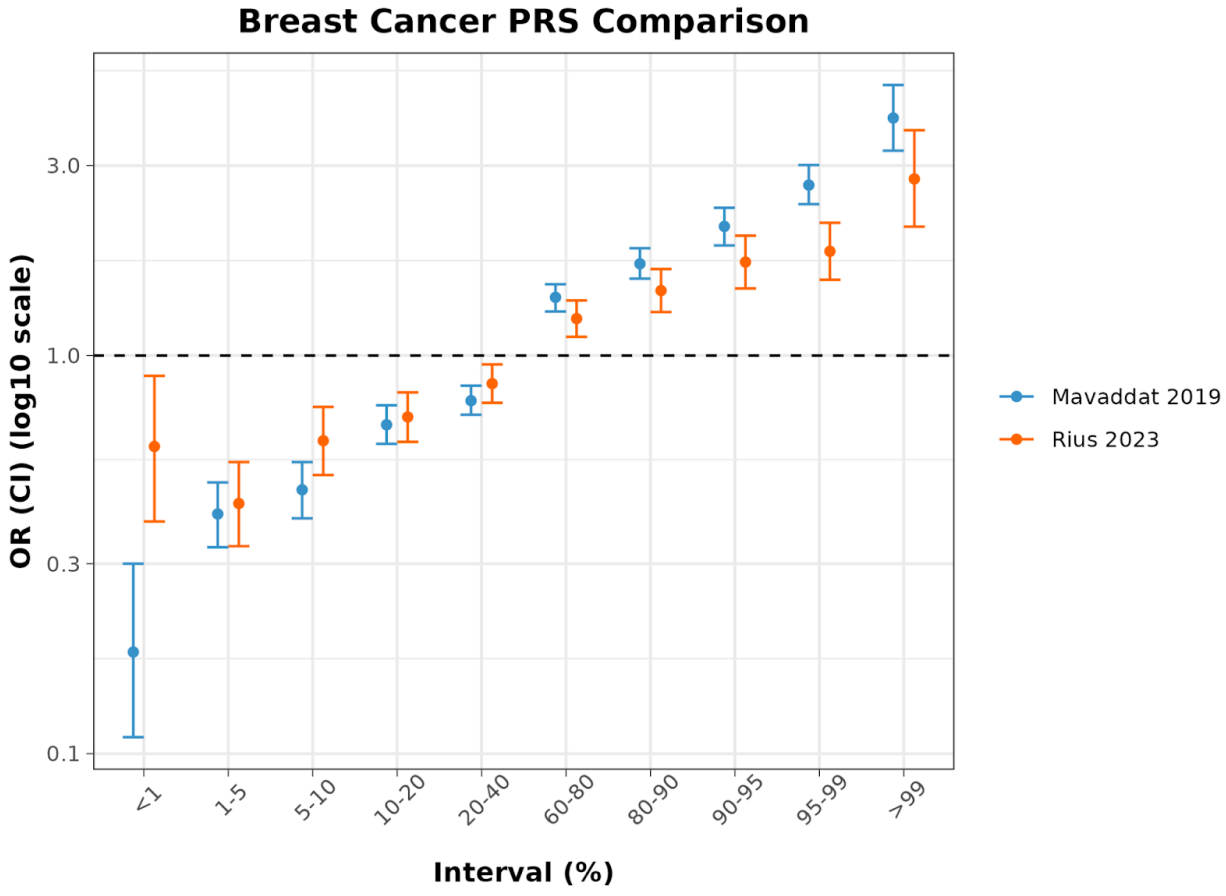


Figure 3. Comparison of PRS_{3820} performance for Europeans and Brazilians. The plot shows the PRS_{3820} adapted in this study (orange), with 2,892 variants, compared with the original from Mavaddat *et al.* study (blue), with 3,820 variants.

