Prediction of melanoma incidence based on combination of genetic variants

Georgios Ntritsos and Evangelos Evangelou
Department of Hygiene and Epidemiology Ioannina, University of Ioannina School of Medicine
Ioannina, Greece
Emails: gntritsos@uoi.gr, vangelis@uoi.gr

Abstract: The occurrence of melanoma is a composite process that implicates the interaction of phenotypic, environmental, and genetic risk factors. We constructed genetic risk models, with the aim to assess their predictive performance on melanoma risk.

Summary level data from the largest meta-analysis of genome-wide association studies for melanoma, up to date, were used for the construction of weighted genetic risk scores. We used six different p-value thresholds for genetic variants inclusion. We evaluated our genetic risk scores in 2,862 events of incident melanoma and 321,789 cancer-free controls from the UK Biobank, a prospective cohort study of 500,000 participants. Using AUCs, we compared the predictive ability of the different genetic risk scores.

Genetic risk scores were strongly associated melanoma risk. Odds Ratios ranged from 1.478 to 1.528. The predictive ability of the genetic risk scores ranged from 0.6234 to 0.6328 showing a moderate performance.

Our study suggests that when the p-value threshold for genetic variants inclusion become more tolerant, the prediction performance of the model improved. Validation of the results in larger populations, as well as Southern European populations is needed.

Keywords—Melanoma, Genetic Risk Score, AUC, Prediction, UK Biobank

I. INTRODUCTION

Melanoma, the most malignant type of skin malignancy, is one of the most common neoplasms with an increasing worldwide incidence of 300,000 cases each year [1]. Melanoma arises due to malignant transformation of melanocytes. When detected in its earliest stages, is highly curable and has a five-year rate of survival at about 98% [2], thus early intervention can reduce mortality, morbidity and health care cost. The aetiology of melanoma is complex and additional research is needed to discover the role of modifiable risk factors on melanomagenesis to reinforce primary prevention strategies.

Environmental exposures (e.g. UV exposure) [3], phenotypic attributes, family history [4] and genetic factors [5] are involved in the development of melanoma.

This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the project "Reinforcement of Postdoctoral Researchers - 2nd Cycle" (MIS-5033021), implemented by the State Scholarships Foundation (IKY).

Genome-wide association studies (GWAS) and candidate gene studies [6] supported the role of genetics in the risk of melanoma. Recently, a large GWAS accumulating 36, 760 cases and 375,188 melanoma-free controls from the UK, USA, Australia, Northern and Western Europe, and the Mediterranean, identified 54 genetic loci associated with the risk of melanoma at the level of genome-wide significance [7].

Currently, the most empirical way to provide an estimate of the genetic predisposition to a disease at the individual level, is using genetic risk scores (GRSs) [8]. GRSs are using the findings from GWASs and combine the effects of numerous single nucleotide polymorphisms (SNPs) into a single score. The GRS, for each individual, is a sum of the effects of the alleles of risk corresponding to a disease, with each one of these alleles weighted by its effect size, as it is estimated from an independent GWAS on the disease. The primary aim of the GRS is to predict the chance of an individual to be affected by a disease and the secondary aim, to aware individuals at high risk, adhering to certain treatments or specific behavioral and lifestyle modifications

The purpose of this study was to construct several GRS models and assess their predictive performance on risk of melanoma. The effect estimates of each genetic variant on melanoma were derived from the largest GWAS meta-analysis of melanoma [7] and we validate the prediction ability of GRSs in UK Biobank, a population-based cohort with participants of European descent [9].

II. METHODS

A. Study population

For the development of the GRSs we used summary level data from the largest GWAS meta-analysis of melanoma up to date consisted of 36,760 cases of melanoma and 375,188 melanoma-free controls from UK, USA, Australia, Northern and Western Europe and Mediterranean [7]. We evaluated our GRSs using data at individual level from the UK Biobank. UK Biobank is a long-term prospective study having recruited ~500,000 people in UK, aged 40–69 years when they joined the study [9]. Participants attended assessment centers at which baseline data were collected on their medical history, lifestyle, body composition and environment. DNA for genotyping was collected as well. For the present study we used a subset of unrelated European-ancestry individuals. In more detail, from the initial set of participants, we excluded individuals without available genetic data and individuals of non-European ancestry. Also, we removed one of each pair of 1st and 2nd degree relatives by using the centrally provided kinship data from UK Biobank [10].

