Polygenic risk scores for cervical HPV infection, neoplasia and cancer show potential for personalised screening: Comparison of two methods

Anna Tisler (anna.tisler@ut.ee)
Institute of Family Medicine and Public Health, University of Tartu

Anneli Uuskula
Institute of Family Medicine and Public Health, University of Tartu

Sven Erik Ojavee
Department of Computational Biology, University of Lausanne

Kristi Läll
Estonian Genome Centre, Institute of Genomics, University of Tartu

Estonian Biobank research team
Estonian Genome Centre, Institute of Genomics, University of Tartu

Triin Laisk
Estonian Genome Centre, Institute of Genomics, University of Tartu

Article

Keywords:

Posted Date: March 23rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2648624/v1

Abstract

The era of precision medicine requires the achievement of accurate risk assessment. Polygenic risk scores (PRSs) have strong potential for increasing the benefits of nationwide cancer screening programs. The current pool of evidence on the role of a PRS as a risk stratification model in actual practice and implementation is limited. To better understand the impact of possible method-induced variance, we constructed and validated two PRSs for cervical cancer (CC) using the Estonian Biobank female population (691 CC cases and 13 820 controls) and evaluated their utility in predicting incident cervical neoplasia (CIN), cancer, and human papillomavirus (HPV) infection using two methods (LDPred and BayesRR-RC). This study demonstrated that two genetic risk scores were significantly associated with CIN, CC, and HPV infection incidence. Independent of the method, we demonstrated that women with elevated PRS values reached the observed cumulative risk levels of CIN or CC much earlier. Our results indicated that the PRS-based discrimination rules could differ substantially when the PRSs contain similar predictive information. In summary, our analysis indicated that PRSs represent a personalized genetic component that could be an additional tool for cervical cancer risk stratification, and earlier detection of abnormalities provides invaluable information for those at high risk.

Introduction

Cervical cancer is the fourth most frequently diagnosed cancer and the fourth leading cause of cancer death among women, with an estimated 604,000 new cases and 342,000 deaths worldwide in 2020[1]. Cervical cancer cumulative risk among women up to age 70 in Eastern Europe is 1.4%, which is higher than that in other high human development index (HDI) countries (1.3%) and more than twice as high as that in Western Europe (0.67%) [2]. Persistent infection with high-risk HPV (hrHPV) is proven to be a causal and necessary factor for cervical cancer development and its preceding lesion. It has been estimated that the average lifetime probability of HPV among those with at least one opposite-sex partner is 84.6%, but the risk of HPV infection progressing to cervical cancer varies according to HPV genotype, preventive behaviour and health risk factors [3, 4]. Women’s behavioural and sexual characteristics associated with a higher risk of HPV infection acquisition, persistence, and progression to precancerous and more advanced cancerous stages are well described [5]. The main risk factors reported are age at first intercourse, hormonal contraception use, number of sexual partners, parity, and smoking. However, it is acknowledged that both various exposures and heritable factors contribute to cervical cancer development [6]. Twin and family studies have estimated the heritability of cervical cancer to be 22–64%, while the common variant heritability (proportion of phenotypic variance explained by common variants) is estimated to be as high as 36% [7]. The genetic component in the HPV-lesion-cervical cancer relationship is understudied and is supported by a modest number of studies.

Genome-wide association studies (GWASs) have become a valuable tool to describe the genetic basis for common human diseases, and in line with this, they have also identified susceptibility loci for cervical cancer [8]. Polygenic risk scores (PRSs) combine the effects of several genetic variants into one variable that can be used to assess the genetic risk of a disease for an individual. Therefore, PRSs allow grouping participants into different risk categories for disease and are also used as a covariate in epidemiological analyses. There are a number of methods for PRS calculation, and the methods differ in terms of two key criteria: which genetic variants to include and what weights to allocate to them. Often, when new PRS methods are introduced, comparisons are made between a limited set of methods, together with application to some real data examples, since there is a need to explore and quantify the variability of PRS values derived using different estimation methods on the same target sample [9].

Although recent GWASs have begun to clarify the genetic background of cervical cancer and preceding HPV infection, further studies explaining genetic susceptibility for prevalent HPV infection and whether there is an overlap in genetic factors for HPV infection and progression of cervical disease (CIN, CC) are needed. Here, to understand the impact of method-induced variance on genomic prediction of cervical cancer and HPV status, we compared two methods (LDPred and BayesRR-RC) for cervical cancer PRSs.

Results

We identified 885 cervical cancer cases (overall mean age at recruitment 51.7 years, SD 13.4), 4,406 CIN cases (mean age at recruitment 38.4 years, SD 10.7), and 83,065 controls (mean age at recruitment 42.6 years, SD 14.2). We first used the prevalent cervical cancer cases (n = 691) and controls (n = 13,820) to select the best-performing PRSs for subsequent analyses, and these individuals were removed from further analyses.

Selecting the best-performing PRSs

We evaluated a total of 14 PRSs calculated with two separate methods to select the best-performing PRS for each method. According to our analyses (Supplementary Table 1), in Score 1, the best score was for LDpred_p3.0000e.03 (OR 1.44, 95% CI 1.33–1.56), which included 2,894,555 variants (causal fraction 0.3%). In Score 2, BayesRR-RC showed the strongest association (OR 1.44, 95% CI 1.33–1.56). In further analyses, we shall refer to these two PRSs as Score 1 and Score 2, respectively.

In the following analyses, the remaining cases/controls were divided as follows: incident cancer 194 cases and 69,245 controls with a mean age of 45.7 (SD 13.6) and 42.6 (SD 14.3) years, respectively; CIN 1,009 cases and 35,275 controls with a mean age of 31.7 (SD 9.8) and 42.6 (SD 14.2) years, respectively; and prevalent CIN 3,397 cases and 33,970 controls with a mean age of 40.0 (SD 10.0) and 42.6 (SD 14.2) years, respectively. Data on 1,347 women for association analysis with HPV infection were used (Fig. 1).

