medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Stratifying Lung Adenocarcinoma Risk with Multi-ancestry Polygenic Risk Scores in East Asian Never-Smokers

Batel Blechter[1*], Xiaoyu Wang[1*], Jianxin Shi[1*], Kouya Shiraishi[2*], Jiyeon Choi[1*], Keitaro Matsuo[3*], Tzu-Yu Chen[4*], Juncheng Dai[5,6*], Rayjean J Hung[7*], Kexin Chen[8*], Xiao-Ou Shu[9*], Young Tae Kim[10], Parichoy Pal Choudhury[11], Jacob Williams[1], Maria Teresa Landi[1], Dongxin Lin[12], Wei Zheng[9], Zhihua Yin[13], Baosen Zhou[14], Jiucun Wang[15,16], Wei Jie Seow[1,17,18], Lei Song[1], I-Shou Chang[19], Wei Hu[1], Li-Hsin Chien[4,20], Qiuyin Cai[9], Yun-Chul Hong[21], Hee Nam Kim[22], Yi-Long Wu[23], Maria Pik Wong[24], Brian Douglas Richardson[1,25], Shilan Li[1,26], Tongwu Zhang[1], Charles Breeze[1], Zhaoming Wang[27], Bryan A Bassig[28], Jin Hee Kim[29], Demetrius Albanes[1], Jason YY Wong[30], Min-Ho Shin[21], Lap Ping Chung[24], Yang Yang[31], She-Juan An[23], Hong Zheng[8], Yasushi Yatabe[32], Xu-Chao Zhang[23], Young-Chul Kim[33,34], Neil E Caporaso[1], Jiang Chang[35], James Chung Man Ho[36], Michiaki Kubo[37], Yataro Daigo[38,39], Minsun Song[40], Yukihide Momozawa[37], Yoichiro Kamatani[41], Masashi Kobayashi[42], Kenichi Okubo[42], Takayuki Honda[43], H Dean Hosgood[44], Hideo Kunitoh[45], Shun-ichi Watanabe[46], Yohei Miyagi[47], Haruhiko Nakayama[48], Shingo Matsumoto[49], Hidehito Horinouchi[46], Masahiro Tsuboi[50], Ryuji Hamamoto[51], Koichi Goto[49], Yuichiro Ohe[46], Atsushi Takahashi[41], Akiteru Goto[52], Yoshihiro Minamiya[53], Megumi Hara[54], Yuichiro Nishida[54], Kenji Takeuchi[55], Kenji Wakai[55], Koichi Matsuda[56], Yoshinori Murakami[57], Kimihiro Shimizu[58], Hiroyuki Suzuki[59], Motonobu Saito[60], Yoichi Ohtaki[61], Kazumi Tanaka[61], Tangchun Wu[62], Fusheng Wei[63], Hongji Dai[8], Mitchell J Machiela[1], Jian Su[23], Yeul Hong Kim[64], In-Jae Oh[33,34], Victor Ho Fun Lee[65], Gee-Chen Chang[66,67,68,69], Ying-Huang Tsai[70,71], Kuan-Yu Che[72], Ming-Shyan Huang[73], Wu-Chou Su[74], Yuh-Min Chen[75], Adeline Seow[17], Jae Yong Park[76], Sun-Seog Kweon[21,77], Kun-Chieh Chen[67], Yu-Tang Gao[78], Biyun Qian[8], Chen Wu[12], Daru Lu[15,16], Jianjun Liu[76,79,80], Ann G Schwartz[81], Richard Houlston[82], Margaret R Spitz[83], Ivan P Gorlov[83], Xifeng Wu[84], Ping Yang[85], Stephen Lam[86], Adonina Tardon[87], Chu Chen[88], Stig E Bojesen[89,90], Mattias Johansson[91], Angela Risch[92,93,94], Heike Bickeböller[95], Bu-Tian Ji[1], H-Erich Wichmann[96,97,98], David C. Christiani[99], Gadi Rennert[100], Susanne Arnold[101], Paul Brennan[91], James McKay[91], John K Field[102], Michael P.A. Davies[102], Sanjay S Shete[103], Loic Le Marchand[104], Geoffrey Liu[105], Angeline Andrew[106], Lambertus A Kiemeney[107], Shan Zienolddiny-Narui[108], Kjell Grankvist[109], Mikael Johansson[110], Angela Cox[111], Fiona Taylor[111], Jian-Min Yuan[112], Philip Lazarus[113], Matthew B Schabath[114], Melinda C Aldrich[115], Hyo-Sung Jeon[116], Shih Sheng Jiang[19], Jae Sook Sung[64], Chung-Hsing Chen[19], Chin-Fu Hsiao[4], Yoo Jin Jung[117], Huan Guo[118], Zhibin Hu[5], Laurie Burdett[1,119], Meredith Yeager[1,119], Amy Hutchinson[1,119], Belynda Hicks[1,119], Jia Liu[1,119], Bin Zhu[1,119], Sonja I Berndt[1], Wei Wu[13], Junwen Wang[120,121], Yuqing Li[79], Jin Eun Choi[116], Kyong Hwa Park[64], Sook Whan Sung[122], Li Liu[123], Chang Hyun Kang[117], Wen-Chang Wang[124], Jun Xu[125], Peng Guan[13,126], Wen Tan[12], Chong-Jen Yu[127], Gong Yang[9], Alan Dart Loon Sihoe[128], Ying Chen[17], Yi Young Choi[116], Jun Suk Kim[129], Ho-Il Yoon[130], In Kyu Park[117], Ping Xu[131], Qincheng He[13], Chih-Liang Wang[132], Hsiao-Han Hung[19,133], Roel C.H. Vermeulen[134], Iona Cheng[135], Junjie Wu[15,16], Wei-Yen Lim[17], Fang-Yu Tsai[19,133], John K.C. Chan[136], Jihua Li[137], Hongyan Chen[15,16], Hsien-Chih Lin[4], Li Jin[15,16], Jie Liu[138], Norie Sawada[139], Taiki Yamaji[140], Kathleen Wyatt[1,119], Shengchao A. Li[1,119], Hongxia Ma[5,6], Meng Zhu[5,6], Zhehai Wang[138], Sensen Cheng[138], Xuelian Li[13,126], Yangwu Ren[13,126], Ann Chao[141], Motoki Iwasaki[139,140], Junjie Zhu[30], Gening Jiang[30], Ke Fei[30], Guoping Wu[63], Chih-Yi Chen[142], Chien-Jen Chen[143], Pan-Chyr Yang[144], Jinming Yu[138], Victoria L. Stevens[145], Joseph F. Fraumeni Jr[1], Nilanjan Chatterjee[1,146,147‡], Olga Y Gorlova[148,149,150‡],

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Christopher I Amos[148,149,150‡], Hongbing Shen[5,6‡], Chao Agnes Hsiung[4‡], Stephen J Chanock[1‡], Nathaniel Rothman[1‡], Takashi Kohno[2‡], Qing Lan[1‡], Haoyu Zhang[1‡]

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA, [2]Division of Genome Biology, National Cancer Research Institute, Tokyo, Japan, [3]Division of Molecular Medicine, Aichi Cancer Center Research Institute, Nagoya, Japan, [4]Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan, [5]Department of Epidemiology, School of Public Health, Nanjing Medical University, Nanjing, China, [6]Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Medicine, Nanjing Medical University, Nanjing, China, [7]Prosserman Centre for Population Health Research, Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, Canada, [8]Department of Epidemiology and Biostatistics, National Clinical Research Center for Cancer, Key Laboratory of Molecular Cancer Epidemiology of Tianjin, Tianjin Medical University Cancer Institute and Hospital, Tianjin Medical University, Tianjin, China, [9]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center and Vanderbilt-Ingram Cancer Center, Nashville, TN, USA, [10]Cancer Research Institute, Seoul National University College of Medicine, Seoul National University Hospital, Seoul, Seoul, Republic of Korea, [11]American Cancer Society, Atlanta, GA, USA, [12]Department of Etiology & Carcinogenesis and State Key Laboratory of Molecular Oncology, Cancer Institute and Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, [13]Department of Epidemiology, School of Public Health, China Medical University, Shenyang, China, [14]Department of Clinical Epidemiology and Center of Evidence Based Medicine, The First Hospital of China Medical University, Shenyang, China, [15]Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai, China, [16]State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai, China, [17]Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore, Singapore, [18]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore, [19]National Institute of Cancer Research, National Health Research Institutes, Zhunan, Taiwan, [20]Department of Applied Mathematics, Chung Yuan Christian University, Chong-Li, Taiwan, [21]Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea, [22]Department of Preventive Medicine, Chonnam National University Medical School, Gwangju, Republic of Kore, [23]Guangdong Lung Cancer Institute, Guangdong Provincial Key Laboratory of Translational Medicine in Lung Cancer, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, China, [24]Department of Pathology, Queen Mary Hospital, Hong Kong, Hong Kong, [25]Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA, [26]Department of Biostatistics, Bioinformatics & Biomathematics, Georgetown University Medical Center, Washington, DC, USA, [27]Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN, USA, [28]Saville Cancer Screening and Prevention Center, Inova Schar Cancer Institute, Inova Health System, Fairfax, VA, USA, [29]Department of Environmental Health, Graduate School of Public Health, Seoul National University, Seoul, Republic of Korea,, [30]Epidemiology and Community Health Branch, National Heart Lung and Blood Institute, Bethesda, MD, USA, [31]Shanghai Pulmonary Hospital, Shanghai, China, [32]Department of Pathology and Clinical Laboratories, National Cancer Center Hospital, Tokyo,