978-1-6654-2742-5/21/$31.00 ©2021 IEEE

B. Melanoma incidence in UK Biobank

Data on cancer incidence in UK Biobank were available for each participant through linkage to the Central Registers of the UK National Health Service (NHS). We identified 2,862 unrelated Europeans with full genetic data available, whose primary cancer diagnosis was melanoma according to ICD-9 (172) and ICD-10 (C43) codes. From the database we selected 321,568 controls of European descent that had never had a diagnosis of cancer, either self - reported and were not registered in the national cancer registry. We therefore included 324,651 participants in our analysis.

C. Genotyping and imputation in UK Biobank

We used quality controlled genotypes from 488,377 UK Biobank participants. DNA samples of 438,427 participants were genotyped at 825,427 variants using a custom Affymetrix UK Biobank Axiom Array chip and DNA samples of the other 49,950 participants were genotyped at 807,411 variants using a custom Affymetrix UK BiLEVE Axiom Array chip from the UK BiLEVE study [11]. A merged reference panel from the 1000 Genomes Phase 3 panel and the UK10K and, in addition, the Haplotype Reference Consortium (HRC) panel were used for the central imputation of the variants [10]. UK Biobank centrally computed genetic principal components in order to account for population stratification

D. Genetic Risk Score

GRS constitutes a sum across SNPs of the number of the alleles of risk at that SNPs, weighted by their effect estimates (beta coefficients). The betas were obtained from meta-analyses of the GWAS described above [11]. For each individual *i* in the UK Biobank the GRS was calculated as:

$$GRS_i = \sum_{j=1}^{K} b_j \cdot g_{ij}$$

where b_j is the beta coefficient for the SNP *j*, g_{ij} is the value of the genotype for SNP *j* for individual *i* (0/1/2, depending on the number of the risk alleles for the certain SNP).

We developed six GRS models, where SNP inclusion was based on different p-value thresholds. The p-value thresholds were chosen as $5x10^{-7}$, $1x10^{-7}$, $5x10^{-8}$, $1x10^{-8}$, $5x10^{-9}$ and $1x10^{-9}$. Values $>5x10^{-8}$ are generally considered genome-wide significant in the field of genetic epidemiology even though more stringent thresholds have been proposed [12]. Clumping was performed for each one of the GRS models, with the aim to keep a subset of SNPs per genetic region that are nearly uncorrelated with each other [13]. More specifically, clumping was performed by selecting the most significant SNP and removing from examination all SNPs in LD with this index SNP ($r^2<0.25$) and withdrawing correlated SNPs up to 250 kilobase pairs from the index SNP. All the GRSs were standardized per unit increase in the cancer-free population. The construction of the GRSs was performed using PLINK 2.0 [14].

E. Statistical Analysis

GRSs were analyzed with respect to melanoma events using logistic regression models and we calculated odds ratios (OR) and 95% confidence intervals (CIs). Logistic regression is a classification algorithm, that is used to forecast a dichotomous outcome based upon a set of non-dependent variables. The logistic regression equation for our dataset is calculated as:

$$Y_i = \frac{e^{bo + b1 \cdot x}}{1 + e^{bo + b1 \cdot x}}$$

where Y_i is the predicted probability of an individual *i* to be a melanoma case, $b0$ is the intercept term of the regression and $b1$ is the regression coefficient for the single input value (x). The equation above can be transformed to

$$log\left(\frac{p_i}{1 - p_i}\right) = bo + b1 \cdot x$$

The term on the parenthesis is called the odds and is a ratio of the probability of the event divided by the probability of not the event. Logistic regression analyses were adjusted for the first ten genetic principal components in order to correct for possible population stratification. The predictive ability of the GRSs were assessed in the UK Biobank participants by calculating the Area Under the Curve (AUC) with bootstrap CIs These bootstrap CIs were calculated after creating 1000 resamples from our population and computing AUC for each one of these samples.Statistical analyses were performed using the statistical package Stata 14.0 software.

