Predicting Lung Cancer in Korean Never-Smokers with Polygenic Risk Scores

Juyeon Kim[1], Young Sik Park[2], Jin Hee Kim[3], Yun-Chul Hong[4], Young-Chul Kim[5], In-Jae Oh[5], Sun Ha Jee[6], Myung-Ju Ahn[7], Jong-Won Kim[8], Jae-Joon Yim[2], Sungho Won[1, 9, 10, 11]*

[1]Department of Public Health Sciences, Seoul National University, Seoul, Korea

[2]Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Seoul National University Hospital, Seoul, Korea

[3]Department of Integrative Bioscience & Biotechnology, Sejong University, Seoul, Korea

[4]Department of Human Systems Medicine, Seoul National University College of Medicine, Seoul, Korea

[5]Department of Internal Medicine, Lung and Esophageal Cancer Clinic, Chonnam National University Hwasun Hospital, Hwasun, Korea

[6]Department of Epidemiology and Health Promotion, Institute for Health Promotion, Graduate School of Public Health, Yonsei University, Seoul, Korea

[7]Division of Hematology-Oncology, Department of Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea

[8]Department of Laboratory Medicine and Genetics, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea

[9]RexSoft Corps, Seoul, Korea

[10]Institute of Health and Environment, Seoul National University, Seoul, Korea.

[11]Interdisciplinary Program of Bioinformatics, Seoul National University, Seoul, Korea.

***Corresponding Author:** Sungho Won

Department of Public Health Science, Seoul National University, 1 Kwanak-ro Kwanak-gu Seoul 151-742 Korea; Email: won1@snu.ac.kr; Tel: +82-2-880-2714; Fax: +82-303-0942-2714.

ABSTRACT

In the last few decades, genome-wide association studies (GWAS) with more than 10,000 subjects have identified several loci associated with lung cancer. Hence, recently, genetic data have been used to develop novel risk prediction tools for cancer. The present study aimed to establish a lung cancer prediction model for Korean never-smokers using polygenic risk scores (PRSs). PRSs were calculated using a thresholding-pruning-based approach based on 11 genome-wide significant single nucleotide polymorphisms (SNPs). Overall, the odds ratios tended to increase as PRSs were larger, with the odds ratio of the top 5% PRSs being 1.71 (95% confidence interval: 1.31–2.23), and the area under the curve (AUC) of the prediction model being of 0.76 (95% confidence interval: 0.747–0.774). The receiver operating characteristic (ROC) curves of the prediction model with and without PRSs as covariates were compared using DeLong's test, and a significant difference was observed. Our results suggest that PRSs can be valuable tools for predicting the risk of lung cancer.

KEYWORDS

Genome-wide association study, lung cancer, never-smokers, polygenic risk score.

1 INTRODUCTION

Lung neoplasms are the leading cause of cancer worldwide (Fitzmaurice et al., 2019), with lung cancer having been the second most commonly diagnosed cancer in 2020, after breast cancer in women (Bray et al., 2018). In Korea, the age-adjusted prevalence of lung cancer in 2018 was 94.1 cases per 100,000 people, accounting for 4.7% of all cancer cases, and its age-adjusted incidence rate was of 28.0 cases per 100,000 people, accounting for 11.7% of all cancers (Korea Central Cancer Registry, 2020). Among men, the incidence of lung cancer in 2018 was 41.9 cases per 100,000 people, which is the second highest in Korea, whereas in Japan, United States, and United Kingdom the mean incidence was of 41.4, 40.1, and 35.5 cases per 100,000 people, respectively (Korea Central Cancer Registry, 2020). Lung cancer is the most common cancer affecting men aged > 65 years in Korea (Korea Central Cancer Registry, 2020), despite the proportion of never-smokers increased to 25.4% in 2009-2012, which was 19.1% in 2004-2008 (Park & Jang, 2016).

Smoking is a major risk factor for the progression of lung cancer and has been associated with over 80% of lung cancer cases in the Western world (Corrales et al., 2020). Indeed, reduced smoking habits has led to a decrease in mortality and incidence of lung cancer (Thandra et al., 2021). Nonetheless, even though most patients are smokers, the proportion of never-smokers with lung cancer has been increasing over time (Couraud et al., 2012), with the World Health Organization estimating that 25% of lung cancer cases worldwide occur in never-smokers (Ferlay et al., 2010). Therefore, the risk profiles of never-smokers are expected to be markedly different from those of smokers, with family history, secondhand smoke, cooking oil fumes, radon exposure, domestic fuel smoke, asbestos, and menopausal hormone replacement therapy being suggested as potential risk factors associated with lung cancer in never-smokers. However, to date, none of these sug gested risk factors have been exclusively identified in never-smokers (Couraud et al., 2012; Hung et al., 2021).

Lung cancer in never-smokers has been considered to be a distinct medical entity from that in ever-smokers, and some clinically important features have been identified. First, it is more frequent in certain regions than in others (Asia > North America > Europe) (Couraud et al., 2012). Second, mutations in the epidermal growth factor receptor (EGFR) are more common in 1) adenocarcinomas than in non-small cell lung cancer and 2) in never-smokers than in ever-smokers (Chapman et al., 2016). Noteworthily, although the somatic variant profile of Asian populations is very similar to that of Europeans, the prevalence of *EGFR* mutations is higher in Asian women than in Caucasian women (Chapman et al., 2016). Therefore, it is unlikely that other risk factors, such as secondhand smoke, could be responsible for the increased lung cancer incidence in the Asian population (Mitsudomi, 2014). Several studies have suggested that *EGFR* mutations occur independently of smoking and that the low frequency of *EGFR* mutations in smokers could be explained by the occurrence of smoking-related lung cancer (Mitsudomi, 2014; Shi et al., 2014; Truong et al., 2010). Moreover, genome-wide data related to *EGFR* mutations failed to provide relevant knowledge on carcinogenesis.