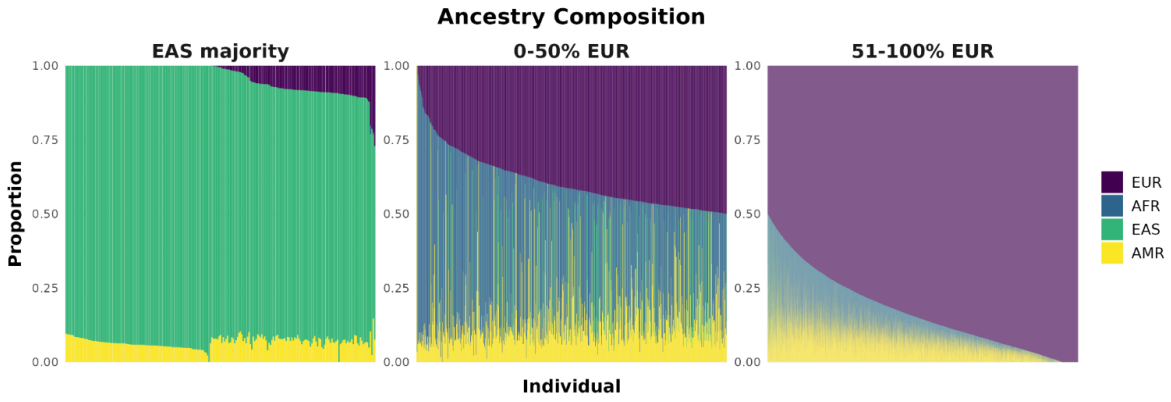
The lower interval, comprehending the lowest 1% of PRS values, showed a smaller decrease in BC risk compared to the original study. This result is probably related to the small sample size of this section, with only 31 cases and 88 controls available to calculate OR. In addition, the 95th to 99th percentile interval exhibited marginal growth in odds ratio (OR) when contrasted with the interval immediately below (OR 90th-95th: 1.75, OR 95th-99th: 1.83). Besides that, both effect sizes show an expressive increase in BC risk due to PRS results. This might be partly due to the cohort sample size. Our study evaluated a total of 14,577 individuals, while Mavaddat's evaluated twice this number in their test dataset composed of joined cohorts (n = 29,751).

PRS evaluation by ancestry composition

Since our sample contains a great majority of EUR ancestry proportion, we decided to evaluate the PRS effect size in different ancestry compositions. We have created three groups: EAS majority (> 50% EAS, n = 217), 0 - 50% EUR (n = 763) and 51 - 100% EUR (n = 13,597). All three bins had statistically significant ($p < 0.001$) ORs above 1.40 (1.54, 1.44 and 1.43, respectively) per PRS standard deviation, showing a positive association of the PRS value with increased BC risk. The EAS majority group shows a wider confidence interval due to the small sample size (cases = 64, controls = 153). Besides that, the lower tail of the 95% confidence interval has an OR of 1.14

(**Supplementary Table 5**), which means at least 14% risk rise for each unit of standardized PRS increase.

A.



B.

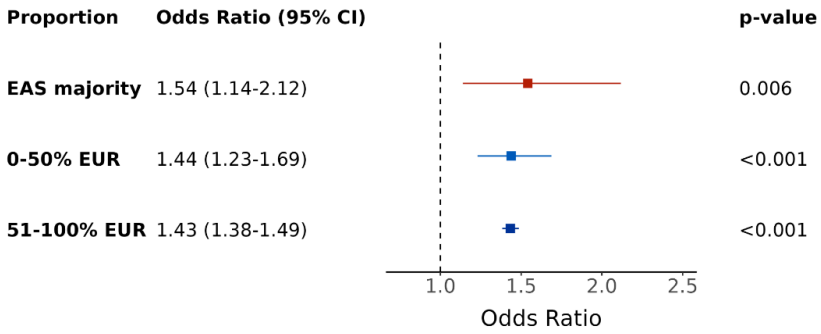


Figure 4. Breast cancer Odds Ratio by ancestry proportion. The cohort was split into three groups based on main ancestry: EAS majority (>50% EAS), 0 - 50% EUR and 51 - 100% EUR (A) Ancestry composition of each group, with colors representing continental ancestries for each subject. (B) Breast cancer ORs by PRS_{3820} standard deviation for the three groups. p-values displayed were corrected for multiple-hypothesis testing using Bonferroni method.

Comparison of PRS derived from genomes imputed from exomes with WGS

A correlation of 0.76 (p value < 2.2e-16) was obtained between BC PRS_{3820} values calculated from imputed genomes and WGS, showing a consistent concordance between both methods (**Supplementary Figure 1A**).