PRS association with CIN

We found that both risk scores were significantly associated with prevalent CIN status in the case–control subset of the EstBB cohort.

As found in the previous step, Scores 1 and 2 performed relatively equally in association with prevalent cervical cancer status. The same applied with respect to prevalent CIN with an OR = 1.32 per SD, 95% CI

1.27–1.38, $p = 1.1 \times 10^{-44}$ with Score 1 and 1.32 (95% CI 1.27–1.37), $p = 1.3 \times 10^{-42}$ with Score 2.

PRS association with incident CC/CIN

Next, we evaluated the performance of the PRSs for incident CC or CIN in EstBB. Both PRSs were associated with both conditions ($p < 0.05$).

For CC, the risk increased 1.32-fold per 1-SD increase in the Score 1 PRS (Harrell’s C-statistic of 0.581, SE 0.020). Score 2 showed a slightly lower HR of 1.25 (Harrell’s C-statistic of 0.566, SE 0.022). On the other hand, Score 2 had a slightly higher HR for CIN of 1.37 (Harrell's C-statistic 0.59, SE 0.009) compared to Score 1 with an HR of 1.34 (Harrell's C-statistic 0.582, SE 0.009). Although we found nominal differences implying that Score 2 might better reflect the CIN risk and Score 1 better reflect the CC risk, it should be noted that the differences were very small, and the clinical significance of those differences is outside the scope of this study.

Women in the highest 20% of genetic risk were estimated to have a 2.32 (Score 2) to 2.50 (Score 1) times greater risk of developing CC than women in the lowest 20% (Table 2). The effect was less pronounced when comparing the top 20% of women with the women below the median, resulting in a 1.58 (Score 2) to 1.66 (Score 1) times greater risk for the top 20% of women. A similar effect was observed when comparing the top 20% of women with the rest, giving hazard ratios from 1.49 (Score 1) to 1.60 (Score 2) (Table 2). Similar to CC, a clear risk gradient was observed within the risk categories for CIN. Women in the top 20% of genetic risk had an HR of 2.38 (Score 2) to 2.50 (Score 1) for incident CIN compared to women in the bottom 20%, HR of 1.66 (Score 1) to 1.91 (Score 2) compared to women below the median and HR of 1.49 (Score 1) to 1.62 (Score 2) compared to the rest of the cohort (Table 2).

Table 2
Hazard ratios of incident cervical cancer and cervical intraepithelial neoplasia for the two evaluated genetic risk scores

	Cervical cancer		**Cervical intraepithelial neoplasia**	
	Score 1 HR (95% CI)	Score 2 HR (95% CI)	Score 1 HR (95% CI)	Score 2 HR (95% CI)
Top 40% versus remaining	1.59 (1.20–2.11)	1.42 (1.07–1.88)	1.61 (1.43–1.82)	1.72 (1.51–1.95)
Top 20% versus remaining	1.49 (1.08–2.04)	1.60 (1.17–2.19)	1.62 (1.42–1.86)	1.68 (1.46–1.93)
Top 20% versus the bottom 20%	2.50 (1.51–4.14)	2.32 (1.43–3.74)	2.38 (1.93–2.95)	2.42 (1.95-3.00)
Top 20% versus below the median	1.66 (1.17–2.35)	1.58 (1.13–2.21)	1.91 (1.64–2.22)	1.98 (1.70–2.32)
Top 10% versus remaining	1.40 (0.93–2.11)	1.62 (1.10–2.39)	1.56 (1.31–1.86)	1.71 (1.44–2.03)
Top 5% versus remaining	1.63 (0.96–2.76)	1.69 (1.01–2.82)	1.68 (1.34–2.11)	2.00 (1.61–2.48)

As seen in Fig. 2, the cumulative incidence of CC by age 70 was estimated to be 5.3% (95% CI 3.7–6.8) for women in the top 20% of genetic risk (as defined using Score 1), 3.7% (95% CI 2.9–4.3) for those between the 20-80th percentiles and 1.8% (95% CI 0.9–1.8) for those in the lowest 20%. The cumulative incidence in risk categories defined using Score 2 was similar (5.3%, 3.5%, and 2.4%, respectively) (Fig. 2b).

As seen in Fig. 3a, the cumulative incidence of CIN by age 50 was estimated to be 37.1% (95% CI 33.3–40.7) for women in the top 20% of genetic risk, while it was 17.2% (95% CI 14.0-20.3) among women in the bottom 20% with Score 1. The results of Score 2 (Fig. 3b) were similar, with a cumulative incidence of 37.4% (95% CI 33.6–40.9) for the top 20%.

Correlation of PRSs

The Pearson correlation between Scores 1 and 2 was 0.76. We then divided all women into two categories (high: PRS in the top 5%, not high: everyone else) based on the two PRSs. Eight percent of women belonged to the high category with at least one PRS, while 1.9% were in the top 5% with both compared PRSs (Fig. 4). Even though the scores were strongly correlated, we observed that the individual classification into the top 5% risk score category depended on a selected score and often did not overlap for a single individual. We also combined Score 1 and Score 2 into a further score called metaPRS (see Methods). When analysing the metaPRS in association with incident CC and CIN using the Cox proportional hazards model, the results mirrored those from the analysis of individual scores (HR 1.31 (SE 0.07), C-statistic 0.578 (SE 0.021); thus, additional results are not shown.