GWAS have discovered many common genetic variants associated with complex traits and disorders (Buniello et al., 2019; Klein et al., 2005; Visscher et al., 2017). Most cancers are highly polygenic (Stahl et al., 2012; Zeng et al., 2018; Zhang et al., 2018; Zhang et al., 2020), with lung cancer being one of the most polygenic (Zhang et al., 2020), along with breast and oropharynx cancers. For these polygenic traits, the effect size associated with each risk variant is small, and individuals with multiple risk variants tend to have an elevated disease risk (Chatterjee et al., 2013). Therefore, PRS can be useful for risk assessment as it combines multiple variants into scores that evaluate genetic susceptibility (Dudbridge, 2013). Several studies have shown that PRS can be used as a predictor of lung cancer in a population that includes both ever-smokers and never-sm okers; however, most of these studies were conducted on non-Hispanic whites, and no study evaluated exclusively for Asian never-smokers. In the present study, a lung cancer prediction model for Korean never-smokers was built using PRSs and its accuracy was evaluated.

2 MATERIALS AND METHODS

2.1 Korean lung cancer cohorts

Never-smoking Korean lung cancer subjects were recruited from five different institutes: Seoul National University Hospital (SNUH), Yeonsei University (YSU), Sejong University (SU), Samsung Medical Center (SMC), and Chonnam National University (CNU) (Ahn et al., 2012; Kim et al., 2013; Lan et al., 2012; Lee et al., 2017). Data from non-smoking controls were obtained from the CAVAS study of the Korean Genome and Epidemiology Study (Kim & Han, 2017). Never-smokers were defined as those who had smoked less than 100 cigarettes in their lifetime or had never smoked. A total of 8,348 individual were included in the study (1,642 cases and 6,706 controls matched according to their principal component (PC) scores calculated using the EIGENSTRAT method (Price et al., 2006)). All participants provided written informed consent, and the study was approved by the institutional review board and ethics committee of the Seoul National University Hospital (approval no. H-1906-126-1042).

2.2 Genotyping, quality control, and imputation

SNUH and YSU cohorts were genotyped using Axiom KoreanChip V1.0 or Axiom KoreanChip V1.1 (Moon et al., 2019), SU and SMC cohorts were evaluated using the Affymetrix Genome-Wide Human SNP Array (5.0 and 6.0, respectively), and CNU cohort was evaluated using an Illumina Human660W-Quad array. Variant calling for SNUH and YSU was performed using the K-medoid approach (Seo et al., 2019). The analysis approach used is summarized in **Figure 1**. As different genotyping platforms can generate substantial numbers of false-positive data, and quality con trols (QC) were carefully performed. SNPs were removed if call rates were < 95% or 99%, *P*-values for Hardy–Weinberg equilibrium were < 10^{-3} or 10^{-5}, or minor allele frequencies (MAFs) were < 5%. Subjects were also excluded if there was sex inconsistency, call rate < 0.95, outlying heterozygosity (heterozygosity rate > mean ± 3 standard deviation [SD]), or estimated identity-by-descent > 0.9. QCs were conducted with cases and controls separately for each participant institution, with cases and controls from the same institute being merged and the same QC being applied to the merged data with the following additional step: SNPs were removed if missing rates between cases and controls differed significantly ($P < 0.01$). Lastly, genotyping platforms of SNUH and YSU, and for SU and SMC were the same, and subjects from institutes with the same genotyping platforms were pooled. The same QCs were applied to pooled subjects. After QCs, the remaining subjects and SNPs were used to impute the untyped SNPs using the Michigan imputation server. Non-Europeans of the Haplotype Reference Consortium (r1.1 2016) were selected as reference panel, and Eagle (v2.4) was used as the phasing program. Imputed SNPs were removed if MAFs were < 0.05, $R^2 < 0.3$, *P*-values for the Hardy–Weinberg equilibrium exact test were < 10^{-3} or 10^{-5}, call rates were < 95% or 99%.

Association analyses were conducted using logistic regression. To adjust for population stratification, PC scores were calculated, and the 10 PC scores corresponding to the 10 largest eigenvalues were included as covariates in the following logistic regressions:

$$logit(P[sex, age, PC, SNPs]) = \beta_0 + \beta_{sex} sex + \beta_{age} age + \beta_{PC} PC + \beta_{SNPs} SNPs \quad (1)$$

The genomic inflation factor and quantile-quantile plot were used to compare the genome-wide distribution of the test statistic for H_0: $\beta_{SNPs} = 0$ with the expected null distribution.

2.3 Polygenic risk score construction

Calculation of PRS requires an effect size estimate of genome-wide significant SNPs. GWAS Cat alog (https://www.ebi.ac.uk/gwas/) and PubMed (https://pubmed.ncbi.nlm.nih.gov) were screened to obtain GWAS summary statistics of lung cancer in never-smokers of Asian ancestry. Seow *et al.* (Seow et al., 2017) conducted a GWAS using the largest East Asian population, reporting 11 genome-wide significant SNPs, among which the genotypes of 10 SNPs were available in each Korean cohort(Table S1); thus, their summary statistics were incorporated to build the PRS. Let $\widehat{\beta}_i$ be the log odds ratios (ORs) obtained from Seow *et al.* (Seow et al., 2017) for SNP *i* (*i* = 1, […], 11) and x_{ij} be the number of risk alleles of SNP *i* for subject *j* in the Korean cohort (x_{ij} = 0, 1, 2); then, the PRS of subject *j* was calculated by a weighted sum of the risk alleles that an individual carries, as follows:

$$\begin{aligned} PRS_j = \; & -0.22 \cdot \mathrm{rs}4488809_j + 0.36 \cdot \mathrm{rs}\ 2736100_j - 0.15 \cdot \mathrm{rs}9387478_j + 0.15 \cdot \mathrm{rs}3817963_j + 0.15 \\ & \cdot \mathrm{rs}2395185_j + 0.16 \cdot \mathrm{rs}2179920_j - 0.16 \cdot \mathrm{rs}7741164_j - 0.27 \cdot \mathrm{rs}72658409_j + 0.22 \\ & \cdot \mathrm{rs}7086803_j - 0.16 \cdot \mathrm{rs}11610143_j - 0.16 \cdot \mathrm{rs}7216064_j \end{aligned} \tag{2}$$