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Japan, [33]Lung and Esophageal Cancer Clinic, Chonnam National University Hwasun Hospital, Hwasuneup, Republic of Korea, [34]Department of Internal Medicine, Chonnam National University Medical School, Gwangju, Republic of Korea, [35]Cancer Institute and Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, [36]Department of Medicine, School of Clinical Medicine, The University of Hong Kong, Queen Mary Hospital, Hong Kong, Hong Kong, [37]Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan, [38]Center for Antibody and Vaccine Therapy, Research Hospital, Institute of Medical Science, The University of Tokyo, Tokyo, Japan, [39]Department of Medical Oncology and Cancer Center, and Center for Advanced Medicine against Cancer, Shiga University of Medical Science, Shiga, Japan, [40]Department of Statistics & Research Institute of Natural Sciences, Sookmyung Women's University, Seoul, 04310, Korea, [41]Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan, [42]Department of Thoracic Surgery, Tokyo Medical and Dental University, Tokyo, Japan, [43]Department of Respiratory Medicine, Tokyo Medical and Dental University, Tokyo, Japan, [44]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, New York, USA, [45]Department of Medical Oncology, Japanese Red Cross Medical Center, Tokyo, Japan, [46]Department of Thoracic Surgery, National Cancer Center Hospital, Tokyo, Japan, [47]Molecular Pathology and Genetics Division, Kanagawa Cancer Center Research Institute, Yokohama, Japan, [48]Department of Thoracic Surgery, Kanagawa Cancer Center, Yokohama, Japan, [49]Department of Thoracic Oncology, National Cancer Center Hospital East, Kashiwa, Japan, [50]Department of Thoracic Surgery, National Cancer Center Hospital East, Kashiwa, Japan, [51]Division of Medical AI Research and Development, National Cancer Center Research Institute, Tokyo, Japan, [52]Department of Cellular and Organ Pathology, Graduate School of Medicine, Akita University, Akita, Japan, [53]Department of Thoracic Surgery, Graduate School of Medicine, Akita University, Akita, Japan, [54]Department of Preventive Medicine, Faculty of Medicine, Saga University, Saga, Japan, [55]Department of Preventive Medicine, Nagoya University Graduate School of Medicine, Nagoya, Japan, [56]Laboratory of Clinical Genome Sequencing, Department of Computational Biology and Medical Science, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan, [57]Department of Molecular Biology, Institute for Advanced Medical Sciences, Nippon Medical School, Tokyo, Japan, [58]Department of Surgery, Division of General Thoracic Surgery, Shinshu University School of Medicine Asahi, Nagano, Japan, [59]Department of Chest Surgery, Fukushima Medical University School of Medicine, Fukushima, Japan, [60]Department of Gastrointestinal Tract Surgery, Fukushima Medical University School of Medicine, Fukushima, Japan, [61]Department of Integrative center of General Surgery, Gunma University Hospital, Gunma, Japan, [62]Institute of Occupational Medicine and Ministry of Education Key Lab for Environment and Health, School of Public Health, Huazhong University of Science and Technology, Wuhan, China, [63]China National Environmental Monitoring Center, Beijing, China, [64]Department of Internal Medicine, Division of Oncology/Hematology, College of Medicine, Korea University Anam Hospital, Seoul, Republic of Korea, [65]Department of Clinical Oncology, The University of Hong Kong, Queen Mary Hospital, Hong Kong, Hong Kong, [66]School of Medicine and Institute of Medicine, Chung Shan Medical University, Taichung, Taiwan, [67]Department of Internal Medicine, Division of Pulmonary Medicine, Chung Shan Medical University Hospital, Taichung, Taiwan, [68]Institute of Biomedical Sciences, National Chung Hsing University, Taichung, Taiwan, [69]Department of Internal Medicine, Division of Chest Medicine, Taichung Veterans General Hospital, Taichung, Taiwan, [70]Department of Respiratory Therapy, Chang Gung University, Taoyuan, Taiwan,

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[71]Department of Pulmonary and Critical Care, Xiamen Chang Gung Hospital, Xiamen, China, [72]Department of Internal Medicine, National Taiwan University Hospital and College of Medicine, Taipei, Taiwan, [73]Department of Internal Medicine, E-Da Cancer Hospital, I-Shou University and Kaohsiung Medical University, Kaohsiung, Taiwan, [74]Department of Oncology, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan, [75]Department of Chest Medicine, Taipei Veterans General Hospital, and school of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan, [76]Lung Cancer Center, Kyungpook National University Medical Center, Daegu, Republic of Korea, [77]Jeonnam Regional Cancer Center, Chonnam National University, Hwasun, Republic of Korea, [78]Department of Epidemiology, Shanghai Cancer Institute, Shanghai, China, [79]Department of Human Genetics, Genome Institute of Singapore, Singapore, Singapore, [80]School of Life Sciences, Anhui Medical University, Hefei, China, [81]Karmanos Cancer Institute, Wayne State University, Detroit, MI, USA, [82]Division of Genetics and Epidemiology, Institute of Cancer Research, London, UK, [83]Department of Medicine, Section of Epidemiology and Population Science, Institute for Clinical and Translational Research, Houston, TX, USA, [84]School of Medicine, Zhejiang University, Hangzhou, Zhejiang, China, [85]Department of Health Sciences Research, Mayo Clinic, Scottsdale, AZ, USA, [86]Department of Integrative Oncology, BC Cancer, Vancouver, BC, Canada, [87]Health Research Institute of Asturias (ISPA), Oviedo and Universidad Nebrija, Madrid, Spain, [88]Program in Epidemiology, Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, Washington, USA, [89]Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark, [90]Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Copenhagen, Denmark, [91]International Agency for Research on Cancer (IARC/WHO), Lyon, France, [92]German Cancer Research Center (DKFZ), Heidelberg, Germany, [93]Translational Lung Research Center Heidelberg (TLRC-H), Member of the German Center for Lung Research (DZL), Heidelberg, Germany, [94]Department of Biosciences and Medical Biology and Center for Tumorbiology and Immunology (CTBI), University of Salzburg and Cancer Cluster Salzburg, Salzburg, Austria, [95]University Medical Center Goettingen, Goettingen, Germany, [96]Institute of Medical Informatics, Biometry and Epidemiology, Ludwig Maximilians University, Munich, Germany, [97]Helmholtz Center Munich, Institute of Epidemiology, Munich, Germany, [98]Institute of Medical Statistics and Epidemiology, Technical University Munich, Munich, Germany, [99]Harvard TH Chan School of Public Health, Boston, Massachusetts, USA, [100]Carmel Medical Center, Israel, [101]Markey Cancer Center, Lexington, KY, USA, [102]Liverpool University, Liverpool, UK, [103]The University of Texas MD Anderson Cancer Center, Houston, Texas, USA, [104]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA, [105]Princess Margaret Cancer Center, Toronto, ON, Canada, [106]Norris Cotton Cancer Center, Lebanon, NH, USA, [107]Radboud University Medical Center, Nijmegen, Netherlands, [108]National Institute of Occupational Health, Oslo, Norway, [109]Department of Medical Biosciences, Umeå University, Umeå, Sweden, [110]Department of Radiation Sciences, Umeå University, Umeå, Sweden, [111]University of Sheffield, Sheffield, UK, [112]UPMC Hillman Cancer Center and Department of Epidemiology, School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA, [113]D, School of Pharmacy and Pharmaceutical Sciences, SUNY University at Buffalo, NY, USA, [114]Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA, [115]Department of Thoracic Surgery, Division of Epidemiology, Vanderbilt University Medical Center, Nashville, TN, USA, [116]Cancer Research Center, Kyungpook National University Medical Center, Daegu, Republic of Korea, [117]Department of Thoracic and