Associations of risk scores with predictors of high-risk HPV infection

.Both PRSs were significantly associated with high-risk HPV (hrHPV) infection, giving an adjusted OR of 1.28 (95% CI 1.11–1.48 for Score 1, 95% CI 1.10–1.47 for Score 2) (Supplementary Tables 2, 3). We further quantified the effect of nongenetic HPV risk factors while adjusting for the PRS value, hence enabling hrHPV risk estimation conditional on genetic factors. Several nongenetic risk factors were associated with hrHPV infection (Scores 1 and 2, respectively): being single, OR 2.71 (95% CI 1.68–4.38) and OR 2.65 (95% CI 1.64–4.30), lower education level, OR 1.44 (95% CI 1.05–1.98) and OR 1.42 (95% CI 1.01–1.98) and increased number of sexual partners, OR 1.04 (95% CI 1.01–1.06) and OR 1.04 (95% CI 1.02–1.07). Overall, the estimates of the nongenetic HPV risk factors were similar regardless of the score used for adjustment.

Discussion

In our study, we demonstrated that two genetic risk scores calculated using different approaches were significantly associated with CC and CIN status in the case–control subset of our cohort. While on average, approximately 1% of women in our dataset were diagnosed with CC by the age of 70, women in the highest five percentiles of our tested PRSs reached the same cumulative risk level by age 55, 15 years earlier. Similarly, on average, approximately 30% of women in our dataset were diagnosed with CIN by the age of 70, but women in the top 20% of genetic risk reached the same cumulative incidence before their 40th birthday. Our results suggest that genetic risk estimation could be an additional tool for CC risk stratification in clinical practice, either for targeted screening or prevention practices. However, there are certain aspects that need to be considered and that are discussed in more detail below. In addition to our main findings, both tested genetic risk scores were also strong predictors of hrHPV infection, comparable to known risk factors such as marital status and the number of partners. Since the genetic risk score summarizes all the genetic risk factors for CC, we speculate that the same genetic factors associated with CC susceptibility are also associated with hrHPV infection, providing insight into HPV genetic susceptibility, which has thus far remained poorly characterized. Our results are in line with the findings by Koel et al. [10], who showed that a large part of the predictive power of PRSs for CC comes from the HLA fraction. HLA-related signalling, on the other hand, plays a central role in the course of HPV infection and may determine whether the infection is successfully cleared or persists and develops into a malignant lesion. It is possible that the PRSs capture different biological pathways or mechanisms, which is also supported by our results, in which the two scores showed very similar results in the analysis of prevalent cases but different results in the analysis of incident cases. Hence, we encourage drawing comparisons separately for prevalent and incident cases, as this could pinpoint different aspects of genetic risk prediction.

The correlation between the two PRSs was substantial (Pearson correlation of 0.76), which is expected given the overlap in the datasets used to estimate genome-wide effects for SNPs that were then later used to develop the scores (namely, UK Biobank data). The main difference in terms of methods is that Score 1 was constructed using multiple datasets combining many marginal SNP effect estimates (one SNP at a time), whereas Score 2 was obtained using a single dataset, but estimates were retrieved jointly (all SNP effects were estimated in one model). Score 1 could better leverage the heterogeneity in samples, and Score 2 could better leverage the genetic architecture (for example, LD structure) implications on genetic prediction, provided the training set (UK Biobank) and test set (Estonian Biobank) have relatively similar genetic architecture profiles. Nevertheless, the similarity of these two scores was further confirmed by the fact that the metaPRS combining Scores 1 and 2 did not yield noticeably improved predictions compared to the scores separately. Despite the large correlation, we observed that the two scores classified high-risk individuals differently. For example, 1.9% of women in our dataset were in the top 5% with both compared PRSs, indicating that approximately 60% of the individuals in the top 5% of Score 1 would not be classified as individuals in the top 5% of risk with Score 2. This demonstrates that PRS-based risk stratification could result in substantial differences across methods when identifying high-risk patients. Due to these differences, it has been suggested to provide a probability of belonging to an increased risk category instead of strict categories to account for the variability of scores [11]. More research with larger and fully characterized samples is needed to assess the utility of combining a PRS with hrHPV status and other clinical risk factors into a complex risk prediction tool.

A major strength of our study is the fact that due to the nature of data at the EstBB, we were able to include only women with a known CIN status, as we could use procedure codes in combination with disease codes to select only those women as controls who had a Pap test during the 5 years prior to this study and did not have diagnosis codes for CC or CIN. Although this approach reduced the heterogeneity of the data, it also biased the prevalence and incidence rates of the evaluated diagnoses, which means that these cannot directly be extrapolated to the general population.

Organized CC screening is a globally recommended public health policy. To date, population-based screening has not been optimal in many countries due to low participation rates. Unfavourable trends in the burden of CC require new approaches. CC screening personalization with the help of risk-based algorithms considering risk factors such as health, sexual behaviours, and genetic components could lead to a more precise and individual-based approach. It is important to develop accurate models for a more personalized approach with screening intervals based on pretest probability. A PRS combining genotypes with phenotypic profiles has been shown to improve the risk prediction of cancers [12, 13] and has the potential to considerably increase the benefits of nationwide screening programs. Our results demonstrated that a PRS for CC has similar potential and paves the way for future studies evaluating this in independent cohorts.

Materials And Methods

Data from the Estonian Biobank (EstBB) were used to compare the performance of previously published cervical cancer PRSs calculated using different approaches and to evaluate their utility in association with CC, CIN and hrHPV.