When we compared imputed (exome) and sequenced genomes (WGS), most of the extreme PRS_{3820} values were concordant (decile 1: 56%; decile 10: 60%) (**Supplementary Figure 1B**). Furthermore, most of the proportion which is not in the same decile is in the surrounding deciles, which indicates a low deviation from the purpose of predicting risk.

Comparison of PRS and breast cancer genes effect size

For the purpose of understanding how the PRS_{3820} effect size compare to known high and moderate risk genes for BC, we have compared OR of the top PRS_{3820} decile (PRS90) with all pathogenic variants located in *TP53*, *BRCA1*, *BRCA2*, *PALB2*, *ATM* and *CHEK2* genes (**Figure 5**) in this cohort of individuals.

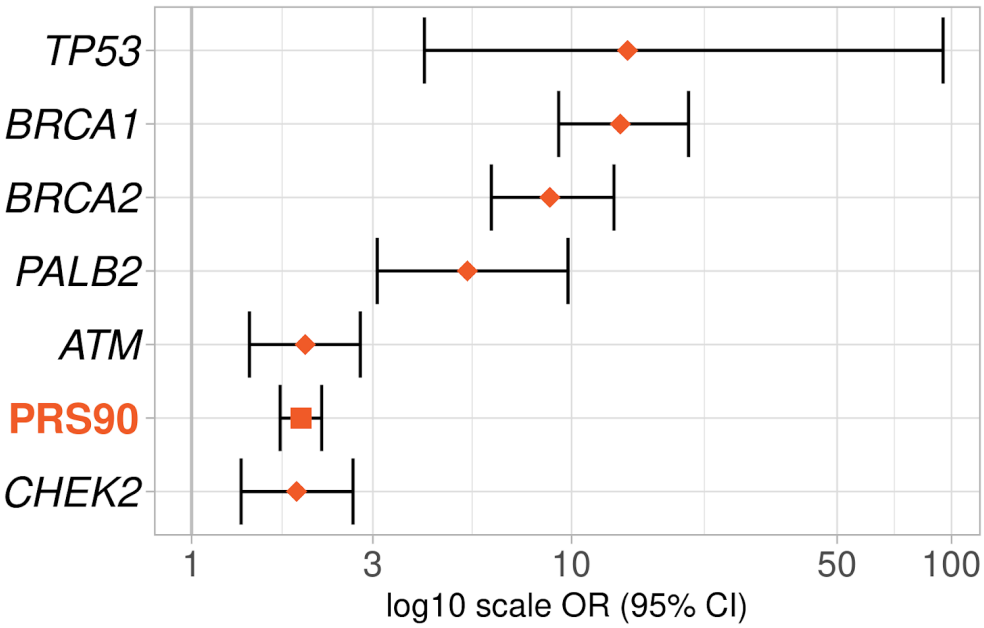


Figure 5. Effect sizes of 90th percentile of PRS and BC genes in BC risk. Effect sizes (OR and CI) were obtained according to the presence of pathogenic variants in the genes *TP53*, *BRCA1*, *BRCA2*, *PALB2*, *ATM* and *CHEK2*, or belonging to the 90th to 100th percentiles of PRS_{3820}.

As expected, *TP53*, *BRCA1* and *BRCA2* present the most extreme BC risks (OR: 14.05, CI: 4.1-95.05; OR: 13.43, CI: 9.25-20.32; and OR: 8.77, CI: 6.15-12.93, respectively). PRS90 risk (OR: 1.94, CI:1.71-2.2) can be compared with moderate risk BC genes *ATM* (OR: 1.99, CI: 1.42-2.78) and *CHEK2* (OR: 1.89, CI: 1.35-2.66). This result indicates how an increased risk for BC due to PRS90 could be interpreted in the clinical context, potentially following the same care protocols as for a moderate risk monogenic variant for BC.

Discussion

In the present study we have validated a breast cancer PRS developed from Europeans in the highly admixed Brazilian population. The PRS adapted from Mavaddat *et al.* study with 2,892 variants[8] showed a statistically significant risk prediction value (OR: 1.41 per SD). Furthermore, individuals classified in the top decile had an expressive effect size (OR: 1.94; CI: 1.71 - 2.20) of almost one-fold increased risk of BC compared to the middle percentiles (40-60%). This PRS highest decile risk is comparable with the previously reported risks for moderate-penetrance monogenic variants in *ATM*, *NF1*, and *CHEK2* genes (1.82, 1.93, and 2.47 OR, respectively)[25], and also with risks in *ATM* and *CHEK2* calculated in our sample (1.99 and 1.89 OR, respectively).