III. RESULTS

A. Population Characteristics of validation dataset

Our validation sample of 324,651 individuals consisted of 169,868 women (52.32%) and 154,783 men (47.68%), with a mean age at recruitment of 56.09 years. The 2,862 melanoma incident events had a mean age of 58.57 years and 1,674 (58.49%) were women (Table I).

TABLE I. Demographic characteristics of our sample

	All participants (N= 324,651)	**Melanoma cases (N= 2,862)**	**Cancer-free individuals (N= 321,789)**
Females, N (%)	169,868 (52.32)	1,674 (50.49)	168,194 (52.27)
Age, mean (SD) years	56.08 (8.01)	58.57 (7.49)	56.06 (8.01)

B. *Associations between GRSs and the risk of melanoma*

As expected, all GRS were strongly associated with melanoma risk in the testing dataset of UK Biobank. When setting the p-value threshold at $5x10^{-7}$, a total of 377 SNPs were included in the model. The OR was 1.528 (95% CI 1.478-1.580) per standard deviation increase. For a p-value threshold for inclusion at $1x10^{-7}$, a total of 318 SNPs were included in the GRS model. The OR was 1.514 (95% CI 1.464-1.566) per standard deviation increase. While setting the p-value threshold at $5x10^{-8}$, 300 SNPs were used for the GRS construction and the OR was 1.509 (95% CI 1.460-1.561). Assuming a p-value threshold at $1x10^{-8}$, a total of 259 SNPs were included in the GRS, and the OR was 1.494 (95% 1.445-1.545) per standard deviation increase. Finally, when making the p-value threshold even stricter, that is $5x10^{-9}$ and $1x10^{-9}$, 243 and 216 SNPs were used for the GRSs construction, respectively, and the ORs were 1.490 (95% 1.441-1.540) and 1.478 (95% CI 1.430-1.528) respectively (Table II).

TABLE II. Association between GRSs and risk of melanoma

Threshold for SNPs inclusion	# SNPs included	OR* (95% CI)	p-value
p-value< $5x10^{-7}$	377	1.528 (1.478, 1.580)	$9.55x10^{-137}$
p-value< $1x10^{-7}$	318	1.515 (1.464, 1.566)	$5.80x10^{-131}$
p-value< $5x10^{-8}$	300	1.509 (1.460, 1.561)	$5.86x10^{-129}$
p-value< $1x10^{-8}$	259	1.494 (1.445, 1.545)	$2.40x10^{-123}$
p-value< $5x10^{-9}$	243	1.490 (1.441, 1.540)	$1.32x10^{-121}$
p-value< $1x10^{-9}$	216	1.478 (1.430, 1.528)	$4.66x10^{-117}$

*ORs were adjusted for ten the first ten genetic principal components

C. *Assessing the predictive performance of the GRSs*

We evaluated the predictive performance of the 6 GRSs in the testing dataset of UK Biobank. When setting the p-value threshold at $5x10^{-7}$, the model's AUC was 0.6328 (95% CI 0.6225-0.6430) (Fig. 1a).