Data source (target population) and genotyping

The Estonian Biobank (EstBB) is a population-based biobank with genotype data and health information for over 200,000 participants recruited between 2002 and 2020 [14] in its latest data freeze, which represents approximately 20% of the Estonian adult population. EstBB women were followed up from the date of EstBB entry until 31.12.2019, which is the date of the last link with the main dataset during the study period. All EstBB participants were genotyped using Illumina Global Screening Array v1.0 and v2.0 at the Genotyping Core Lab of the Institute of Genomics, University of Tartu. After genotyping, PLINK format files were created using Illumina GenomeStudio v2.0.4. The exclusion criteria included an individual call rate < 95% and sex mismatch. Before imputation, variants were filtered by a call rate < 95%, HWE p value < 1e-4 (autosomal variants only), and minor allele frequency (MAF) < 1%. Prephasing was performed using Eagle v2.3 software (the number of conditioning haplotypes Eagle2 uses when phasing each sample was set to –Kpbwt = 20000), and Beagle v.28Sep18.79339 with an effective population size ne = 20,000 was used for imputation. A population-specific imputation reference of 2297 WGS samples was used [15].

The health data for EstBB participants were obtained from regular linking with the Estonian Health Insurance Fund (EHIF), the Estonian Cancer Registry (ECR), and the Causes of Death Registry, which are population-based and nationwide health/administration registries [16]. The EHIF is the core purchaser of health care services in Estonia, covering health care costs for insured people and managing services for uninsured citizens. The information on health status is stored as diagnostic codes based on the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) and codes relating to medical services and procedures with corresponding dates. As the EHIF reimburses health care providers on a fee-for-service basis, the database is considered to be relatively complete. As of December 2021, the EHIF contained information on 1 265 601 individuals or 94% of the Estonian population with insurance coverage [17]. The ECR is a population-based registry with nationwide coverage that has reliable cancer incidence data from 1968. It is compulsory for all physicians and pathologists working in Estonia to report cancer cases to the ECR. Additionally, the ECR uses multiple sources to ascertain cancer cases, including regular linkages with two cancer centres and trace-back of cases identified via death certificates. The completeness of case reporting is high, as evidenced by data quality indicators [18]. HPV data originated from a study in which an age-stratified (30–33, 57–60, 67–70) random sample of EstBB female gene donors was invited to a biobehavioural survey utilizing a self-administered survey on risk factors for cervical cancer and self-collected vaginal swabs for high-risk HPV detection (Supplementary Note).

Case definitions

Cervical intraepithelial neoplasia (CIN 2,3)

Phenotypes CIN 2 and 3 were defined using data from EHIF with ICD-10 codes N87.1 (CIN2), N87.2 (CIN3), and procedure/Nomesco codes corresponding to histological evaluation or biopsy of the cervix on the same medical claim (Supplementary Note).

Cervical cancer

Cervical cancer was defined using ECR and EHIF data with ICD-10 codes C53 and D06 and all their subcodes. Prevalent cervical cancer cases were defined as individual cancer patients who had received their diagnosis before joining EstBB. Incident cases of CIN and cancer were defined as individuals who were free of cervical cancer or CIN diagnosis at recruitment but received the corresponding diagnosis during the follow-up period.

Women who tested positive for any of the *high-risk HPV types* were considered to be infected.

Control group women

Control group women were defined as women without cervical pathology who had a known Pap test status (normal cytological finding) (Supplementary Note).

Statistical analysis

Polygenic risk scoring methods

Score 1 (LDpred)

We used the PRSs developed by Koel et al. [10], which were calculated using the LDPred software and a GWAS meta-analysis of UKBB, Kaiser Permanente, and FinnGen data (discovery sample). In brief, LDPred is a PRS software that adjusts GWAS summary statistics for the effects of linkage disequilibrium and produces different PRS profiles that differ in the expected proportion of causal SNPs and the adjusted weights given to individual SNPs. This set of PRSs included ten scores.

Score 2 (Bayesian whole-genome regression)

Score 2 used Bayesian whole-genome regression approaches. We compared the PRSs based on the BayesRR-RC model [19] using case–control data and the BayesW model [20] using time-to-event data. In contrast to Score 1, the SNP weights are recovered by analysing individual-level data. Hence, these models simultaneously estimate the effects of all variants, potentially giving an optimal predictor. The models were estimated for cervical cancer using only UK Biobank data of N = 248,798 European ancestry women (discovery sample), including 8,680 cervical cancer cases and 2,174,071 SNPs [21]. All PRSs were standardized, and effect sizes corresponded to an increase by one standard deviation.

Selecting the best-performing PRS from each score

The best PRSs were evaluated in a prevalence cervical cancer dataset comprising 691 prevalent cervical cancer case subjects and 13,820 control subjects. We tested the association between the PRS and the phenotype using age-adjusted logistic regression models. Based on the obtained odds ratios (ORs), we selected the best-performing PRSs from each test set and named them Score 1 and Score 2.

MetaPRS

To test the potential joint effect of Scores 1 and 2, we additionally combined them into a metaPRS [22], which was a weighted sum of the two scores. To construct the metaPRS, log odds ratios of PRSs from the logistic regression model in the prevalent analysis step were used as weights.

PRS association analysis with prevalent CIN and CC

Association with prevalent cervical intraepithelial neoplasia

Prevalent and incident cases represented slightly different aspects of the phenotype and were therefore analysed separately. Prevalent cases can be biased towards those with better survival, while incident cases represent the likelihood of getting the disease. Therefore, analysis of incident cases separately can allow a more thorough characterization of the predictive power of the PRS. We used the two best-performing PRSs identified in the first step of the analysis and tested their association with prevalent CIN cases in EstBB data. To test the association between the PRSs (Score 1 and 2) and prevalent CIN, we used an age-adjusted logistic regression model comparing prevalent cases and controls (individuals without cervical pathology and with a known PAP test status). We then compared the p values and odds ratios (ORs) between the two scores.