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Cardiovascular Surgery, Cancer Research Institute, Seoul National University College of Medicine, Seoul, Republic of Korea, [118]Department of Occupational and Environmental Health and Ministry of Education Key Lab for Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [119]Cancer Genomics Research Laboratory, Frederick National Laboratory for Cancer Research, Leidos Biomedical Research Inc, Rockville, MD, USA, [120]Division of Applied Oral Sciences & Community Dental Care, Faculty of Dentistry, The University of Hong Kong, Hong Kong, China, [121]State Key Laboratory of Pharmaceutical Biotechnology, The University of Hong Kong, Hong Kong, Chinaboratory of Pharmaceutical Biotechnology, The University of Hong Kong, Hong Kong, China, [122]State Key Laboratory of Pharmaceutical Biotechnology, The University of Hong Kong, Hong Kong, China, [123]Department of Oncology, Cancer Center, Union Hospital, Huazhong University of Science and Technology, Wuhan, China, [124]The Ph.D. Program for Translational Medicine, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan, [125]School of Public Health, Li Ka Shing (LKS) Faculty of Medicine, The University of Hong Kong, Hong Kong, China, [126]Key Laboratory of Cancer Etiology and Intervention, University of Liaoning Province, Shenyang, China, [127]Department of Internal Medicine, National Taiwan University Hospital Hsin-Chu Branch, Hsinchu, Taiwan, [128]Gleneagles Hong Kong Hospital, Hong Kong, China, [129]Department of Internal Medicine, Division of Medical Oncology, College of Medicine, Korea University Guro Hospital, Seoul, Republic of Korea, [130]Department of Internal Medicine, Seoul National University Bundang Hospital, Seongnam, Republic of Korea, [131]Department of Oncology, Wuhan Iron and Steel (Group) Corporation Staff-Worker Hospital, Wuhan, China, [132]Division of Pulmonary Oncology and Interventional Bronchoscopy, Department of Thoracic Medicine, Chang Gung Memorial Hospital, Taoyuan, Taiwan, [133]Division of Thoracic Surgery, Department of Surgery, Chung Shan Medical University Hospital, Taichung, Taiwan, [134]Division of Environmental Epidemiology, Institute for Risk Assessment Sciences (IRAS), Utrecht University, Utrecht, The Netherlands, [135]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA, [136]Department of Pathology, Queen Elizabeth Hospital, Hong Kong, China, [137]Qujing Center for Diseases Control and Prevention, Qujing, China, [138]Department of Oncology, Shandong Cancer Hospital and Institute, Shandong Academy of Medical Sciences, Jinan, China, [139]Division of Cohort Research, National Cancer Center Institute for Cancer Control, Tokyo, Japan, [140]Division of Epidemiology, National Cancer Center Institute for Cancer Control, Tokyo, Japan, [141]Center for Global Health, National Cancer Institute, Bethesda, MD, USA, [142]Institute of Medicine, Chung Shan Medical University, Taichung, Taiwan, [143]Genomic Research Center, Academia Sinica, Taipei, Taiwan, [144]Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan, [145]Laboratory Services, American Cancer Society, Georgia, USA, [146]Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD, USA, [147]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA, [148]Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX, USA, [149]Section of Epidemiology and Population Sciences, Department of Medicine, Baylor College of Medicine, Houston, TX, USA, [150]Dan L Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX, USA

[*] These authors contributed equally to this work

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[‡] These authors jointly directed this work

Correspondence to: Batel Blechter (batel.blechter@nih.gov), Qing Lan (qingl@mail.nih.gov), and Haoyu Zhang (haoyu.zhang2@nih.gov)

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Abstract

Polygenic risk scores (PRSs) are promising for risk stratification but have mainly been developed in European populations. This study developed single- and multi-ancestry PRSs for lung adenocarcinoma (LUAD) in East Asian (EAS) never-smokers using genome-wide association study summary statistics from EAS (8,002 cases; 20,782 controls) and European (2,058 cases; 5,575 controls) populations. A multi-ancestry PRS, developed using CT-SLEB, was strongly associated with LUAD risk (odds ratio=1.71, 95% confidence interval (CI):1.61,1.82), with an area under the receiver operating curve value of 0.640 (95% CI:0.629,0.653). Individuals in the highest 20% of the PRS had nearly four times the risk compared to the lowest 20%. Individuals in the 95th percentile of the PRS had an estimated 6.69% lifetime absolute risk. Notably, this group reached the average population 10-year LUAD risk at age 50 (0.42%) by age 41. Our study underscores the potential of multi-ancestry PRS approaches to enhance LUAD risk stratification in EAS never-smokers.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Key Words: Polygenic risk scores, Lung adenocarcinoma, Genome-wide association studies, East Asian never smokers, Lifetime absolute risk

Abbreviations:

EAS, East Asian; LUAD, lung adenocarcinoma; GWAS, genome-wide association study; EUR, European; PRS, polygenic risk score; SNP, single nucleotide polymorphism; FLCCA, Female Lung Cancer Consortium in Asia; LD, linkage disequilibrium; AUC, area under the receiver operating curve; PC, principal component; OR, odds ratio; CI, confidence interval; USPSTF, United States Preventive Services Taskforce

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Introduction

Lung cancer is a major global health challenge, responsible for about 20% of all cancer deaths in 2020[1]. While smoking is the primary etiologic factor, around 25% of lung cancer cases occur in never-smokers, with significant geographical variations[2] . Notably, in East Asia (EAS), never-smoking women exhibit high incidences of lung adenocarcinoma (LUAD), the predominant histologic subtype[3].

Beyond smoking, risk factors for LUAD include environmental and occupational exposures, lifestyle, family history, and genetic susceptibility[4–7]. Genome-wide association studies (GWAS) for never-smoking lung cancer across EAS and European (EUR) populations have identified 28 susceptibility loci at 25 independent regions[8–13]. These GWAS findings enhance our understanding of LUAD's genetic architecture, yet the translation of these findings into clinical practice requires further investigation.

Polygenic risk scores (PRS) aggregate the effects of individual single nucleotide polymorphisms (SNPs) to estimate genetic risk for diseases, playing a crucial role in cancer prevention and screening. Previous efforts primarily utilized EUR populations for lung cancer PRS development[14,15] to better identify high-risk individuals. Nonetheless such PRSs often fall short for EAS never-smokers, reflecting a bias in screening guidelines and the focus of studies on populations with EUR ancestry. Despite ongoing efforts to integrate PRS into clinical practices for complex diseases, as seen in initiatives like the electronic Medical Records and Genomics (eMERGE) network[16], Veterans Affairs Genomic Medicine at Veteran Affairs (GenoVA) study[17], and the Women Informed to Screen Depending On Measures of risk (WISDOM) study[18], lung cancer has been excluded. The predominant global burden of lung cancer, driven primarily by smoking, overlooks the significant impact and potential benefits of PRS in never-smoking populations.

The recent expansion of GWAS across diverse populations, combined with advancements in PRS methodology, underscores a pivotal shift towards enhancing disease outcome prediction beyond EUR populations[19]. Notably, the development of multi-ancestry PRS methods marks a

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

considerable advancement, leveraging data from various ancestral backgrounds to enrich the predictive accuracy and robustness of PRSs[20–26]. By jointly modeling genetic data from multiple populations, these approaches improve the predictive power of PRSs, thereby enhancing disease outcome predictions in non-EUR populations.

While the development and validation of PRSs are critical, the application of an established PRS in estimating the absolute risk of a disease offers valuable insights into risk stratification, potentially guiding clinical interventions, such as lung cancer screening. Moreover, projecting necessary sample sizes for future studies to achieve varying levels of PRS predictive accuracy is an underexplored area that could significantly impact epidemiological research design and cost-effectiveness.

This study aims to evaluate the predictive performance of single- and multi-ancestry PRSs for LUAD in never-smoking EAS individuals using state-of-the-art methodologies. By integrating multiple datasets, we constructed PRSs based on summary statistics for 5,622 never-smoking cases and 21,813 never-smoking controls from EAS and EUR ancestries. We assessed the performance of the PRSs using independent, individual level data of 4,438 never-smoking EAS cases and 4,544 never-smoking EAS controls. Furthermore, we estimated the lifetime and 10-year absolute risks of LUAD using the most accurate PRSs developed. Lastly, we projected the sample sizes needed in future research to achieve specific levels of prediction accuracy with PRSs in EAS never-smokers (**Figure 1**).

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Methods

Study design and data sources

Training data compilation: study population and genotyping

The studies, genotyping protocols, and quality control for PRS construction are previously described in detail[27]. Briefly, EAS single-ancestry PRSs for LUAD were constructed using summary data from 3,564 never-smoking LUAD cases and 16,238 never-smoking controls of EAS ancestry from the Nanjing Lung Cancer Study (NJLCS)[9,28], National Cancer Center of JAPAN (NCC), and the Research Institute and Aichi Cancer Center (ACC). The NJLCS study combined data from several cities, genotyped by Affymetrix Genome-Wide Human SNP Array 6.0, as well as the Oncoarray GWAS[28–30]. The NCC study included lung cancer patients and control data from multiple Japanese, with genotyping conducted on Illumina HumanOmniExpress and HumanOmni1-Quad genotyping platforms. Similarly, the ACC study pooled data from the multiple Japanese medical institutions, and the Nagahama Study, with genotyping performed using Illumina 610k and Illumina660k platforms. To construct the multi-ancestry PRS, we further incorporated GWAS summary statistics from 2,058 never-smoking LUAD cases and 5,575 never-smoking controls of EUR ancestry, genotyped using Illumina Infinium OmniExpress-24 v1.2 BeadChips and Illumina Human660W-Quad BeadChip[13].