2.4 Logistic regression and its prediction accuracy

The prediction model was built with using logistic regression models. Lung cancer status was used as the response variable. PRSs were categorized into nine different groups based on PRSs of the subject percentiles of the controls: < 5%, 5–10%, 10–20%, 20–40%, 40–60%, 60–80%, 80–90%, 90–95%, and > 95%, which were indicated as 1, 2, (…), and 9, respectively. PRSs were incorporated as covariates to estimate their ORs after adjusting for sex, age, 10 PC scores, and genotyping array as follows:

$$logit\left(P[PRS, sex, age, array, PC]\right) = \beta_0 + \beta_{PRS} PRS + \beta_{sex} sex + \beta_{age} age + \beta_{array} array + \beta_{PC} PC \tag{3}$$

In this study, 10 PC scores were used to adjust the population stratification. To assess the ability o f the PRS to identify high-risk cases, we considered an alternative model in which the PRS was coded as 1 for the top 1% PRSs, otherwise was coded as 0.

The prediction accuracy of the logistic regression was evaluated using the AUC. The confidence interval and *P*-value were obtained using the DeLong's test. Subjects from SU and CNU overlapped with those of Seow *et al*. (Seow et al., 2017), and most never-smokers were females. Therefore, the accuracy of the prediction model was evaluated according to three different scenarios: Dataset 1, all subjects (SNUH, YSU, SU, SMC, and CNU); Dataset 2, only females; and dataset 3, subjects from SNUH, YSU, and SMC. Moreover, ROC curves of the prediction model with and without PRSs as covariates were compared using DeLong's test. The ORs of the PRSs were estimated and adjusted for the first 10 PC scores, sex, age, and dataset. All the analyses were performed using Plink (v1.9 and v2.0), ONETOOL (Song et al., 2018), R (v3.6.3), and Python (v2.7.17).

2.5 Meta analyses

Meta-analyses were conducted to calculate the combined effect sizes for each SNP using METAL (Willer et al., 2010). The effect sizes of each SNP were combined using weighted means. Forest plots were obtained using R (v3.6.3)(Figure S1).

3 RESULTS

3.1 Descriptive statistics

The descriptive statistics of the subjects included in the study are shown in **Table 1**. Among the 8,348 individuals evaluated, 72.4% were females and a total of 84.6% of the patients were pathologically diagnosed with adenocarcinoma. Significant differences in age were observed among the different study cohorts.

3.2 Odds ratios of PRS and prediction accuracy

The PRSs were calculated for each dataset. The Kolmogorov-Smirnov test showed that the PRSs were normally distributed, and that the cases had significantly higher PRSs than controls in all datasets (**Table 2**, **Figure S2**). Moreover, Table 2 also shows that cases always have significantly higher means than controls in Dataset1, which includes all subjects (P = 4.50 × 10^{-10} for KoreanChip; P = 6.34 × 10^{-9} for Affymetrix; P = 3.63 × 10^{-8} for Illumina array). the estimated ORs of the PRSs tended to increase in higher percentile groups of the PRSs, compared with the reference percentile group (40–60%) (**Figure 2**). OR of the top 5% PRS group was 1.71 (95% confidence interval [CI]: 1.31–2.23; P = 7.40 × 10^{-5}), and for females the OR was 1.66 (1.25–2.19; P = 3.76 × 10^{-4}). The OR of the top 5% PRS group was maximized for Dataset 3 (OR = 2.45 [1.74–3.44]; P = 2.21 × 10^{-7}) (**Table S2**). No significant differences between men and women were observed in Dataset1, which includes all subjects. The ORs of the bottom 5% PRS group were less than 1, indicating their protective effect against lung cancer (**Table S2**). Some PRS percentile groups were not significantly different from the reference group, but an increasing tendency in ORs was observed for all datasets (**Table S2**).

Table 3 shows the ORs of the top 1% PRSs compared with the other PRS subgroups. Overall, the ORs were significant only for Dataset 3 and the OR of the top 1% PRS group was generally not higher than that of the top 5% group (**Table 3**, **Table S2**). For comparison, we referred to the ORs reported in the study by Fritsche *et al.* (Fritsche, 2020), which included data from lung cancer patients obtained from the UK biobank (UKB) and Michigan Genomics Initiative (MGI) (**Table 3**). Most of these patients were Caucasian, and both smokers and never-smokers were included. Although the ethnicity and smoking status were different, OR of the top 1% PRSs of Dataset3 was higher than those of UKB and MGI.

3.3 Lung cancer prediction with polygenic risk scores

PRSs, age, and sex were considered covariates, and a prediction model was built. The pre dictors included sex, age, and continuous PRSs. The highest AUC was 0.764 (95% CI: 0.750–0.778; $P = 7.02 \times 10^{-280}$) from Dataset 2, followed by Dataset 1 (**Figure 3**, **Table 4**). AUCs from Dataset 1 and Dataset 2 were similar ($P = 0.73$), whereas those from Dataset 1 and Dataset 3 differed significantly ($P = 1.80 \times 10^{-6}$). Moreover, significant differences were observed concerning the ROC of the prediction model with and without PRSs as covariates in every scenario (**Table 4**).

4 DISCUSSION

Genetics for lung cancer in never-smokers has been considered one of the most important risk factors, with several studies having been performed to predict lung cancer while considering specific genetic factors. Recently, PRS has been applied to predict lung cancer; however, there no such studies have been conducted exclusively for Asian never-smokers. In this study, we constructed PRSs based on recent meta-analyses and evaluated their prediction accuracy for lung cancer in never-smokers in Korea. Our results show that individuals with PRSs higher than the reference percentile group(40-60%) have a much higher probability of developing lung cancer; thus, these PRSs can be considered valuable predictors of lung cancer.