Association with incident cervical cancer and cervical intraepithelial neoplasia

Both PRSs (Scores 1 and 2) were evaluated in the analysis of incident CC and CIN cases and controls. This validation set was used to test the predictive ability of the PRSs. The PRSs were standardized and categorized into different groups of percentiles. We used Cox proportional hazard models to estimate the hazard ratios (HRs) corresponding to one standard deviation of the continuous PRS for the validation dataset. Harrell's C-statistic was used to characterize the discriminative ability of each PRS estimated from the same Cox proportional hazard models. Cumulative incidence estimates accounting for competing events (mortality) were computed using the "cmprsk" R library. While comparing different PRS groups with each other, age was used as a timescale (using both age at entry and age at the end of follow-up/diagnosis) to properly account for left truncation in the data.

Polygenic risk score analysis and its association with hrHPV infection

A total of 1347 women responded to the questionnaire, 207 of whom were hrHPV positive. Medians and interquartile ranges are presented for the numerical variables along with p values from the Mann–Whitney U test with a null hypothesis that the probability that the numeric value in the hrHPV-positive class is higher than that in the hrHPV-negative class, which is higher than 0.5. Percentages for each of the categories are presented within either hrHPV-positive or hrHPV-negative classes. For the categorical variables with only two classes, Fisher's exact test was performed, and for the categorical variables with more than two classes, a chi-squared test was performed, both testing a null hypothesis that category frequencies are not dependent on hrHPV status. We evaluated the properties of Scores 1 and 2 on hrHPV status prediction by comparing respective odds ratios. We present adjusted odds ratios such that the adjustment was made using several sociodemographic variables that could also stratify for hrHPV status: financial condition, marital status, age group and nationality (Supplementary Table 4).

We used R 4.1.1 for analysis [23].

Declarations

Ethical approval

The study was carried out under ethical approval 1.1-12/624 from the Estonian Committee on Bioethics and Human Research (Estonian Ministry of Social Affairs) and data release N05 from the EstBB. The study has been conducted according to the Declaration of Helsinki. All participants provided written informed consent to participate in the Estonian Biobank.

The UK Biobank study was approved by the North West Multi-Centre Research Ethics Committee (reference for UK Biobank is 16/NW/0274). All participants provided written informed consent to participate in the UK Biobank study.

The cross-sectional survey was approved by the Research Ethics Committee of the University of Tartu (protocols 300/T-17 20.01.2020, 332/M-7 21.12.2020) and by the Estonian Committee on Bioethics and Human Research (protocol 1.1-12/660) 14.01.2021.

Data Availability

The individual level data from Estonian Biobank are available under restricted access for containing sensitive information from healthcare registers, access can be obtained through the Estonian biobank upon submission of a research plan and signing a data transfer agreement. All data access to the Estonian Biobank must follow the informed consent regulations of the Estonian Committee on Bioethics and Human Research, which are clearly described in the Data Access section at https://genomics.ut.ee/en/content/estonian-biobank. A preliminary request for raw genetic and phenotype data must first be submitted via the email address releases@ut.ee. UKBB summary statistics can be accessed from http://www.decode.com/summarydata.

Acknowledgements

This work was supported through grant EMP416 from the EEA (European Economic Area) and Norway Grants. We thank Matthew R. Robinson for his valuable feedback. This work was supported by the Estonian Research Council grants PSG776 and PRG1911 and comments by the European Union through the European Regional Development Fund Project No. 2014-2020.4.01.15-0012 GENTRANSMED. This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101016775.

Consortium

Andres Metspalu, Lili Milani, Tõnu Esko, Reedik Mägi, Mari Nelis and Georgi Hudjashov

EstBBresearch@ut.ee

Author contributions:

Anna Tisler: Study design and concept, data collection during the cross-sectional study, drafting and critical revision of the initial manuscript

Anneli Uusküla: Study design and concept, data collection during the cross sectional study, drafting and critical revision of the initial manuscript

Sven Erik Ojavee: Study design and conception, data analysis and interpretation, drafting and critical revision of the initial manuscript

Kristi Läll: Study design and conception, data analysis and interpretation, drafting and critical revision of the initial manuscript

Triin Laisk: Study design, data analysis, drafting and critical revision of the initial manuscript

References

1. Sung H, *et al*. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. **71**:209-249(2021).
2. International Agency for Research on Cancer. Global cancer observatory. https://gco.iarc.fr/
3. Chesson, H.W., Dunne, E.F., Hariri, S., Markowitz, L.E. The estimated lifetime probability of acquiring human papillomavirus in the united states. *Sex Trans Dis*. **41**:660-664 (2014).
4. Castellsagu´e, X., Munoz, N. Cofactors in human papillomavirus carcinogenesis—role of parity, oral contraceptives, and tobacco smoking. *J Natl Cancer Inst Monogr*.**31**:20–28 (2003).
5. Stensen, S, *et al*. Factors associated with type-specific persistence of high-risk human papillomavirus infection: A population-based study. *Int J Cancer*.**138**: 361–368 (2016).
6. Ramachandran D, Dörk T. Genomic Risk Factors for Cervical Cancer. *Cancers* (Basel). **13**:5137 (2021).
7. Leo PJ, *et al.* Defining the genetic susceptibility to cervical neoplasia-A genome-wide association study. *PLoS Genet*.**13**:e1006866 (2017).
8. Hindorff, L.A, *et al*. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*. **106**: 9362–9367 (2009).
9. Läll K *et al*. Polygenic prediction of breast cancer: comparison of genetic predictors and implications for risk stratification. *BMC Cancer*.**19**:557 (2019).
10. Koel, M *et al*. Gwas meta-analysis and gene expression data link reproductive tract development, immune response and cellular proliferation/apoptosis with cervical cancer and clarify overlap with

other cervical phenotypes. Preprint at: https://www.medrxiv.org/content/10.1101/2021.06.18.21259075v2 (2021).