Tuning and validation: study population and genotyping

For PRS tuning and validation, we used 4,438 never-smoking LUAD cases and 4,544 never-smoking controls from the Female Lung Cancer Consortium in Asia (FLCCA)[10,11], an international consortium composed of never-smoking EAS women from regions including Mainland China, Singapore, Taiwan, Hong Kong, South Korea and Japan [10,11]. All samples were genotyped using the Illumina 660W, 370K or 610Q microarrays. The FLCCA data were randomly and equally divided for tuning (2,219 cases, 2,272 controls) and validation (2,219 cases, 2,272 controls). Informed consent was obtained from all participants, and the recruitment and data collection procedures were approved by the local ethics review committees of all participating research institutes.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Construction of PRS

Single-ancestry PRS methods

PRS-25: We constructed a simple PRS using the 25 independent SNPs (PRS-25) that reached genome-wide significance (i.e., $P<5x10^{-8}$) in the largest GWAS for LUAD in EAS population to date[8]. The PRS-25 was calculated by summing the risk alleles weighed by their effect sizes (i.e., per allele log-odds ratio) obtained from a meta-analysis including only never-smokers.

PRS-CT: We generated a PRS using the CT method[31] using PLINK 1.90[32]. This involved the clumping of SNPs based on linkage disequilibrium (LD) with a threshold of $r^2=0.1$ within a 500kb window. The LD were estimated from half of the FLCCA dataset used for tuning dataset. We created nine SNP subsets by applying incremental P-value thresholds (i.e., $P<5x10^{-8}$, $P<5x10^{-7}$,…, $P<5x10^{-1}$, 1), and calculated their respective PRSs using PLINK2[32] command "--score cols=+scoresums,-scoreavgs no-mean-imputation". Using the tuning dataset, the AUC was calculated for each threshold to identify the most predictive P-value cutoff.

LDpred2-PRS: We applied the LDpred2 method, implemented in the R package bigsnpr[33], using a Bayesian framework to estimate SNP effect sizes through a shrinkage estimator. This method leverages GWAS summary statistics, incorporating a prior for effect sizes, while also accounting for LD across SNPs. Our analysis was constrained to HapMap3 variants, and we calibrated the model using a range of hyperparameters: the proportion of causal SNPs was set across a 21-point logarithmic sequence from 10^{-5} to 1 , and the per-SNP heritability was set as a fraction (0.3, 0.7, 1 or 1.4) of the total heritability estimated by LD score regression[34] divided by the number of causal SNPs. Lastly, we used the "sparse" option to set weaker effects to zero. The optimal tuning parameters were selected based on the highest AUC achieved on the tuning dataset.

PRS-EUR 128: We evaluated the performance of a European-derived PRS (PRS-EUR 128)[35], which consists of 128 variants obtained from a GWAS of predominately active smokers, on EAS populations to assess its cross-population applicability.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Multi-ancestry PRSs methods:

LDpred2 PRS + PRS-EUR 128: For the multi-ancestry PRSs, we applied the weighted-PRS approach by linearly combining the most predictive single-ancestry PRS from the EAS population (Ldpred2 PRS) with the EUR population-specific PRS (PRS-EUR 128). The weights of this combined PRS were calculated by applying a logistic regression on the tuning dataset in R version 4.2.0.

PRS-CSx: We applied PRS-CSx, a multi-ancestry polygenic prediction method that uses a Bayesian framework with a continuous shrinkage prior to estimate SNP effect sizes from GWAS summary statistics across different populations[26]. LD reference panels for EUR and EAS, provided by the PRS-CSx software, were constructed from the 1000 Genomes Project (1KG) samples [36]. We used the default setting with the gamma-gamma prior hyperparameters (a and b) at 1 and 0.5, respectively. The shrinkage parameter ϕ was assessed at 10^{-6}, 10^{-4}, 10^{-2}, and 1 within the tuning dataset to select the value with highest AUC. With the optimal ϕ , we calculated the PRS using posterior effect-sizes from both EAS and EUR populations. Weights were then estimated to linearly combine the two PRSs on the tuning dataset. The final performance was then evaluated on an EAS independent validation dataset.

CT-SLEB: We used the recently developed CT-SLEB method to derive a PRS using EAS and EUR datasets [24]. This method extends the standard CT method by designing a two-dimensional approach to select SNPs for EAS PRS construction. It incorporates an empirical Bayesian (EB) framework to model genetic correlations between EAS and EUR. Following this, a super-learning (SL) model is then applied to integrate multiple PRSs, each generated using distinct p-value thresholds and clumping parameters[37]. Our implementation of CT-SLEB followed the default setting with p-value thresholds $p_t = 5\text{x}10^{-8}, 5\text{x}10^{-7}, \ldots, 5\text{x}10^{-1}$, or 1, and genetic distances $d = 50/r^2$ or $100/r^2$, where $r^2 = 0.01, 0.05, 0.1, 0.2, 0.5\ or\ 0.8$. The clumping process for EAS was based on the LD reference derived from the FLCCA tuning dataset, and the EUR LD reference was based on the 1KG EUR samples.

Relative and absolute risk calculation for PRSs

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

To evaluate each PRS's performance, we standardized the scores to a unit standard deviation, based on the distribution among the control group in the EAS validation dataset. We then quantified the association between the standardized PRSs and LUAD risk using the OR and 95% CI, via logistic regression, adjusting for age and the first 10 principal components (PCs). We evaluated the predictive performance of the different PRS models through the adjusted AUC values, accounting for age and top 10 PCs, using the R package RISCA[37]. We further estimated the ORs of each PRS for risk of LUAD based on PRS percentiles, setting the middle quantile (40-60%) as the reference category. Lastly, we used the likelihood ratio test to evaluate the interaction between the PRS and age (<40, 40-49, 50-59, 60-69, ≥70).

For absolute risk estimates, we used the iCARE software[38] to calculate the cumulative lifetime (age 30-80) and 10-year absolute risks of LUAD among never-smoking female controls in FLCCA (N=4,544). Absolute risks were derived by applying the Cox proportional hazard model with the top performing PRS (CT-SLEB) and age-specific lung cancer incidence and mortality rates in Taiwan[39]. We estimated the absolute risk of LUAD overall, as well as stratified by first-degree family history of lung cancer.

Projections of CT and LDpred2 PRS performance by sample size

We used GENESIS package [40] to estimate the sample sizes needed for PRS to reach various AUC levels in the EAS population. This method estimates the expected number of SNP discoveries and their explained heritability in future studies. Using the GWAS summary statistics from our EAS training dataset, and the provided LD scores for EAS populations from the 1KG dataset, we projected the AUC for CT PRS across various case-control ratios, from 1:1 to 1:10, and case numbers ranging from 5,000 to 200,000. Given GENESIS's specific design for CT PRS, we developed a extend its application to LDpred2 PRS projections. This involved modeling the relationship between effective sample sizes with the phenotypic variance ratio between LDpred2 and CT PRS (**Supplementary Figure 1)**, elaborated in the **Supplementary Note**.

Results

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Development and validation of the PRS

We applied several cutting-edge single-and multi-ancestry PRS methods (**Methods, Supplementary Figure 2**), evaluating their performance in terms of relative risk and AUC within the EAS validation dataset (**Methods**, **Table 1, Supplementary Figure 2-3)**. Among single-ancestry PRS methods, the LDpred2 PRS, incorporating 942,591 SNPs, outperformed other methods, with an odds ratio (OR) per unit standard deviation (SD) of 1.62 (95% confidence interval (CI): 1.52, 1.73) alongside an adjusted AUC of 0.629 (95% CI: 0.618, 0.641). In contrast, the PRS designed exclusively for EUR populations underperformed in our EAS population, with an estimated adjusted AUC of 0.489 (95% CI: 0.477, 0.501), possibly attributed to the inclusion of smokers in the development of the existing EUR PRS.

Among multi-ancestry PRS methods, the weighted-PRS of EAS and EUR yielded an OR per unit SD of 1.62 (95% CI: 1.52, 1.72), with PRS-CSx and CT-SLEB showing even stronger association with ORs of 1.67 (95% CI: 1.57, 1.78) and 1.71 (95% CI: 1.61, 1.82), respectively. Among all the PRS methods, CT-SLEB had the highest AUC of 0.640 (95% CI: 0.629, 0.653). Setting individuals in the middle PRS quantile (40th to 60th percentile) as the reference category, those in the highest 5% of risk for the top performing PRS, CT-SLEB, had 4.17 (95% CI: 3.20, 5.47)-fold risk, whereas those in the lowest 5% had 0.33 (95% CI: 0.21, 0.51)-fold risk of developing LUAD (**Figure 2**).