To date, the largest studies exploring PRS in patients with lung cancer were based on the MGI and UKB cohorts, which consist of non-Hispanic white European populations, including ever and never-smokers. Their data suggested that the top 1% PRS represented an increased risk of lung cancer, with ORs of 1.75 for MGI (95% CI 0.796–3.85) and 1.94 for UKB (95% CI 1.22–3.1) groups (Fritsche, 2020). In our analyses, the OR of the top 1% PRS in never-smoker subjects of the Dataset 3 was 2.22 and was significantly associated with increase lung cancer risk ($P = 0.03$). For Datasets 1 and 2, the results were not statistically significant, but the data suggested a tendency for increased risk (OR > 1). These results indicate that the PRS can be utilized as a prognostic tool for lung cancer, regardless of the smoking status of the patient. Moreover, PRS can be useful for identifying never-smoking individuals with a high risk of lung cancer.

Our data showed that the predictive potential of PRS was similar between women and men. Multiple studies have shown substantial differences in lung cancer incidence according to sex. For example, studies based on data from The Cancer Genome Atlas showed that 15% of autosomal genes have sex-biased copy number alterations in several cancers, which can be associated with different mRNA expression profiles (Lopes-Ramos et al., 2020). Sex-related behavior and exposure can also affect gene mutations. In non-small cell lung cancer, the mutational spectrum of *EGFR* and *TP53* is influenced by sex. Indeed, the frequency of transversion mutations on *TP53* is 40% among women, which is higher than among men (25-28%) (Lopes-Ramos et al., 2020). Moreover, different methylation patterns by sex have been observed in various human tissues, such as blood, brain, and muscle (Lopes-Ramos et al., 2020). Therefore, it is expected that a combination of genetic mechanisms can contribute to epidemiological sex differences. However, in the present study, no sex-specific differences were observed. The largest difference in MAFs between men and women was 0.002, and there were no SNPs with minor allele frequencies that were significantly different between women and men.

This study had some limitations. First, although more than 1,500 lung cancer patients were considered, genetic analyses usually require more than 10,000 subjects, and the sample size may not be sufficient for evaluating the accuracy of the risk prediction model estimates using PRSs. Second, lung cancer consists of etiologically heterogeneous subtypes; however, this information was not available for this study. If etiological subtypes are to be considered, prediction accuracy may be greatly improved. Thus, further studies considering the subtype-specific genetic architecture of lung cancer with adequate sample size are still needed to further confirm our findings. Third, summary statistics were only available for some SNPs. Some studies showed that the AUC difference between the prediction model with 592,000 SNPs and 10 SNPs was 0.03, with inclusion of more SNPs not substantially improving prediction accuracy (Liyanarachchi et al., 2020). However, this conclusion was obtained based on non-Hispanic white populations; hence, further investi gations are necessary for East Asians. Fourth, it has been shown that secondhand smoke significantly affects lung cancer; however, its effect was not considered in this study. The mean age of the study participants was 60.2 years old. Laws to ban or stop smoking in all enclosed workplaces or cessation of health have been recently adopted, and the controls of our study participants may have also been affected by secondhand smoke.

Lung cancer has been widely known to be a highly heterogeneous disease that can occur and progress due to the interplay between permanent genetic mutations and epigenetic alterations (Dong et al., 2017). The lung cancer prediction accuracy can be improved by combining other clinical or lifestyle risk factors. In this study, we demonstrate that PRS can be a valuable tool for identifying individuals at a high risk of lung cancer. However, the predictive accuracy of PRSs is still not sufficiently good so it can be used in clinical practice; hence, further studies are warranted to improve it.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

AUTHOR CONTRIBUTIONS

Juyeon Kim conceived the project, carried out data analysis, and drafted the manuscript. Sungho Won reviewed the literature and provided feedback. Young Sik Park, Jin Hee Kim, Yun-Chul Hong, Young-Chul Kim, In-Jae Oh, Sun Ha Jee, Myung-Ju Ahn, Jong-Won Kim, and Jae-Joon Yim conducted the epidemiological studies, contributed samples to the genotyping. All authors contributed to the writing of the manuscript.

ACKNOWLEDGEMENTS

This work was supported by the Technology Innovation Program (20016417) funded by the Ministry of Trade, Industry, and Energy (Korea).

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

ORCID

Juyeon Kim http://orcid.org/0000-0001-8919-4793

Sungho Won http://orcid.org/0000-0001-5751-5089

REFERENCES

Ahn, M. J., Won, H. H., Lee, J., Lee, S. T., Sun, J. M., Park, Y. H., . . . Park, K. (2012). The 18p11.22 locus is associated with never smoker non-small cell lung cancer susceptibility in Korean populations. *Hum Genet*, *131*(3), 365-372. https://doi.org/10.1007/s00439-011-1080-z

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*, *68*(6), 394-424. https://doi.org/10.3322/caac.21492

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., . . . Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*, *47*(D1), D1005-d1012. https://doi.org/10.1093/nar/gky1120

Chapman, A. M., Sun, K. Y., Ruestow, P., Cowan, D. M., & Madl, A. K. (2016). Lung cancer mutation profile of EGFR, ALK, and KRAS: Meta-analysis and comparison of never and ever smokers. *Lung Cancer*, *102*, 122-134. https://doi.org/10.1016/j.lungcan.2016.10.010

Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S. J., & Park, J. H. (2013). Project ing the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat Genet*, *45*(4), 400-405, 405e401-403. https://doi.org/10.1038/ng.2579

Corrales, L., Rosell, R., Cardona, A. F., Martín, C., Zatarain-Barrón, Z. L., & Arrieta, O. (2020). Lung cancer in never smokers: The role of different risk factors other than tobacco smoking. *Crit Rev Oncol Hematol*, *148*, 102895. https://doi.org/10.1016/j.critrevonc.2020.102895

Couraud, S., Zalcman, G., Milleron, B., Morin, F., & Souquet, P. J. (2012). Lung cancer in never smokers--a review. *Eur J Cancer*, *48*(9), 1299-1311. https://doi.org/10.1016/j.ejca.2012.03.007