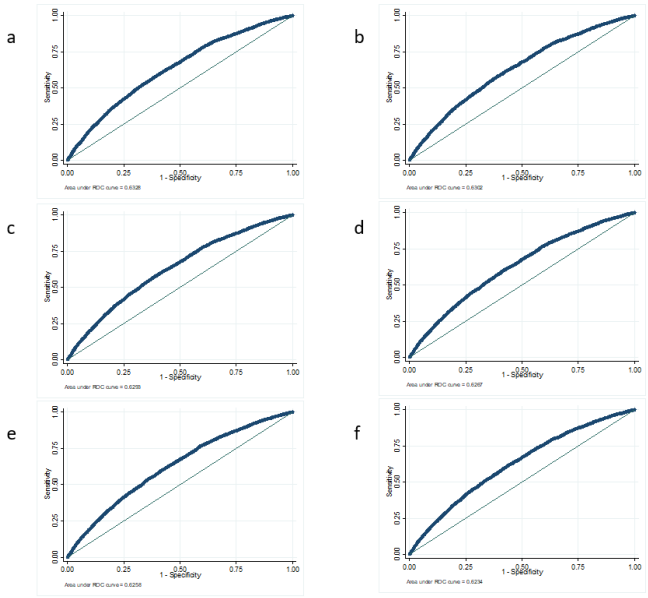

Figure 1. AUCs for the different p-value thresholds

For a p-value threshold at $1x10^{-7}$, the AUC was 0.6302 (95% CI 0.6202-0.6402) (Fig. 1b). While setting the p-value threshold at $5x10^{-8}$, the AUC was 0.6293 (95% CI 0.6198-0.6387) (Fig. 1c). For the p-value threshold at $1x10^{-8}$ the AUC was 0.6267 (95% CI 0.6168-0.6365) (Fig. 1d). For the p-value threshold at $5x10^{-9}$ the AUC was 0.6258 (95% CI 0.6138-0.6344) (Fig. 1e). Finally, for the p-value threshold at $1x10^{-8}$ the AUC was 0.6234 (95% CI 0.6133-0.6334) (Fig. 1f) (Table III).

TABLE III. Predictive ability of GRSs

p-value threshold	AUC	95% Bootstrap CI
p-value< $5x10^{-7}$	0.6328	0.6225, 0.6430
p-value< $1x10^{-7}$	0.6302	0.6202, 0.6402
p-value< $5x10^{-8}$	0.6293	0.6198, 0.6387
p-value< $1x10^{-8}$	0.6267	0.6168, 0.6365
p-value<$5x10^{-9}$	0.6258	0.6138, 0.6344
p-value< $1x10^{-9}$	0.6234	0.6133, 0.6334

IV. CONCLUSION

In this study we thoroughly evaluated the predictive ability of six melanoma GRSs in a testing dataset of 2,862 melanoma cases and 321,789 controls. We compared six GRSs models using various p-value thresholds for variants inclusion. The number and the effect size of the SNPs, as well as, the sample size of the training dataset are factors, on which the ideal p-value threshold for SNPs inclusion for a disease risk, depends [15]. We observed that all GRSs were strong predictors of melanoma incidence, with ORs ranged from 1.478 to 1.528, showing the strong association of the genetic predisposition with the risk of melanoma. The strongest association was observed when the p-value threshold was more tolerant, that is $5x10^{-7}$, including 377 SNPs. Moreover, the predictive ability of those GRSs ranged from 0.6234 to 0.6328, showing a moderate discrimination power. The best predictive performance was observed when the p-value threshold was set to $5x10^{-7}$, showing that, when more SNPs were included in the model, the predictive ability on melanoma incidence was improved even though these SNPs are not necessarily considered genome-wide significant. It has been proved that although the effects of single SNPs on several occasions do not reach the genome-wide significant thresholds, due to various reasons, such as lack of power, the combined effect of those SNPs into a score, could enhance the prediction of the disease risk [16]. This suggests the polygenic aetiology that implicit the risk melanoma.

Our study has several strengths. For GRSs construction, we used as training set, the largest melanoma GWAS data to date, a crucial factor of the accuracy of GRS prediction [14]. We intensively examined different selection criteria for variants inclusion and compared their predictive performance on melanoma incidence. A limitation in our study, was the lack of more summary data, which did not allow us to examine more models with less strict p-value threshold for inclusion.

In conclusion, our study suggests that when examining the predictive ability of prognostic models, that contain several SNPs, on melanoma incidence, it appears that when the p-value threshold for SNPs inclusion become more tolerant, the prediction performance of the model is improved. Further

research needs to be accomplished in datasets with more melanoma cases and even more lenient p-value thresholds. Nonetheless, most of the developed GRSs include only genome-wide significant SNPs, as they largely increase the chance of integrating true positive signals, several studies have shown that GRSs including millions of SNPs far outperform GRSs build on the most strongly association SNPs showing substantially greater predictive power [17]. Validation of the results in larger populations, as well as Southern European populations is needed.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, and A. Jemal. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018 Nov;68(6):394-424.