Interaction between CT-SLEB PRS and age

We observed a significant multiplicative interaction between the CT-SLEB PRS and age at diagnosis for LUAD risk (p-interaction = 0.002). As illustrated in **Figure 3**, the OR per SD of PRS was higher in younger individuals, particularly those under 40 years of age (OR=1.70, 95% CI: 1.42, 2.05) compared to individuals aged 70 and older (OR=1.53, 95% CI: 1.37, 1.72). The association followed a roughly linear pattern across middle-age categories, peaking between ages 60-69 (OR = 1.93, 95% CI: 1.79-2.08). This pattern suggests a stronger impact of genetic risk factors in younger age groups. A similar trend was observed when analyzing the association between LUAD risk and a dichotomized PRS (upper 90th percentile) by age categories.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Absolute risk of developing lung adenocarcinoma by CT-SLEB PRS

Using the iCARE package[41], we estimated the absolute risk of LUAD for never-smoking EAS women, utilizing the CT-SLEB PRS which demonstrated the highest AUC (**Table 1**). The cumulative lifetime risk of LUAD, assessed between age 30-80, varied markedly across PRS percentiles, ranging from a minimal mean of 0.78% in the 5th percentile to a substantial mean of 6.69% in the 95th percentile (**Figure 4)**. Additionally, the 10-year absolute risk for LUAD in never-smoking 50-year-old women, varied from a mean of 0.13% in the 5th percentile to mean of 1.11% in the 95th percentile of the PRS distribution. While the average 10-year absolute risk for this age group is 0.42%, never-smoking women in the >95 and 90-95th percentiles PRS reached this risk earlier, at age of 41 and 42, respectively.

To evaluate the combined effect of PRS and family history, we further modeled the absolute risk for LUAD in individuals with (n=86) and without (n=1,414) family history of lung cancer in a first-degree relative (**Figure 5, Supplementary Figure 4**). We observed that those with family history of lung cancer have higher absolute lifetime risk of LUAD than those without family history. The lifetime absolute risk between age 30-80 for women at the bottom 5th and top 95th percentile of the PRS with family history of lung cancer were 2.77% and 9.39%, respectively (Supplementary Figure 4). Additionally, the 10-year absolute risk for women in the 95th PRS percentile with a family history reached the population average 10-year absolute risk at age 50 (0.42%) much earlier, at age 38. Formal tests for interaction between PRS and family history were not statistically significant (p=0.45), suggesting that family history and PRS independently contribute to LUAD risk stratification.

Projections of polygenic risk score performance by sample size

Using the GENESIS model, we projected the expected AUC of PRS-CT and LDpred2 PRS under varying GWAS sample sizes and case-control ratios for LUAD in never smoking EAS women (**Figure 6**). Our analysis, using the EAS training dataset, estimated approximately 1,772 (s.e.=1,641) susceptibility variants that are independently associated with LUAD[27]. This high

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

number of variants underscores the extensive polygenic nature of LUAD, implying relatively small effect sizes for individual SNPs. Based on our current data, the expected AUC for the LDpred2 PRS stands at 0.631 (95% CI: 0.618, 0.641), aligning closely with the actual performance of the LDpred2 PRS in the validation dataset. The upper boundary of our predictions, based solely on PRS, suggests an AUC potential of up to 0.731 (95% CI: 0.640,0.786). As sample sizes increase in our projections, the AUC is estimated to rise to 0.673 (95% CI: 0.593–0.725) with 55,000 EAS LUAD cases and a 1:1 case-control ratio. With the same number of cases but a 1:10 case-control ratio, the AUC is projected to further increase to 0.698 (95% CI: 0.621–0.727), nearing the theoretical upper limit of predictive accuracy.

Explaining genetic variance through PRS across different sample sizes

We also evaluated the genetic variance explained by the PRS under different sample sizes given the projected results (**Supplementary Note, Table 2**). The current LDpred2 PRS explained 26.6% of the genome-wide chip heritability, contributing to approximately 16.5% of the 1.84-fold familial relative risk for lung cancer among EAS never-smokers. Concurrently, genome-wide chip variants heritability explained 61.9% of the 1.84-fold familial relative risk associated with the disease. With an expanded sample size of 35,000 cases and 350,000 controls, the constructed PRS is projected to account for 57.9% of the genome-wide chip variants heritability, and 35.8% of the 1.84-fold familial relative risk for the disease.

Discussion

We developed and validated single- and multi-ancestry PRSs for LUAD in never-smoking EAS individuals using the largest GWAS dataset of never-smokers to date. The multi-ancestry PRS method, CT-SLEB, integrating summary data from EAS and EUR never-smokers, emerged as the best-performing PRS. It exhibited a dose-response relationship with LUAD risk and achieved higher AUC than all other evaluated PRSs. Further, our analysis demonstrates the potential of PRS in stratifying individuals' 10-year and lifetime risk of developing LUAD. Lastly, we

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

projected the expected discriminatory accuracy of the PRS across a range of sample sizes and case-control ratios.

To date, GWAS and subsequent PRS models have largely centered on EUR populations [42,43], rendering them less precise when applied to non-EUR populations and risking the exacerbation of health disparities [44,45]. Consistent with prior research [15], our study observed that a lung cancer PRS constructed using EUR data, which included both smokers and never-smokers, significantly underperformed when applied to EAS never-smoking individuals (AUC = 0.489, 95% CI: (0.477, 0.501)), further highlighting the heterogeneity of the disease across populations and the need to expand risk assessment efforts to non-EUR populations. Notably, by utilizing novel methods to integrate GWAS data from both EUR and EAS populations, we demonstrated that a multi-ancestry PRS enhances the precision of risk stratification for LUAD among EAS never-smokers.

Large-scale efforts to integrate PRS into clinical practice have focused on conditions with well-established PRS prediction performance[16–18], such as Type 2 Diabetes, breast cancer, and cardiovascular diseases. Lung cancer has been notably absent from these efforts, primarily due to the focus on smoking as a risk factor. This has limited the exploration and application of PRS in lung cancer risk stratification, particularly among never-smokers. Recent efforts, such as the TALENT study in Taiwan[46], have evaluated low-dose computed tomography (LDCT) screening for never-smokers with traditional risk factors like family history, passive smoking exposure, and chronic lung diseases. However, a PRS has not yet been integrated into these screening protocols. Further, a recent study in Taiwan incorporating genetic susceptibility in a risk model with lifestyle and environmental risk factors observed an improvement in the AUC from 0.697 to 0.714[39]. Notably, the study found that even by incorporating only 11 susceptibility loci, the model's discriminative power had improved. Our study demonstrates the potential of a genome-wide PRS to complement these efforts, enhancing risk stratification and identifying high-risk individuals at younger ages, who may benefit from earlier interventions.

We address this gap by presenting a PRS that signifies a 3.92-fold increase in lung cancer risk for individuals in the top 20% risk quantile, a marked improvement over the 2.09-fold increase

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

observed in a prior study by Wei et al. [47] for a cohort of Chinese never-smoking women. Further, we identified a significant interaction between the PRS and age, observing heterogeneity in effects of PRS across age categories. Similar to findings in a study on prostate cancer[48] , we observed higher associations between the PRS and LUAD among younger individuals compared to those diagnosed at a later age. While the pattern was not strictly linear, and competing risks of mortality may influence these results, our findings suggest that the relative impact of genetic susceptibility may decrease as non-genetic risk factors accumulate with age. This highlights the potential utility of PRS in younger populations, where genetic risk may play a more pronounced role in disease onset.

We estimated the lifetime and 10-year absolute risk of LUAD using the PRS, which can be used in risk stratification efforts and to identify those who are at high risk for the disease. Specifically, we found that the 10-year absolute risk for LUAD in a never-smoking 50-year-old woman, a critical age for initiating recommended annual lung cancer screening as per the United States Preventive Services Taskforce (USPSTF)[49], varied from 0.13% in the 5th percentile to 1.11% in the 95th percentile of the PRS distribution. Notably, women in the highest risk categories (>95th and 90-95th percentiles) reached the average 10-year absolute risk (0.42%) much earlier, at ages 41 and 42, respectively. This finding suggests that the PRS has the potential to identify younger women who may benefit from earlier risk-based interventions. While limited in sample size, we also found that the lifetime absolute risk of LUAD was even higher among women in the top percentile of the PRS and with first-degree family history of lung cancer, suggesting the two risk factors may act independently for their effect on LUAD risk. Although the absolute risk values are modest, the PRS's ability to stratify women into higher-risk categories at younger ages suggests potential utility in guiding early screening decisions. Early identification through PRS could be particularly valuable for never-smoking individuals, who are traditionally excluded from lung cancer screening programs.