11. Ding Y et al. Large uncertainty in individual PRS estimation impact PRS-based risk stratification. Preprint at: https://www.biorxiv.org/content/10.1101/2020.11.30.403188v3 (2020).
12. He, YQ, *et al.* A polygenic risk score for nasopharyngeal carcinoma shows potential for risk stratification and personalized screening. *Nat Commun*. **13**, 1966 (2022).
13. Sipeky, C, *et al.* Prostate cancer risk prediction using a polygenic risk score. *Sci Rep*. **10**, 17075 (2020).
14. Leitsalu L *et al*. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int J Epidemiol*. **44**:1137-47 (2015).
15. Mitt M *et al*. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *Eur J Hum Genet*. **25**:869-876 (2017).
16. Leitsalu L *et al*. Linking a population biobank with national health registries-the estonian experience. *J Pers Med*. **5**:96-106 (2015).
17. Statistics Estonia. https://www.stat.ee/en
18. Orumaa M, Lang K, Magi M, Parna K, Aarelaid T, Innos K. The validity of Estonian Cancer Registry data in 1995-2008. *Eesti Arst*. **94**:339–346 (2015).
19. Patxot, M, *et al*. Probabilistic inference of the genetic architecture underlying functional enrichment of complex traits. *Nat Commun*. **12**:6972 (2021).
20. Ojavee, S.E, et al. Genomic architecture and prediction of censored time-to-event phenotypes with a Bayesian genome-wide analysis. *Nat Commun.* **12**:2337 (2021).
21. Ojavee S.E, et al. Novel discoveries and enhanced genomic prediction from modelling genetic risk of cancer age-at-onset. Preprint at: https://www.medrxiv.org/content/10.1101/2022.03.25.22272955v2 (2022).
22. Inouye M et al. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. *J Am Coll Cardiol*. **72**:1883-1893 (2018).
23. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/

Figures

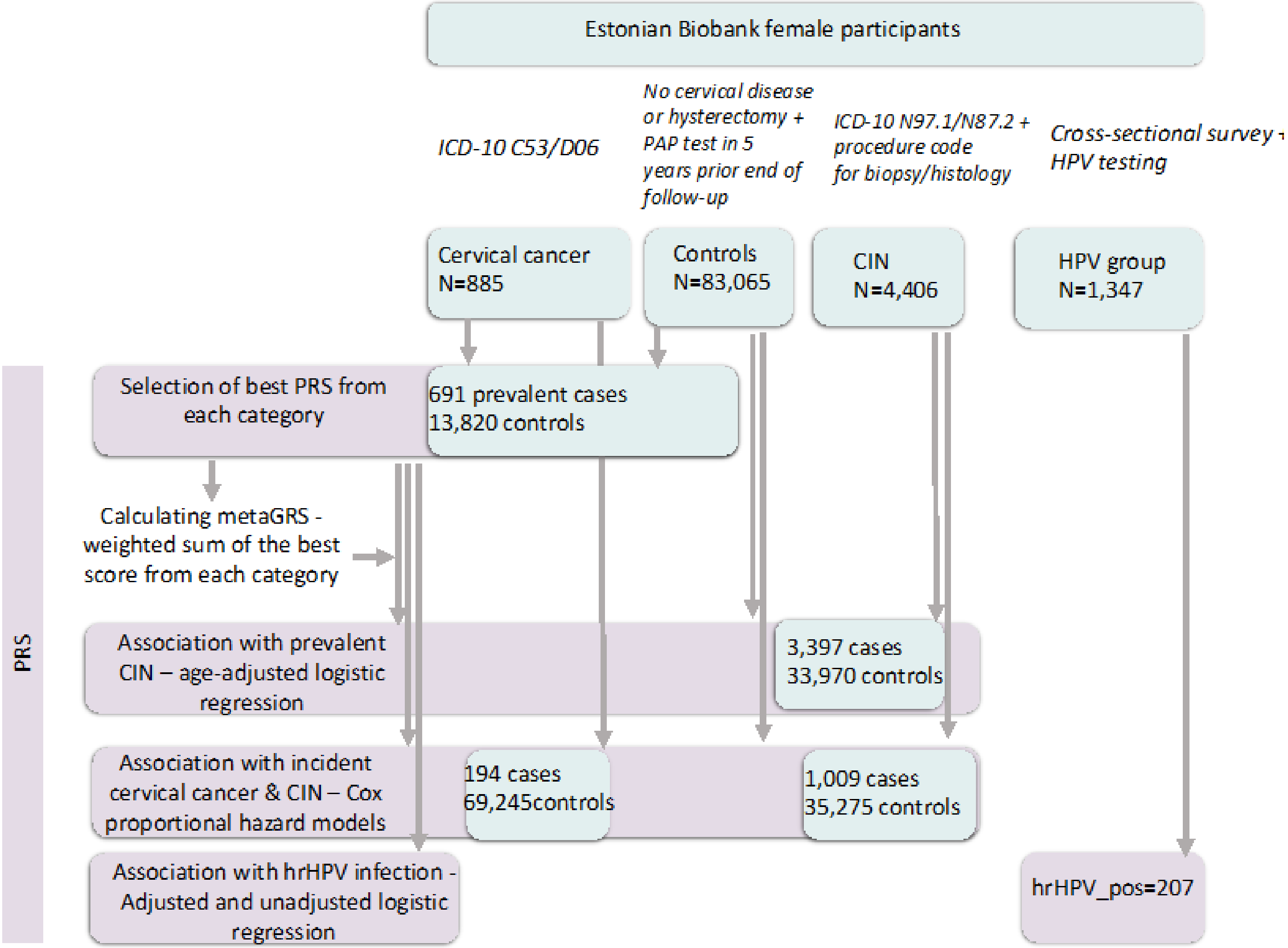
Estonian Biobank female participants
ICD-10 C53/D06
No cervical disease or hysterectomy + PAP test in 5 years prior end of follow-up
ICD-10 N97.1/N87.2 + procedure code for biopsy/histology
Cross-sectional survey + HPV testing
Cervical cancer N=885
Controls N=83,065
CIN N=4,406
HPV group N=1,347
Selection of best PRS from each category
691 prevalent cases 13,820 controls
Calculating metaGRS - weighted sum of the best score from each category
Association with prevalent CIN – age-adjusted logistic regression
3,397 cases 33,970 controls
Association with incident cervical cancer & CIN – Cox proportional hazard models
194 cases 69,245controls
1,009 cases 35,275 controls
Association with hrHPV infection - Adjusted and unadjusted logistic regression
hrHPV_pos=207
PRS