This study is based on a previous study from Mavaddat et al. 2019, which developed and validated a PRS with 3,820 variants evaluating aggressive BC risk (metastatic BC). For all BC subtypes (ER+ and ER-) they found an OR of 1.71 per SD (CI: 1.64 1.79) in the validation set (n = 29,751; cases = 11,428), and OR 1.66 per SD (CI: 1.61 - 1.70) in the prospective set (n = 190,040; cases = 3,215). These values are even greater compared to the widely used 313 PRS (OR: 1.65 per SD; CI: 1.59 - 1.72 in validation set). However, they included a *CHEK2* gene pathogenic variant in the PRS and worked with only aggressive BC, which may have led to overestimating their OR values. A study from Liu and colleagues has evaluated another modification of the same PRS with 3,820 variants developed from Mavaddat *et al.* for African, Latin, and European

346 populations[26]. According to the study, the effect size of this PRS to a BC risk in an
347 European sample (n = 33,594) was 1.40 per standard deviation, a result very similar to
348 ours for a Brazilian sample (OR 1.41 per SD; n = 14,477). They deliberately have
349 included women with *in situ* ductal BC as well as women with metastatic BC, what they
350 claim to be a reason for OR decline compared to the original study, which included only
351 metastatic BC women both in their discovery and validation sets. Our study, however,
352 does not distinguish BC types, therefore we hypothesize that both metastatic and *in situ*
353 BC are included, which may be a factor, together with genetic population structure, that
354 decreased the OR compared to the original study.

355

356 Furthermore, significant effect sizes per PRS standard deviation were obtained for
357 distinct ancestry compositions within our sample. Due to the high proportion of EUR
358 (median 84%, std. dev. 18%), we separated the sample into groups with different
359 ancestry compositions. Despite the small sample size (n = 217) of the EAS majority
360 group (**Supplementary Table 5**), there was a statistical significance (adjusted p-value:
361 0.006) for the effect size in this group, which had similar magnitude (OR: 1.54, CI:
362 1.40-2.12) of the full sample (OR: 1.41, CI: 1.36-1.47, p-value: < 0.0001). Also, PRS_{3820}
363 had significant and expressive effect sizes on BC risk for both EUR proportion groups
364 (0-50% EUR OR: 1.44, CI: 1.23-1.69; 51-100% EUR OR: 1.43, CI: 1.38-1.49). These
365 results evidence that, for individuals with a more prominent East Asian ancestry, for
366 admixed individuals, and for predominantly Europeans, PRS_{3820} is still effective in
367 stratifying BC risk.

368

All of our PRS values were calculated according to a new methodology: the imputation of genomes from exomes. This approach has demonstrated to be very successful for PRS calculation and assessment of BC risk in our study, and could be very interesting for laboratories that already perform exome sequencing as a cost-effective methodology to identify P/LP variants for BC. A variety of studies have compared low-pass genome sequencing with arrays for different applications, such as pharmacogenetics, GWAS, CNV detection, and PRS calculation[27,28,29] . The study of Li *et al.*[28] reported improved accuracy for polygenic risk prediction of imputed low-pass genome compared to array imputation for both coronary artery disease and BC. Despite the slight difference we found between PRS values calculated from sequenced genomes and imputed genomes from exomes (Spearman correlation: 0.76), decile classification showed satisfactory concordance between both methods for the majority of results in the extreme deciles (1 and 10th), which are the most important to define decreased or increased risk. Unfortunately, it was not possible to assess the predictive power of PRS values calculated from genomes of BC patients due to unavailability of paired exome and genome data.