The CT-SLEB method's improved performance can be attributed to its ability to utilize diverse genotyping arrays, thereby including population-specific variants. This aspect, along with the inclusion of EUR samples in the training set, broadens the genetic diversity and leverages larger sample sizes for improved model accuracy. Moreover, CT-SLEB excels in modeling genetic

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

correlations across ancestries, efficiently using both shared and unique genetic markers for refined disease risk estimations across populations. Lastly, we set up our study with a robust three-sample split design, dividing datasets into training, tuning, and validation. This ensures that PRS efficacy is validated independently, effectively reducing the overfitting risk.

Current PRS studies predominantly focus on evaluating relative and absolute risks[8,12,15,47,50], yet our research introduces a crucial forecast of sample sizes required for differing levels of PRS accuracy. Our projection analyses indicated that achieving an AUC of 0.70 would approach the maximized prediction potential of the PRS, given the estimated genetic variance explained by GWAS chip variants. To reach this level of accuracy, a future study would need to include 55,000 cases with a 1:10 case-control ratio. Notably, accessing large biobanks with publicly available controls could reduce the number of required cases[51].-These projections offer a strategic framework for planning and designing future genetic studies on lung cancer, establishing clear benchmarks for PRS model performance, which is essential for the scientific community in advancing genetic risk prediction.

Our study has several limitations. First, it focuses on the genetic susceptibility of LUAD without considering additional questionnaire-based risk factors. However, to be able to utilize data from the largest GWAS conducted to date, this work aimed primarily to identify the most effective PRS model within genetic contexts that can then be used in future efforts. Further, we plan to conduct integrative analyses that merge PRS with other risk indicators to refine LUAD risk predictions for never-smokers. Second, our projections are tailored to single-ancestry PRS models, not fully addressing the intricacies of multi-ancestry approaches, including diverse sample sizes and genetic correlations across populations. Third, our validation cohort, derived from the FLCCA and covering various EAS regions, relied on Taiwanese incidence data for absolute risk estimations. Consequently, our absolute risk findings may not extend universally to all EAS never-smokers, though our PRS performance assessment and sample size projections remain applicable to our population of interest. Lastly, the clinical application of our PRS findings, particularly in screening and risk counseling, is still in preliminary stages. Despite advancements in using PRS for smoking cessation trials[52], its implementation for never-smokers in clinical settings has been limited, presenting a significant area for future research.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

In summary, our study evaluates various PRS models to capture the genetic predisposition to LUAD among never-smoking EAS individuals. It extends beyond risk prediction by estimating both 10-year and cumulative lifetime absolute risks, and by projecting the sample sizes required for future GWAS to refine the predictive power of PRSs in future GWAS. Additionally, we quantify the phenotypic variance captured by PRSs across different sample sizes. Future studies are crucial to further improve these PRS models, aiming to enhance genetic risk predictions while integrating a wider array of risk factors. Such efforts will develop more accurate and comprehensive risk models for LUAD in never-smoking individuals across diverse populations.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Acknowledgements

This work utilized the computational resources of the high-performance computation Biowulf cluster at National Institutes of Health, USA (http://hpc.nih.gov).

Female Lung Cancer Consortium in Asia (NCI**):** This study was supported by a Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Science, Sports, Culture and Technology of Japan, a Grant-in- Aid for the Third Term Comprehensive 10-Year Strategy for Cancer Control from the Ministry Health, Labor and Welfare of Japan, by Health and Labor Sciences Research Grants for Research on Applying Health Technology from the Ministry of Health, Labor and Welfare of Japan, by the National Cancer Center Research and Development Fund, the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (grant No. 2011-0016106), a grant of the National Project for Personalized Genomic Medicine, Ministry for Health & Welfare, Republic of Korea (A111218-11-GM04), the Program for Changjiang Scholars and Innovative Research Team in University in China (IRT_14R40 to K.C.), the National Science & Technology Pillar Program (2011BAI09B00), MOE 111 Project (B13016), the National Natural Science Foundation of China (No. 30772531, and 81272618), Guangdong Provincial Key Laboratory of Lung Cancer Translational Medicine (No. 2012A061400006), Special Fund for Research in the Public Interest from the National Health and Family Planning Commission of PRC (No. 201402031), and the Ministry of Science and Technology, Taiwan (MOST 103-2325-B-400-023 & 104-2325-B-400-012). The Japan Lung Cancer Study (JLCS) was supported in part by the Practical Research for Innovative Cancer Control from Japan Agency for Medical Research and Development (15ck0106096h0002) and the Management Expenses Grants from the Government to the National Cancer Center (26-A-1) for Biobank. BioBank Japan was supported by the Ministry of Education, Culture, Sports, Sciences and Technology of the Japanese government. The Japan Public Health Center-based prospective Study (the JPHC Study) was supported by the National Cancer Center Research and Development Fund (23-A- 31[toku], 26-A-2, 29-A-4, 2020-J-4, and 2023-J-4) (since 2011) and a Grant-in-Aid for Cancer Research from the Ministry of Health, Labour and Welfare of Japan (from 1989 to 2010). The Taiwan GELAC Study (Genetic Epidemiological Study for Lung AdenoCarcinoma) was sup- ported by grants from the National Research Program on Genomic Medicine in Taiwan (DOH99-TD-G-111-028), the National Research Program for Biopharmaceuticals in Taiwan (MOHW 103-TDUPB-211-144003, MOST 103-2325-B-400-023)

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

and the Bioinformatics Core Facility for Translational Medicine and Biotechnology Development (MOST 104-2319-B-400-002). This work was also supported by the Jinan Science Research Project Foundation (201102051), the National Key Scientific and Technological Project (2011ZX09307-001-04), the National Natural Science Foundation of China (No.81272293), the State Key Program of National Natural Science of China (81230067), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF- 2014R1A2A2A05003665), Sookmyung Women's University Research Grants, Korea (1-1603-2048), Agency for Science, Technology and Research (A*STAR), Singapore and the US National Institute of Health Grant (1U19CA148127-01). The overall GWAS project was supported by the intramural program of the US National Institutes of Health/National Cancer Institute. The following is a list of grants by study center: SKLCS (Y.T.K.)—National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (2011-0016106). (J.C.) – This work was supported by a grant from the National R&D Program for Cancer Control, Ministry of Health &Welfare, Republic of Korea (grant no. 0720550-2). (J.S.S) – grant number is A010250. WLCS (T.W.)—National Key Basic Research and Development Program (2011CB503800). SLCS (B.Z.)—National Nature Science Foundation of China (81102194). Liaoning Provincial Department of Education (LS2010168). China Medical Board (00726). GDS (Y.L.W.)—Foundation of Guangdong Science and Technology Department (2006B60101010, 2007A032000002, 2011A030400010). Guangzhou Science and Information Technology Bureau (2011Y2-00014). Chinese Lung Cancer Research Foundation, National Natural Science Foundation of China (81101549). Natural Science Foundation of Guangdong Province (S2011010000792). TLCS (K.C., B.Q)—Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT), China (IRT1076). Tianjin Cancer Institute and Hospital. National Foundation for Cancer Research (US). FLCS (J.C.W., D.R., L.J.)—Ministry of Health (201002007). Ministry of Science and Technology (2011BAI09B00). National S&T Major Special Project (2011ZX09102-010-01). China National High-Tech Research and Development Program (2012AA02A517, 2012AA02A518). National Science Foundation of China (30890034). National Basic Research Program (2012CB944600). Scientific and Technological Support Plans from Jiangsu Province (BE2010715). NLCS (H.S.)—China National High-Tech Research and Development Program Grant (2009AA022705). Priority Academic Program Development of Jiangsu Higher Education Institution. National Key Basic Research Program