Figure 1

Flowchart of the study design and analysed groups

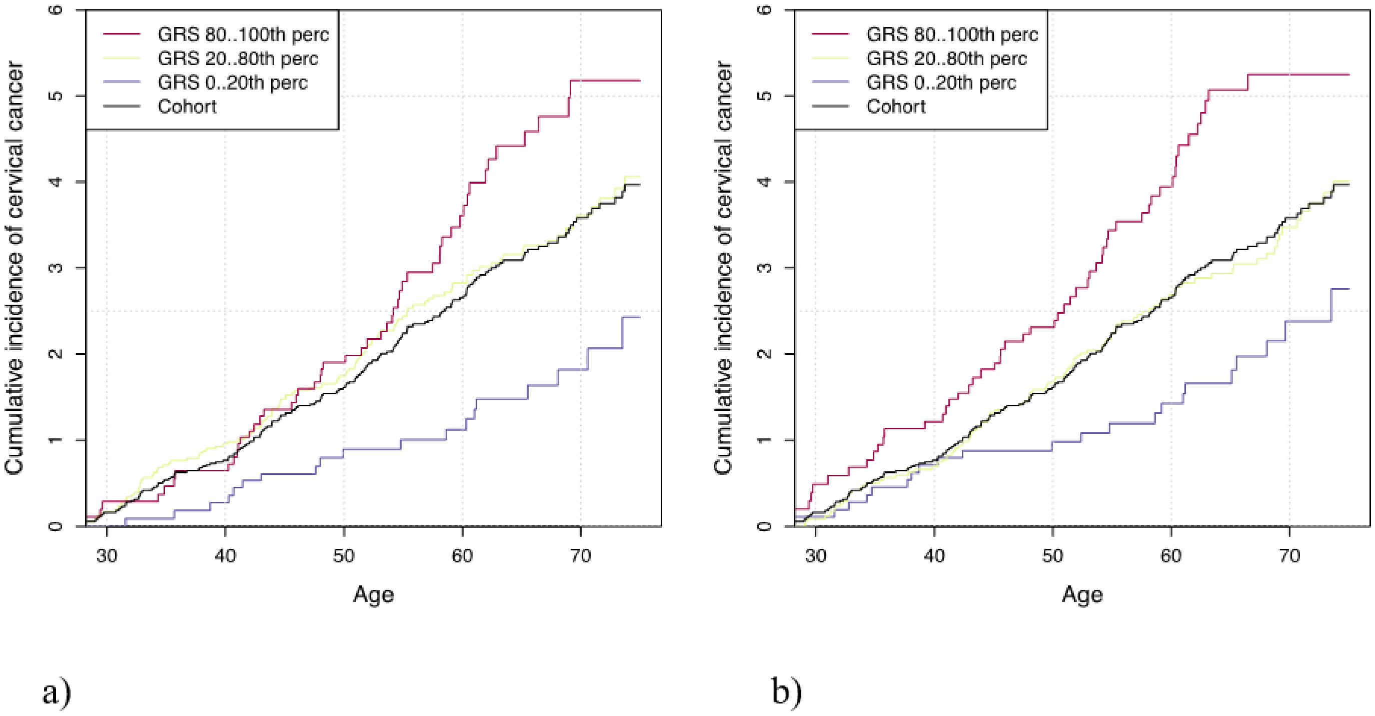

Figure 2

Cumulative incidence of cervical cancer (accounting for competing risks) in a) Score 1 and b) Score 2 risk categories among women aged 30-75 years.