Among familial BC cases, approximately 25% have a P/LP germline variant reported[30]. In the Brazilian population, a robust study with 1,663 breast cancer patients detected 20.1% of P/LP germline variants using multigene panel testing[4,6]. A 2017 study reported that 18% of the hereditary BC can be explained by a polygenic effect of variants discovered in a GWAS[31]. Therefore, employing this PRS in the clinical practice might bring an elucidation to BC Brazilian families without high or moderate-effect germline

variants detected. Moreover, women without prior knowledge of their familial BC condition, or even those with a high PRS risk by chance, will have the possibility to be informed of their results and share them with their physicians to adopt preventive actions accordingly to their risk strata, such as intensifying surveillance adding breast magnetic resonance imaging to mammography screening[32].

In conclusion, our work was able to validate a PRS developed in Europeans in the Brazilian population, using imputed genomes from exomes. The top decile of this PRS presents a risk comparable to moderate-risk monogenic variants for BC. Future studies will be required to evaluate the combination of PRS with P/LP variants and clinical factors in order to deliver more informative results to patients, thus physicians can recommend prevention strategies based on their combined polygenic and monogenic BC risk.

Ethics Statement

This work was approved by the Ethics Committee Comissão para análise de projeto de pesquisa of Hospital das Clínicas da FMUSP - CAPPesq under the CAAE number 70112423.3.0000.0068.

Acknowledgements

We thank all individuals once sequenced in Mendelics laboratory who have consented to participate in this research. We also thank all UKBB participants for their contribution to the PRS hereby analyzed, and all authors from previous studies on BC PRSs in which we based our validation (Khera *et al.* 2018 and Mavaddat *et al.* 2019). Maria Aparecida Azevedo Koike Folgueira received research support from Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil (CNPq-308052/2022-6).

Data Availability

All variants and betas which compose the four evaluated PRSs are available as Supplementary Information. Individual cases and controls data are not publicly available due to the confidentiality consentment agreement signed by all included in the study.

Competing Interests

Flávia Eichemberger Rius, Danilo Viana, Júlia Salomão, Laila Gallo, Renata Freitas, Cláudia Bertolacini, Lucas Taniguti, Danilo Imparato, Flávia Antunes, Gabriel Sousa, Renan Achjian, Eric Fukuyama, Cleandra Gregório, Iuri Ventura, Juliana Gomes, Nathália Taniguti, and David Schlesinger are currently employed by Mendelics, or were employed at the time of the study.

Rodrigo Guindalini acted as a consultant for AstraZeneca, Janssen Oncology, Roche/Genentech and Igenomix; received speaker honoraria from AstraZeneca, Bristol Myers Squibb, GlaxoSmithKline, Merck Sharpe & Dohme Brasil, Novartis, and Roche outside the submitted work; and has equity in Mendelics Análise Genômica.

Olufunmilayo I. Olopade is co-founder at CancerIQ; serves as scientific advisor at Tempus; and has received research funding from Color Genomics and Roche/Genentech.

José Eduardo Krieger, Yonglan Zheng, Dezheng Huo, Simone Maistro and Maria Aparecida Koike declare no competing interests.

Author Contributions

Generated Main Data: Flávia Eichemberger Rius, Danilo Viana, Júlia Salomão, Laila Gallo, Renata Freitas, Cláudia Bertolacini, Lucas Taniguti, Danilo Imparato, Flávia Antunes, Gabriel Sousa, Renan Achjian, Eric Fukuyama, David Schlesinger.

Analyzed Data: Flávia Eichemberger Rius, Rodrigo Guindalini, Danilo Viana, Lucas Taniguti, Danilo Imparato, Flávia Antunes, Gabriel Sousa, Renan Achjian, Eric Fukuyama, Yonglan Zheng, Dezheng Huo, Olufunmilayo I. Olopade, Maria Aparecida Koike, David Schlesinger.

Other Contributions: Cleandra Gregório, Iuri Ventura, Juliana Gomes, Nathália Taniguti, Simone Maistro, José Eduardo Krieger.

References

1. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence

452 and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**,
453 209–249 (2021).

454 2. Instituto Nacional de Câncer. *Estimativa 2023 : incidência de câncer no Brasil*.
455 (Ministério da Saúde, 2023).

456 3. Nielsen, F. C., van Overeem Hansen, T. & Sørensen, C. S. Hereditary breast and
457 ovarian cancer: new genes in confined pathways. *Nat. Rev. Cancer* **16**, 599–612
458 (2016).