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Grant (2011CB503805). GEL-S (A.S.)—National Medical Research Council Singapore grant (NMRC/0897/2004, NMRC/1075/2006). (J.Liu)—Agency for Science, Technology and Research (A*STAR) of Singapore. GELAC (C.A.H.)—National Research Program on Genomic Medicine in Taiwan (DOH98-TDG-111-015). National Research Program for Biopharmaceuticals in Taiwan (DOH 100- TD-PB-111-TM013). National Science Council, Taiwan (NSC 100- 2319-B-400-001). YLCS (Q.L.)—Supported by the intramural pro- gram of U.S. National Institutes of Health, National Cancer Institute. SWHS (W.Z., W-H.C., N.R.)—The work was supported by a grant from the National Institutes of Health (R37 CA70867, UM1 CA182910) and the National Cancer Institute intramural research program, including NCI Intramural Research Program contract (N02 CP1101066). JLCS (K.M., T.K.)—Grants-in-Aid from the Ministry of Health, Labor, and Welfare for Research on Applying Health Technology and for the 3rd-term Comprehensive 10-year Strategy for Cancer Control; by the National Cancer Center Research and Development Fund; by Grant-in-Aid for Scientific Research on Priority Areas and on Innovative Area from the Ministry of Education, Science, Sports, Culture and — Technology of Japan. (W.P.)—NCI R01-CA121210. HKS (J.W.)— General Research Fund of Research Grant Council, Hong Kong (781511M). The Environment and Genetics in Lung Cancer Etiology (EAGLE), Prostate, Lung, Colon, Ovary Screening Trial (PLCO), and Alpha-Tocopherol, Beta-Carotene Cancer Prevention (ATBC) studies were supported by the Intramural Research Program of the National Institutes of Health, National Cancer Institute (NCI), Division of Cancer Epidemiology and Genetics. ATBC was also supported by U.S. Public Health Service contracts (N01-CN-45165, N01-RC-45035, and N01-RC-37004) from the NCI. PLCO was also supported by individual contracts from the NCI to the University of Colorado Denver (NO1-CN-25514), Georgetown University (NO1-CN-25522), the Pacific Health Research Institute (NO1-CN-25515), the Henry Ford Health System (NO1-CN-25512), the University of Minnesota, (NO1-CN- 25513), Washington University (NO1-CN-25516), the University of Pittsburgh (NO1-CN-25511), the University of Utah (NO1-CN- 25524), the Marshfield Clinic Research Foundation (NO1-CN- 25518), the University of Alabama at Birmingham (NO1-CN- 75022), Westat, Inc. (NO1-CN-25476), and the University of California, Los Angeles (NO1-CN-25404). The Carotene and Retinol Efficacy Trial (CARET) is funded by the National Cancer Institute, National Institutes of Health through grants U01-CA063673, UM1-CA167462, and U01-CA167462. The Cancer Prevention Study-II (CPS-II) Nutrition

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Cohort was supported by the American Cancer Society. The NIH Genes, Environment and Health Initiative (GEI) partly funded DNA extraction and statis- tical analyses (HG-06-033-NCI-01 and RO1HL091172-01), genotyping at the Johns Hopkins University Center for Inherited Disease Research. This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A2C4002236)

Female Lung Cancer Consortium in Asia (Tianjin): Tianjin Science and Technology Committee Foundation, 18YFZCSY00520.

Female Lung Cancer Consortium in Asia (Taiwan): The Ministry of Health and Welfare grants DOH97-TD-G-111-028 (ISC), DOH98-TD-G-111-017 (ISC), DOH99-TD-G-111-014 (ISC); DOH97-TD-G-111-026 (CAH), DOH98-TD-G-111-015 (CAH), DOH99-TD-G-111-028 (CAH); National Health Research Institutes grants NHRI-PH-110-GP-01, NHRI-PH-110-GP-03; and the Ministry of Science and Technology grants MOST108-2314-B-400-038(CAH), MOST109-2740-B-400-002(CAH), MOST 111-2740-B-400-002 (CAH), MOST 111-2314-B-400-020 (CAH).

The GWAS of lung cancer in European never smokers was supported by NIH R01 CA149462 (OYG).

OncoArray study in Europeans: The OncoArray data and analysis from INTEGRAL-ILCCO were supported by NIH U19 CA203654, and U19 CA148127. The data harmonization for ILCCO was supported by Canadian Institute for Health Research (CIHR) Canada Research Chair to R.J.H, and CIHR FDN 167273).

European never-smoking lung cancer study: CIA is a Research Scholar of the Cancer Prevention Institute of Texas (CPRIT) and supported by CPRIT grant RR170048.

Taiwan eQTL study: This study was supported by the Ministry of Health and Welfare grants DOH97-TD-G-111-028 (ISC), DOH98-TD-G-111-017 (ISC), DOH99-TD-G-111-014 (ISC);

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

DOH97-TD-G-111-026 (CAH), DOH98-TD-G-111-015 (CAH), DOH99-TD-G-111-028 (CAH); National Health Research Institutes grants NHRI-PH-110-GP-01, NHRI-PH-110-GP-03; and the Ministry of Science and Technology grants MOST108-2314-B-400-038(CAH), MOST109-2740-B-400-002(CAH).

B.B., K.W., J.C., J.S. N.R., Q.L. and H.Z. are supported by NIH intramural Research Program. N.C. is supported by NIH grant 1R01HG010480. P.Y. is supported by Mayo Clinic Foundation Research Funds, NIH-CA77118 and CA80127. G.L. is supported is supported by the Alan Brown Chair and Lusi Wong Fund of the Princess Margaret Cancer Foundation. D.C.C is supported by U01CA209414. O.Y.G. is supported by NIH R01 CA231141.

Author contributions

B.B., N.R., Q.L. and H.Z. organized and designed the study. B.B., K.W., H.Z. carried out the statistical analyses. B.B., K.W., Q.L. N.R. and H.Z. wrote the first draft. Q.L., M.T.L., B.A.B., W.H., N.E.C., BT.J., M.Song, H.P., D.A., C.C.C., L.B., M.Y., A.H., B.H., J.Liu, B.Zhu, S.I.B., C.K., K.Wyatt, S.A.L., A.Chao, J.F.F.J., S.J.C., N.R., Z.Wang, C..L., J.C., C.W., W.T., D.Lin, SJ.A., XC.Z., J.S., YL.W., M.P.W., L.P.C., J.C.M.H., V.H.F.L., Z.H., K.M., J.Y..P., Jia.Liu, HS.J., J.E..C., Y.Y.C., H.N.K., MH.S., SS.K., YC.K., IJ.O., S.W.S., HI.Y., Y.T.K., YC.H., J.H.K., Y.H.K., J.S..S., Y.J..J., K.H.P., C.H.K., J.S.K., I.K.P., B..S., Jie.L, Z.W., S.C., J.Y., JC.W., Y.Y., YB.X., YT.G., D.L., J.Y.Y.W., H.C., L..J., J.Z., G.J., K..F., Z.Y., B.Z., W.W., P.G., Q.H., X.L., Y.R., A.S., Y.L., Y.C., WY.L., W.Z., XO.S., Q.C., G.Y., B.Q., T.W., H.G., L..L., P.X., F.W., G.W., J.X., J.L., R.CH.V., B.B., H.D.III.H., J.Wang, A.D.L..S., J.KC.C., V.L.S., K.C., H.Z., H.D., C.A.H., TY.C., LH.C..., IS.C., CY.C., S.S.J ., CH.C., GC.C., CF.H., YH.T., WC.W., KY.C., MS.H., WC.S., YM.C., CL.W., KC.C., CJ.Y., HH.H., FY.T.., HC.L., CJ.C., PC.Y., K.Shiraishi, T.K., H.K., S.M., H.H., K.Goto, Y.Ohe, S.W., Y.Yatabe, M.T., R.Hamamoto, A.Takahashi, Y.Momozawa, M.Kubo, Y.K., Y.D., Y.Miyagi, H.N., T.Y., N.S., M.I., M.H., Y.N., K.Takeuchi, K.W., K.Matsuda, Y.Murakami, K.S., K.T., Y.O., M.S., H.Suzuki, A.G., Y.M., T.H., M.K., K.O., H.S., J.D., H.M., M.Z., R.J.H., S.L., A.T., C.C., S.E.B., M.Johansson, A.R., H.Bö., HE.W., D.C., G.R., S.A., P.B., J.MK., J.K.F., S.S.S., L.L.M., O.M.., H.Bö., G.L., A.A., L.A.K., S.ZN., K.G., M.J., A.C., JM.Y., P.L., M.B.S., M.C.A., C.I.A., A.G.S., R.H., M.R.S., O.Y.G., I.P.G., X.W., P.Y. conducted epidemiology studies and

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

contributed samples to GWAS and/or conducted initial genotyping. All authors reviewed and approved the final review of the manuscript.

Competing interests

The authors declare no competing interests.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table 1. Prediction performance of different methods for generating polygenic risk scores for lung cancer in never-smoking East Asian populations.

Model type	Number of SNPs	OR (95% CI)[1]	AUC (95% CI)[2]
Single-ancestry method[3]			
PRS-25[4]	25 (24 available)	1.53 (1.44, 1.63)	0.621 (0.612, 0.637)
PRS-CT	8	1.37 (1.29, 1.46)	0.591 (0.585, 0.609)
LDpred2 PRS	942,591	1.62 (1.52, 1.73))	0.629 (0.618, 0.641)
PRS-EUR 128[5]	128 (109 available)	0.97 (0.91, 1.03)	0.489 (0.477, 0.501)
Multi-ancestry method[6]			
LDpred2 PRS + PRS-EUR 128	942,700	1.62 (1.52, 1.72)	0.629 (0.617, 0.640)
PRS-CSx	969,720	1.67 (1.57, 1.78)	0.637 (0.625, 0.647)
CT-SLEB	2,127,229	1.71 (1.61, 1.82)	0.640 (0.629, 0.653)

Abbreviations: SNP, single nucleotide polymorphism; OR, odds ratio; CI, confidence interval; AUC, area under the curve; PC, principal component; CT, clumping and thresholding

[1]ORs are provided per one unit standard deviation of PRS.