[2] R.L. Siegel, K.D. Miller, and A. Jemal, Cancer statistics, 2017. CA: A Cancer Journal for Clinicians, 2017. **67**(1): p. 7-30.

[3] S. Gandini, M. Montella, F. Ayala, L. Benedetto, C.R. Rossi, A. Vecchiato, M.T. Corradin, V. DE Giorgi, P. Queirolo, G. Zannetti, et al. Sun exposure and melanoma prognostic factors. Oncol Lett. 2016 Apr;11(4):2706-2714.

[4] L. Titus-Ernstoff, A.E. Perry, S.K. Spencer, J.J. Gibson, B.F. Cole, and M.S. Ernstoff. Pigmentary characteristics and moles in relation to melanoma risk. Int J Cancer. 2005 Aug 10;116(1):144-9.

[5] K. Antonopoulou, I. Stefanaki, C.M. Lill, F. Chatzinasiou, K.P. Kypreou, F. Karagianni, E. Athanasiadis, G.M. Spyrou, J.P.A. Ioannidis, L. Bertram, E. Evangelou, A.J. Stratigos. Updated field synopsis and systematic meta-analyses of genetic association studies in cutaneous melanoma: the MelGene database. J Invest Dermatol. 2015 Apr;135(4):1074-1079.

[6] E. Evangelou, and A.J. Stratigos. Lessons from genome-wide studies of melanoma: towards precision medicine. Expert Review of Precision Medicine and Drug Development 1, 443-449 (2016).

[7] 9M.T. Landi, D.T. Bishop, S. MacGregor, M.J. Machiela, A.J. Stratigos, P. Ghiorzo, M. Brossard, D. Calista, J. Choi, M.C. Fargnoli, et al. Genome-wide association meta-analyses combining multiple risk phenotypes provide insights into the genetic architecture of cutaneous melanoma susceptibility. Nat Genet. 2020 May;52(5):494-504.

[8] 7N. Amin, CM. van Duijn, A.C. Janssens. Genetic scoring analysis: a way forward in genome wide association studies? Eur J Epidemiol. 2009;24(10):585-7.

[9] 8C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, 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. 2015 Mar 31;12(3):e1001779.

[10] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, *et al.* The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018 Oct;562(7726):203-209.

[11] L.V. Wain, N. Shrine, S. Miller, V.E. Jackson, I. Ntalla, M. Soler Artigas, C.K. Billington, A.K. Kheirallah, R. Allen, J.P. Cook, *et al*. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. Lancet Respir Med. 2015 Oct;3(10):769-81. doi: 10.1016/S2213-2600(15)00283-0. Epub 2015 Sep 27. Erratum in: Lancet Respir Med. 2016 Jan;4(1):e4. Erratum in: Lancet Respir Med. 2016 Jan;4(1):e4.

[12] S.L. PuliT, S.A. de With, P.I. de Bakker. Resetting the bar: Statistical significance in whole-genome sequencing-based association studies of global populations. Genet Epidemiol. 2017 Feb;41(2):145-151.

[13] B.A. Goldstein, L. Yang, E. Salfati and T.L. Assimes. Contemporary Considerations for Constructing a Genetic Risk Score: An Empirical Approach. Genet Epidemiol. 2015;39(6):439-445.

[14] C.C. Chang, C.C. Chow, L.C.A.M. Tellier, S. Vattikuti, S.M. Purcell and J.J. Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets, GigaScience, Volume 4, Issue 1, December 2015, s13742–015–0047–8

[15] F. Dudbridge. Power and predictive accuracy of polygenic risk scores. PLoS Genet. 2013 Mar;9(3):e1003348.

[16] M.A. Simonson, A.G. Wills, M.C. Keller, & M.B. McQueen. Recent methods for polygenic analysis of genome-wide data implicate an important effect of common variants on cardiovascular disease risk. BMC Med Genet 12, 146 (2011).

[17] A.V. Khera, M. Chaffin, K.H. Wade, S. Zahid, J. Brancale, R. Xia, M. Distefano, O. Senol-Cosar, M.E. Haas, A. Bick, *et al*. Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. Cell. 2019 Apr 18;177(3):587-596.e9.