459 4. Guindalini, R. S. C. *et al.* Detection of germline variants in Brazilian breast cancer
460 patients using multigene panel testing. *Sci. Rep.* **12**, 4190 (2022).

461 5. Shiovitz, S. & Korde, L. A. Genetics of breast cancer: a topic in evolution. *Ann.*
462 *Oncol.* **26**, 1291–1299 (2015).

463 6. Melchor, L. & Benítez, J. The complex genetic landscape of familial breast cancer.
464 *Hum. Genet.* **132**, 845–863 (2013).

465 7. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify
466 individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224
467 (2018).

468 8. Mavaddat, N. *et al.* Polygenic risk scores for prediction of breast cancer and breast
469 cancer subtypes. *Am. J. Hum. Genet.* **104**, 21–34 (2019).

470 9. Zhang, H. *et al.* Genome-wide association study identifies 32 novel breast cancer
471 susceptibility loci from overall and subtype-specific analyses. *Nat. Genet.* **52**,
472 572–581 (2020).

473 10. Morra, A. *et al.* Association of germline genetic variants with breast cancer-specific
474 survival in patient subgroups defined by clinic-pathological variables related to

tumor biology and type of systemic treatment. *Breast Cancer Res.* **23**, 86 (2021).

11. Mars, N. *et al.* Genome-wide risk prediction of common diseases across ancestries in one million people. *Cell Genomics* **2**, None (2022).

12. Salzano, Freire-Maia, F. M. N. As origens. in *Populações Brasileiras: Aspectos Demográficos, Genéticos e Antropológicos* (1967).

13. Souza, A. M. de, Resende, S. S., Sousa, T. N. de & Brito, C. F. A. de. A systematic scoping review of the genetic ancestry of the Brazilian population. *Genet. Mol. Biol.* **42**, 495–508 (2019).

14. Naslavsky, M. S. *et al.* Whole-genome sequencing of 1,171 elderly admixed individuals from São Paulo, Brazil. *Nat. Commun.* **13**, 1004 (2022).

15. Sudlow, C. *et al.* UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).

16. Pedersen, B. S. *et al.* Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. *Genome Med.* **12**, 62 (2020).

17. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

18. Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J. & Delaneau, O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* **53**, 120–126 (2021).

19. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, (2020).

20. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and

richer datasets. *Gigascience* **4**, 7 (2015).

21. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

22. Kuhn, M. Building Predictive Models in *R* Using the caret Package. *J. Stat. Softw.* **28**, (2008).

23. R Foundation for Statistical Computing. *R: A Language and Environment for Statistical Computing*. (https://www.R-project.org/, 2023).

24. Aragon, T. J., Fay, M. P., Wollschlaeger, D. & Omidpanah, A. *epitools: Epidemiology Tools. Tools for training and practicing epidemiologists including methods for two-way and multi-way contingency tables.* (CRAN, 2020).

25. Hu, C. *et al.* A Population-Based Study of Genes Previously Implicated in Breast Cancer. *N. Engl. J. Med.* **384**, 440–451 (2021).

26. Liu, C. *et al.* Generalizability of polygenic risk scores for breast cancer among women with european, african, and latinx ancestry. *JAMA Netw. Open* **4**, e2119084 (2021).

27. Wasik, K. *et al.* Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics. *BMC Genomics* **22**, 197 (2021).

28. Li, J. H., Mazur, C. A., Berisa, T. & Pickrell, J. K. Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. *Genome Res.* **31**, 529–537 (2021).

29. Chaubey, A. *et al.* Low-Pass Genome Sequencing: Validation and Diagnostic Utility from 409 Clinical Cases of Low-Pass Genome Sequencing for the Detection of Copy Number Variants to Replace Constitutional Microarray. *J. Mol. Diagn.* **22**,

823–840 (2020).

30. Bahcall, O. Common variation and heritability estimates for breast, ovarian and prostate cancers. *Nat. Genet.* (2019) doi:10.1038/ngicogs.1.

31. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).

32. Monticciolo, D. L., Newell, M. S., Moy, L., Lee, C. S. & Destounis, S. V. Breast Cancer Screening for Women at Higher-Than-Average Risk: Updated Recommendations From the ACR. *J. Am. Coll. Radiol.* (2023) doi:10.1016/j.jacr.2023.04.002.