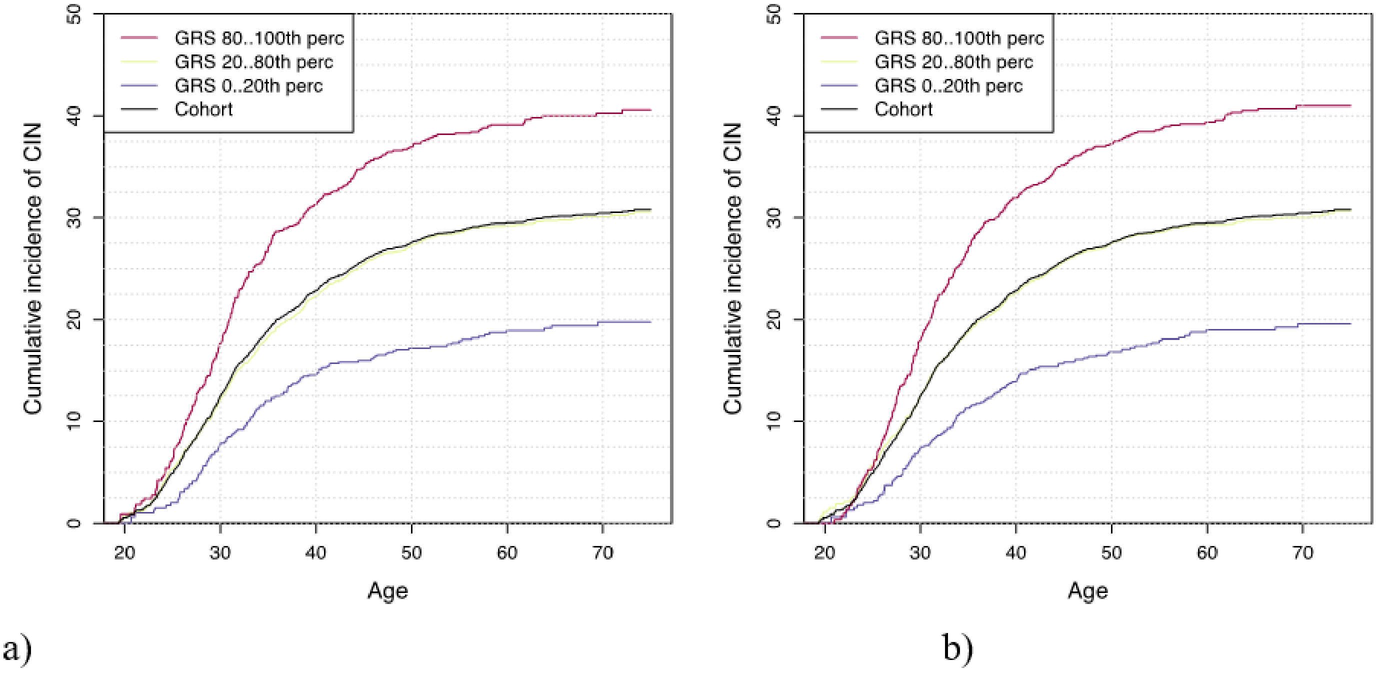

Figure 3

Cumulative incidence of cervical intraepithelial neoplasia (accounting for competing risks) in a) Score 1 and b) Score 2 risk categories among women aged 20-75.

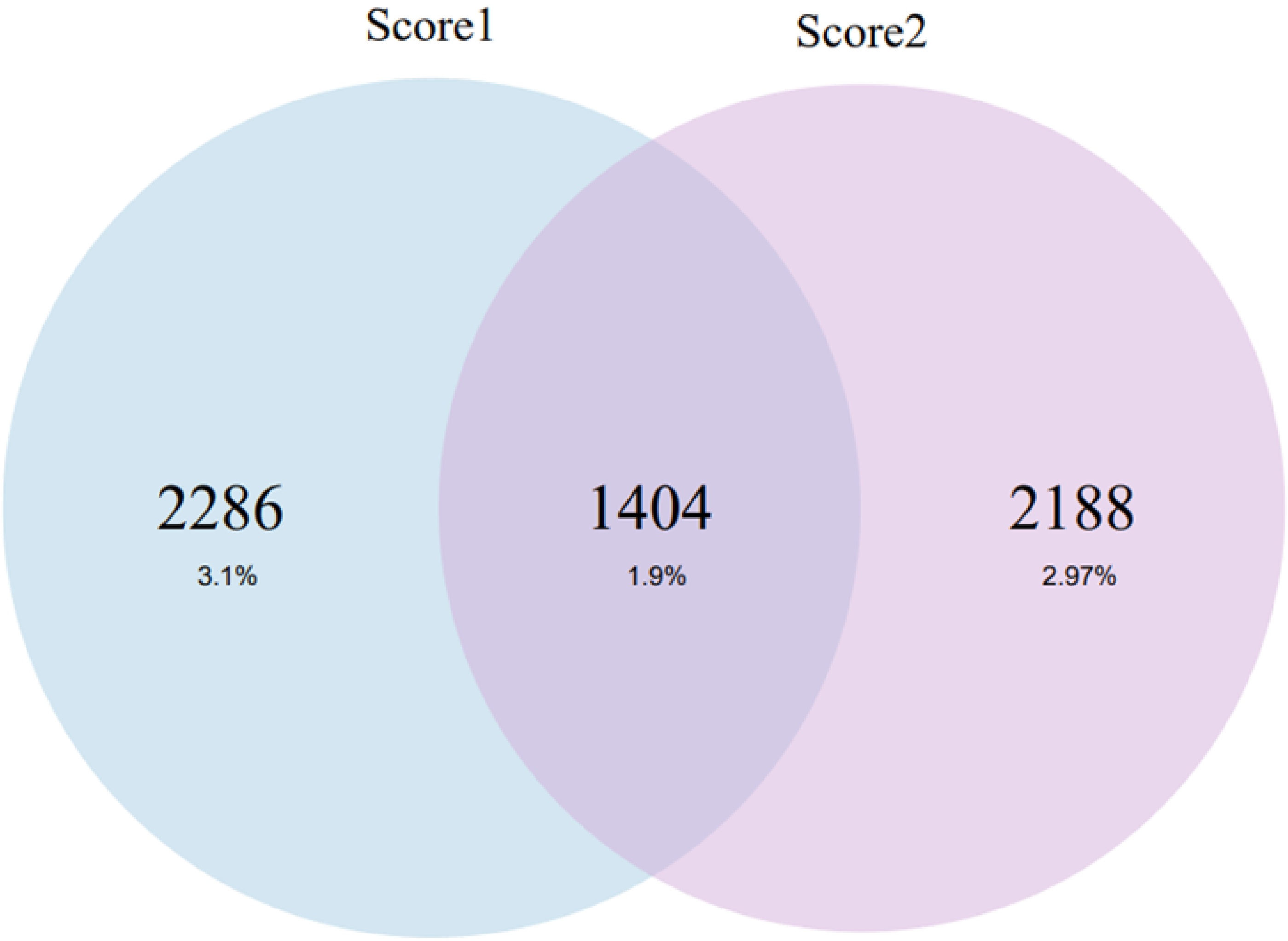

Figure 4

The overlap among highest-risk women (top 5%) in the Estonian Biobank according to two genetic risk scores for cervical cancer. The graph shows women who were classified as being in the top 5% with at least one of the genetic risk scores (Score 1 and 2).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryNote.rtf
- SuuplementaryTables.rtf