Dong, N., Shi, L., Wang, D. C., Chen, C., & Wang, X. (2017). Role of epigenetics in lung cancer heterogeneity and clinical implication. *Semin Cell Dev Biol*, *64*, 18-25. https://doi.org/10.1016/j.semcdb.2016.08.029

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genet*, *9*(3), e1003348. https://doi.org/10.1371/journal.pgen.1003348

Ferlay, J., Shin, H. R., Bray, F., Forman, D., Mathers, C., & Parkin, D. M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer*, *127*(12), 2893-2917. https://doi.org/10.1002/ijc.25516

Fitzmaurice, C., Abate, D., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdel-Rahman, O., . . . Murray, C. J. L. (2019). Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-Years for 29 Cancer Groups, 1990 to 2017: A Systematic Analysis for the Global Burden of Disease Study. *JAMA Oncol*, *5*(12), 1749-1768. https://doi.org/10.1001/jamaoncol.2019.2996

Fritsche, L. G., Patil, S., Beesley, L. J., VandeHaar, P., Salvatore, M., Ma, Y., Mukherjee, B. (2020). Cancer PRSweb: An Online Repository with Polygenic Risk Scores for Major Cancer Traits and Their Evaluation in Two Independent Biobanks. Am J Hum Genet, 107(5), 815-836. https://doi.org/10.1016/j.ajhg.2020.08.025

Hung, R. J., Warkentin, M. T., Brhane, Y., Chatterjee, N., Christiani, D. C., Landi, M. T., . . . Amos, C. I. (2021). Assessing Lung Cancer Absolute Risk Trajectory based on a Polygenic Risk Model. *Cancer Res*. https://doi.org/10.1158/0008-5472.Can-20-1237

Kim, J. H., Park, K., Yim, S. H., Choi, J. E., Sung, J. S., Park, J. Y., . . . Hong, Y. C. (2013). Genome-wide association study of lung cancer in Korean non-smoking women. *J Korean Med Sci*, *28*(6), 840-847. https://doi.org/10.3346/jkms.2013.28.6.840

Kim, Y., & Han, B. G. (2017). Cohort Profile: The Korean Genome and Epidemiology Study (KoGES) Consortium. *Int J Epidemiol*, *46*(2), e20. https://doi.org/10.1093/ije/dyv316

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., . . . Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, *308*(5720), 385-389. https://doi.org/10.1126/science.1109557

Korea Central Cancer Registry, N. C. C. (2020). *Annual report of cancer statistics in Korea in 2018* M. o. H. a. Welfare.

Lan, Q., Hsiung, C. A., Matsuo, K., Hong, Y. C., Seow, A., Wang, Z., . . . Rothman, N. (2012). Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. *Nat Genet*, *44*(12), 1330-1335. https://doi.org/10.1038/ng.2456

Lee, S. J., Jee, Y. H., Jung, K. J., Hong, S., Shin, E. S., & Jee, S. H. (2017). Bilirubin and Stroke Risk Using a Mendelian Randomization Design. *Stroke*, *48*(5), 1154-1160. https://doi.org/10.1161/strokeaha.116.015083

Liyanarachchi, S., Gudmundsson, J., Ferkingstad, E., He, H., Jonasson, J. G., Tragante, V., . . . de la Chapelle, A. (2020). Assessing thyroid cancer risk using polygenic risk scores. *Proc Natl Acad Sci U S A*, *117*(11), 5997-6002. https://doi.org/10.1073/pnas.1919976117

Lopes-Ramos, C. M., Quackenbush, J., & DeMeo, D. L. (2020). Genome-Wide Sex and Gender

Differences in Cancer [Review]. *Frontiers in Oncology*, *10*. https://doi.org/10.3389/fonc.2020.597788

Mitsudomi, T. (2014). Molecular epidemiology of lung cancer and geographic variations with special reference to EGFR mutations. *Transl Lung Cancer Res*, *3*(4), 205-211. https://doi.org/10.3978/j.issn.2218-6751.2014.08.04

Moon, S., Kim, Y. J., Han, S., Hwang, M. Y., Shin, D. M., Park, M. Y., . . . Kim, B.-J. (2019). The Korea Biobank Array: Design and Identification of Coding Variants Associated with Blood Biochemical Traits. *Scientific Reports*, *9*(1), 1382. https://doi.org/10.1038/s41598-018-37832-9

Park, J. Y., & Jang, S. H. (2016). Epidemiology of Lung Cancer in Korea: Recent Trends. *Tuberc Respir Dis (Seoul)*, *79*(2), 58-69. https://doi.org/10.4046/trd.2016.79.2.58

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904-909. https://doi.org/10.1038/ng1847

Seo, S., Park, K., Lee, J. J., Choi, K. Y., Lee, K. H., & Won, S. (2019). SNP genotype calling and quality control for multi-batch-based studies. *Genes & Genomics*, *41*(8), 927-939. https://doi.org/10.1007/s13258-019-00827-5

Seow, W. J., Matsuo, K., Hsiung, C. A., Shiraishi, K., Song, M., Kim, H. N., . . . Lan, Q. (2017). Association between GWAS-identified lung adenocarcinoma susceptibility loci and EGFR mutations in never-smoking Asian women, and comparison with findings from Western populations. *Hum Mol Genet*, *26*(2), 454-465. https://doi.org/10.1093/hmg/ddw414

Shi, Y., Au, J. S., Thongprasert, S., Srinivasan, S., Tsai, C. M., Khoa, M. T., . . . Yang, P. C. (2014). A prospective, molecular epidemiology study of EGFR mutations in Asian patients with advanced non-small-cell lung cancer of adenocarcinoma histology (PIONEER). *J Thorac Oncol*, *9*(2), 154-162. https://doi.org/10.1097/jto.0000000000000033

Song, Y. E., Lee, S., Park, K., Elston, R. C., Yang, H.-J., & Won, S. (2018). ONETOOL for the analysis of family-based big data. *Bioinformatics (Oxford, England)*, *34*(16), 2851-2853. https://doi.org/10.1093/bioinformatics/bty180

Stahl, E. A., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B. F., . . . Plenge, R. M. (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat Genet*, *44*(5), 483-489. https://doi.org/10.1038/ng.2232

Thandra, K. C., Barsouk, A., Saginala, K., Aluru, J. S., & Barsouk, A. (2021). Epidemiology of lung cancer. *Contemp Oncol (Pozn)*, *25*(1), 45-52. https://doi.org/10.5114/wo.2021.103829

Truong, T., Hung, R. J., Amos, C. I., Wu, X., Bickeböller, H., Rosenberger, A., . . . Spitz, M. R. (2010). Replication of lung cancer susceptibility loci at chromosomes 15q25, 5p15, and 6p21: a pooled analysis from the International Lung Cancer Consortium. *J Natl Cancer Inst*, *102*(13), 959-971. https://doi.org/10.1093/jnci/djq178

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet*, *101*(1), 5-22. https://doi.org/10.1016/j.ajhg.2017.06.005

Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics (Oxford, England)*, *26*(17), 2190-2191. https://doi.org/10.1093/bioinformatics/btq340

Zeng, J., de Vlaming, R., Wu, Y., Robinson, M. R., Lloyd-Jones, L. R., Yengo, L., . . . Yang, J. (2018). Signatures of negative selection in the genetic architecture of human complex traits. *Nature Genetics*, *50*(5), 746-753. https://doi.org/10.1038/s41588-018-0101-4

Zhang, Y., Qi, G., Park, J.-H., & Chatterjee, N. (2018). Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 com plex traits. *Nature Genetics*, *50*(9), 1318-1326. https://doi.org/10.1038/s41588-018-0193-x

Zhang, Y. D., Hurson, A. N., Zhang, H., Choudhury, P. P., Easton, D. F., Milne, R. L., . . . Testicular Cancer, C. (2020). Assessment of polygenic architecture and risk prediction based on common variants across fourteen cancers. *Nature Communications*, *11*(1), 3353. https://doi.org/10.1038/s41467-020-16483-3

Table 1. Descriptive statistics.

	SNUH	**YSU**	**SU**	**SMC**	**CNU**	**Total**	***P*-value**
Overall, *N*	1,694	1,018	3,038	1,575	1,023	8,348	
Case	206	172	276	407	581	1,642	< 0.001
Control	1,488	846	2,762	1,168	442	6,706	
Sex, *N*							
Male	389	336	318	339	0	1,382	< 0.001
Female	1,305	682	2,720	1,236	1,023	6,966	
Mean age (SD), years	53.5 (10.0)	47.2 (11.6)	53.3 (9.4)	60.0 (8.1)	60.5 (10.7)		< 0.001
Histology, *N*							
AD	192	NA	240	359	453		< 0.001
Non-AD	14	NA	36	48	128		

Abbreviations: AD, adenocarcinoma; CNU, Chonnam National University; NA, not available; SD, standard deviation; SMC, Samsung Medical Center; SNUH, Seoul National University Hospital; SU, Sejong University; YSU, Yeonsei University.

Table 2. Mean differences between cases and controls

	Polygenic risk score, mean (SD)		*P*-values		
	Cases	Controls	KS test	*t*-test for phenotype	*t*-test for sex
Dataset 1[†]					
KoreanChip	0.30 (0.97)	−0.05 (1.0)	0.33	4.50×10^{-10}	0.12
Affymetrix	0.20 (0.98)	−0.04 (1.0)	0.08	6.34×10^{-9}	0.13
Illumina	0.15 (0.98)	−0.20 (0.99)	0.66	3.63×10^{-8}	NA[†]
Dataset 2[‡]					
KoreanChip	0.30 (0.95)	−0.04 (0.99)	0.24	1.96×10^{-8}	NA[†]
Affymetrix	0.20 (0.98)	−0.05 (1.0)	0.11	2.45×10^{-8}	NA[†]
Illumina	0.15 (0.98)	−0.20 (0.99)	0.66	3.63×10^{-8}	NA[†]
Dataset 3[§]					
KoreanChip	0.30 (0.97)	−0.05 (1.0)	0.33	4.50×10^{-10}	0.12
Affymetrix	0.30 (1.04)	−0.10 (0.97)	0.15	2.88×10^{-10}	0.80
Illumina			NA		

[†] Included all subjects from Chonnam National University (CNU), Samsung Medical Center (SMC), Seoul National University Hospital (SNUH), Sejong University (SU), and Yeonsei University (YSU).

[‡] Included only females from SNUH, YSU, SU, SMC, and CNU.

[§] Included subjects from SNUH, YSU, and SMC.

Abbreviations: KS, Kolmogorov-Smirnov, NA, not available; SD, standard deviation.

Table 3. Odds ratios of top 1% polygenic risk scores (PRSs) compared with the other PRS subgrou ps

	OR (95% CI)	*P*-value
Dataset 1	1.29 (1.26–1.42)	0.38
Dataset 2	1.20 (1.26–1.43)	0.53
Dataset 3	2.22 (1.36–1.60)	0.03
UK Biobank	1.75 (0.796–3.85)	NA
MGI	1.94 (1.22–3.1)	

Abbreviations: CI, confidence interval; MGI, Michigan Genomics Initiative; NA, not available; OR, odds ratio.

Table 4. Comparison of the area under the curves (AUC) of the prediction models with and with out polygenic risk scores (PRSs) as a covariate

	AUC with PRSs (95% confidence interval)	AUC without PRSs (95% confidence interval)	*P*-value, DeLong test
Dataset 1	0.760 (0.747–0.774)	0.750 (0.736–0.765)	1.87×10^{-5}
Dataset 2	0.764 (0.750–0.778)	0.754 (0.739–0.769)	4.01×10^{-5}
Dataset 3	0.703 (0.684–0.722)	0.673 (0.654–0.692)	1.09×10^{-7}

Figure Legends

Figure 1. Flowchart of data collection and analysis protocols. Abbreviations: CNU, Chonnam National University; GWAS, genome-wide association study; IBS, identity by state; PC, principal component; PRS, polygenic risk score; QC, quality control; Rsq, R squared; SNP, single nucleotide polymorphism; SMC, Samsung Medical Center; SNUH, Seoul National University Hospital; SU, Sejong University; YSU, Yeonsei University.

Figure 2. Odds ratios depending on percentiles of polygenic risk scores. Percentiles were defined in control subjects. Dots and vertical red lines represent the odds ratios and their 95% confidence intervals (CI), respectively. Middle quintile (40–60%) was considered as reference group. **(a)** Dataset 1 included all subjects from Chonnam National University (CNU), Samsung Medical Center (SMC), Seoul National University Hospital (SNUH), Sejong University (SU), and Yeonsei University (YSU). **(b)** Dataset 2 included only females from SNUH, YSU, SU, SMC, and CNU. **(c)** Dataset 3 included subjects from SNUH, YSU, and SMC.

Figure 3. Receiver operating characteristic curves of the different datasets. (a) Dataset 1 included all subjects from Chonnam National University (CNU), Samsung Medical Center (SMC), Seoul National University Hospital (SNUH), Sejong University (SU), and Yeonsei University (YSU). **(b)** Dataset 2 included only females from SNUH, YSU, SU, SMC, and CNU. **(c)** Dataset 3 included subjects from SNUH, YSU, and SMC. AUC, area under the curve.

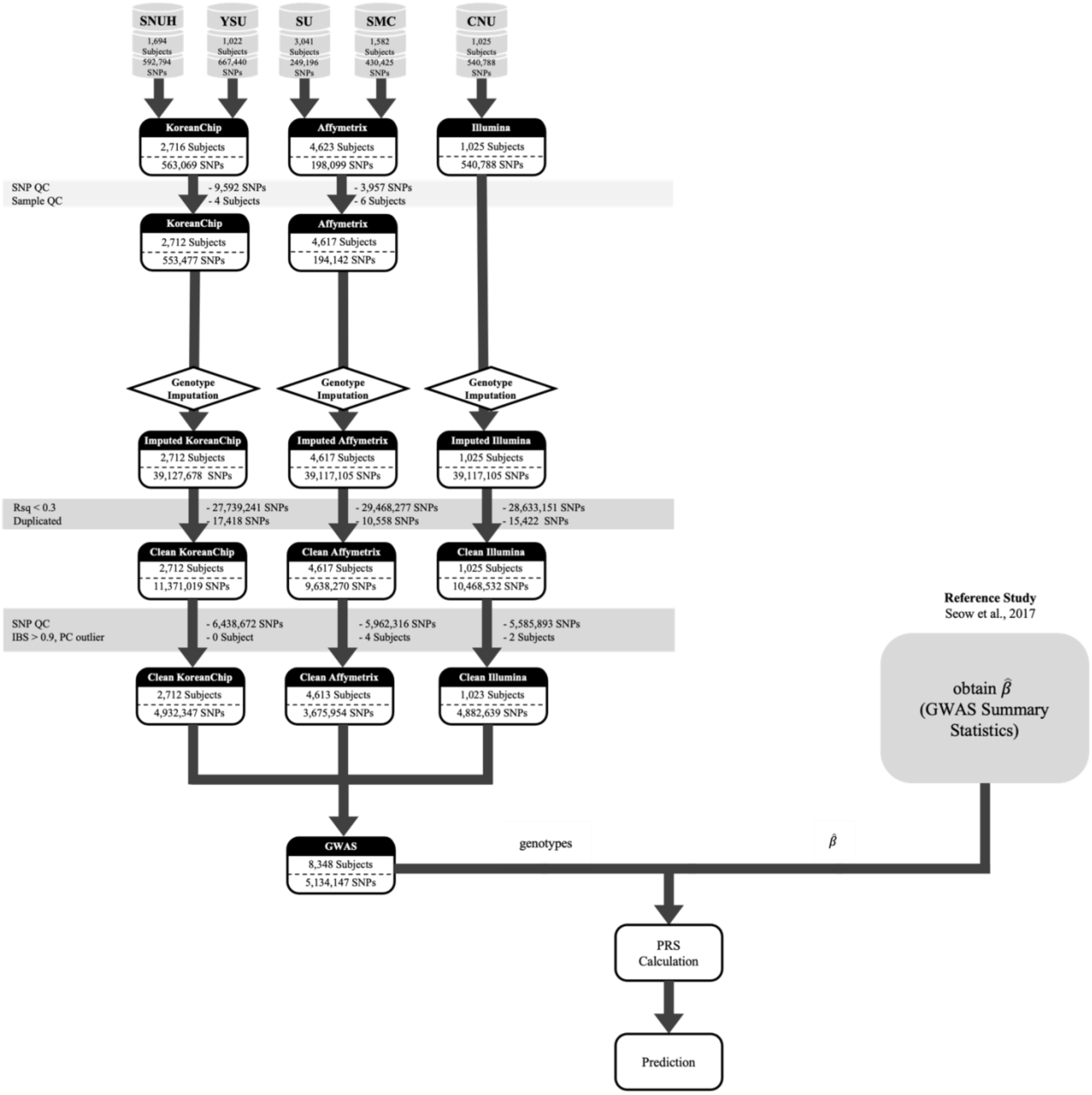

Figure 1. Flowchart of data collection and analysis protocols. Abbreviations: CNU, Chonnam National University; GWAS, genome-wide association study; IBS, identity by state; PC, principal component; PRS, polygenic risk score; QC, quality control; Rsq, R squared; SNP, single nucleotide polymorphism; SMC, Samsung Medical Center; SNUH, Seoul National University Hospital; SU, Sejong University; YSU, Yeonsei University.

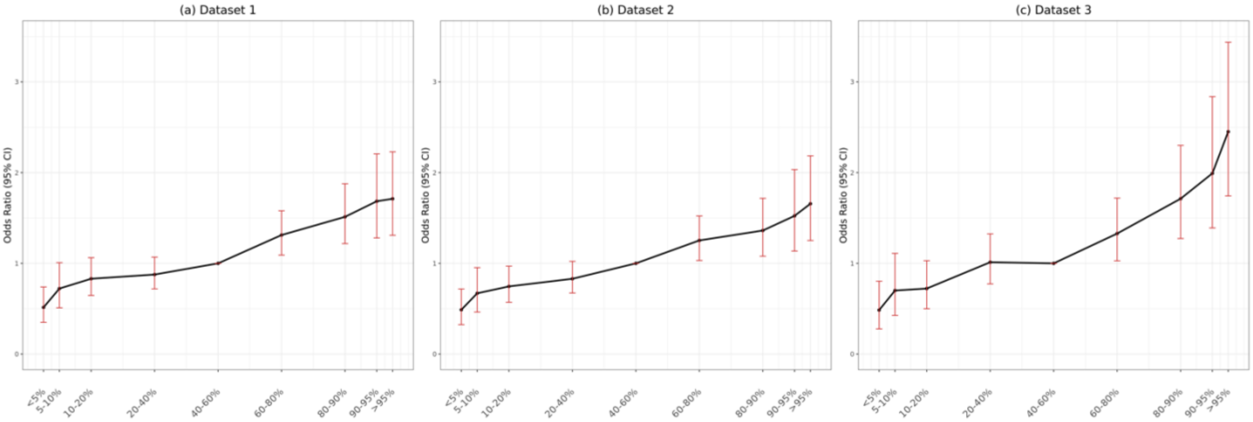

Figure 2. Odds ratios depending on percentiles of polygenic risk scores. Percentiles were defined in control subjects. Dots and vertical red lines represent the odds ratios and their 95% confidence intervals (CI), respectively. Middle quintile (40–60%) was considered as reference group. **(a)** Dataset 1 included all subjects from Chonnam National University (CNU), Samsung Medical Center (SMC), Seoul National University Hospital (SNUH), Sejong University (SU), and Yeonsei University (YSU). **(b)** Dataset 2 included only females from SNUH, YSU, SU, SMC, and CNU. **(c)** Dataset 3 included subjects from SNUH, YSU, and SMC.

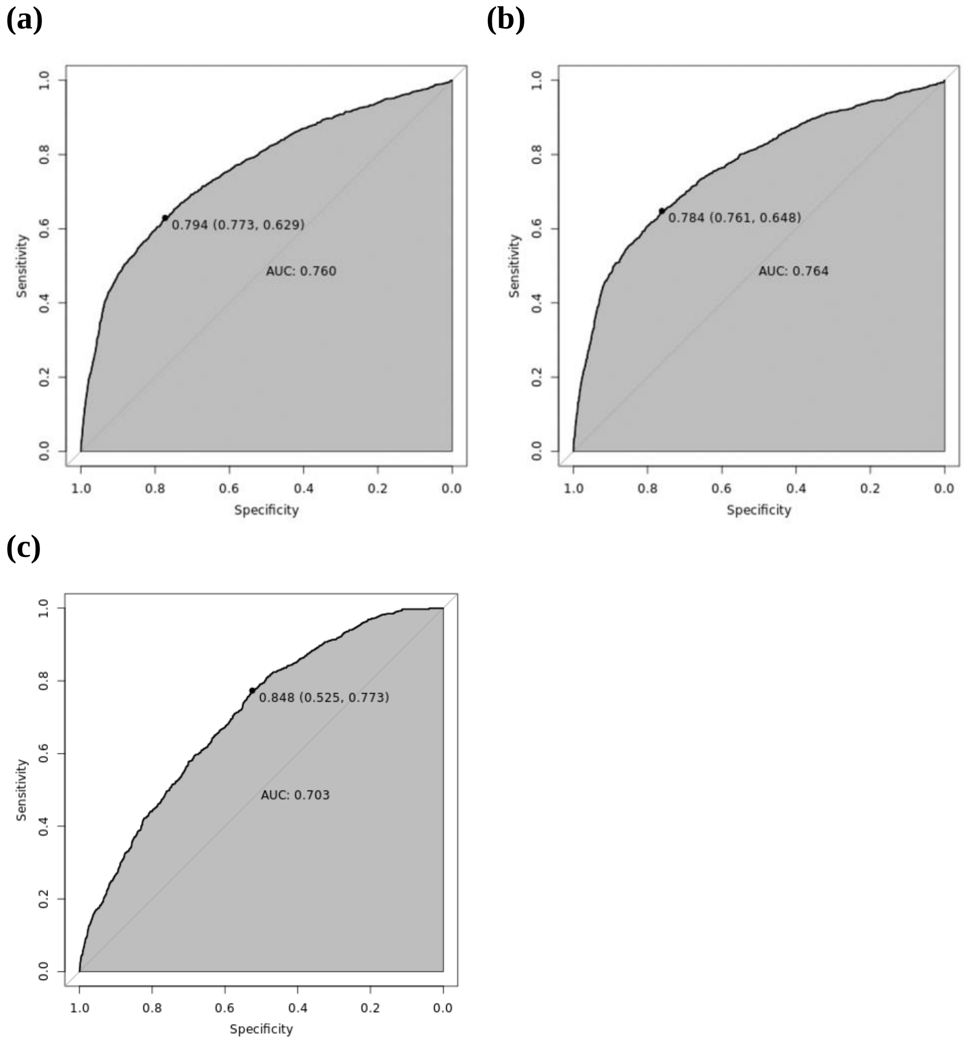

Figure 3. Receiver operating characteristic curves of the different datasets. (a) Dataset 1 included all subjects from Chonnam National University (CNU), Samsung Medical Center (SMC), Seoul National University Hospital (SNUH), Sejong University (SU), and Yeonsei University (YSU). **(b)** Dataset 2 included only females from SNUH, YSU, SU, SMC, and CNU. **(c)** Dataset 3 included subjects from SNUH, YSU, and SMC. AUC, area under the curve.