[2]Adjusted for age (continuous) and first 10 principal components.

[3]Sample size for single-ancestry PRS methods included 3,564 cases and 16,238 controls of East Asian (EAS) ancestry for training, 2,219 cases and 2,272 controls of EAS ancestry for tuning and 2,219 cases and 2,272 controls of EAS ancestry for validation.

[4]SNPs that reached genome-wide significance in Shi et al., 2023.

[5]SNPs obtained from genome-wide association study conducted in individuals with European ancestry (Hung et al., 2021, Cancer Res.).

[6]Sample size for multi-ancestry PRS methods included 3,564 cases and 16,238 controls of East Asian (EAS) ancestry, as well as 2,058 cases and 5,575 controls of European ancestry for training, 2,219 cases and 2,272 controls of EAS ancestry for tuning and 2,219 cases and 2,272 controls of EAS ancestry for validation.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table 2. Genetic variance in East Asian lung cancer among never smokers explained by LDpred2 PRS

Sample size[1]	Genetic variance of PRS[2]	Proportion of all-GWAS variants genetic variance explained by PRS[3]	Proportion of Familial risk explained by PRS[4]
Current	0.201	26.6%	16.5%
35,000 cases	0.437	57.9%	35.8%
55,000 cases	0.540	71.5%	44.3%

[1] *Sample Size*: The current sample comprises 3,564 cases and 16,238 controls. Future projections assume a 1:10 case-control ratio for sample sizes of 35,000 and 55,000 cases. Genetic variance projections for the LDpred2 PRS are based on the GENESIS method, originally designed for the CT method (refer to Nat. Genet. 50, 1318-1326 (2018)), extended to include LDpred2 by modeling the variance ratio between LDpred2 and CT (**Supplementary Note**).

[2] *Genetic Variance of PRS*: This corresponds to the heritability on the frailty scale, assuming a polygenic log-additive model underpins this relationship. It quantifies the proportion of the phenotype variation that can be attributed to genetic factors in the context of PRS.

[3] *Proportion of Genetic Variance from All-GWAS Variants Explained by PRS*: This represents the variance of all genome-wide imputable variants as established through LD-score regression (refer to Nat. Genet. 47, 291-5 (2015) and Nat. Genet. 47, 1236-41 (2015)). On the frailty scale, the genetic variance of all GWAS variants is calculated as $\sigma^2_{GWAS} = Var(\sum_{m=1}^{M} \beta_m G_m)$, where G_m is the standardized genotype for the mth SNP, β_m is the true log odds ratio for the mth SNP and M is the total number of causal SNPs among the GWAS variants. For East Asian never smokers, the estimated genetic variance of all GWAS variants is 0.755.

[4] Proportion of Familial Risk Explained by PRS: This calculates the familial risk in terms of genetic variance using the formula $\lambda_s^2 = \exp(\sigma^2)$, where λ_s is the familial risk when a first-order sibling has the disease, and σ^2 is the genetic variance on frailty-scale. Further details of this calculation can be found in Nat. Genet. 31, 33-36 (2002). For lung cancer in East Asian never-smokers, the familial risk is a 1.84-fold increase, and the genetic variance of all GWAS variants, as estimated through LD-score regression, explains 61.9% of this increased familial relative risk.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

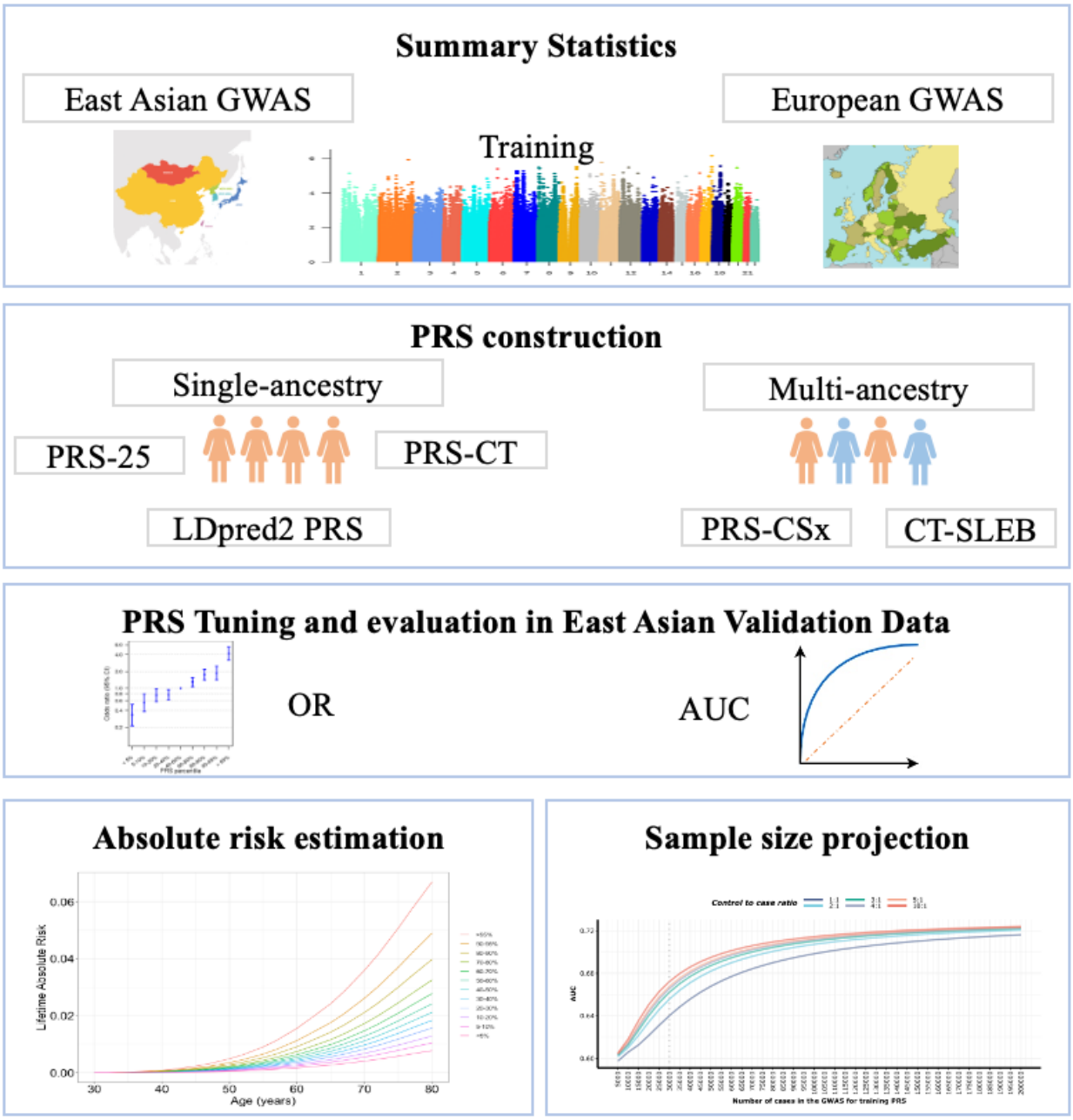


Figure 1. Overview of data structure, polygenic risk score (PRS) development, validation and application. Summary statistics from East Asian (EAS) genome-wide association studies (GWAS) were used to develop single-ancestry PRS using methods such as a simple PRS constructed using 25 SNPs that have previously reached genome-wide significance (i.e., $P<5x10^{-8}$) (PRS-25), a PRS using the clumping and thresholding (CT) method (PRS-CT) incorporating 8 SNPs, and a PRS using a genome-wide Bayesian-based approach, LDpred2 (LDpred2 PRS)

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

incorporating close to a million SNPs. For the multi-ancestry PRS development, we also used summary statistics from European (EUR) GWAS, applying the PRS-CSx method that leveraged genome-wide association summary statistics for close to a million SNPs with a Bayesian continuous shrinkage prior to model SNP effect sizes across populations, as well as CT-SLEB method, which enhances the standard CT methods with a two-dimensional approach to select SNPS for EAS PRS construction by incorporating over 2 million SNPs. Tuning and validation of each PRS was conducted in an independent EAS individual-level data. Relative risk per PRS quantile was calculated as an odds ratio (OR) with the middle quantile (40th to 60th percentile) set as the reference, and the area under the receiver operating curve (AUC) was estimated for each PRS. CT-SLEB PRS was used to estimate 10-year and lifetime cumulative absolute risk, and PRS-CT and PRS-LDpred2 were used for sample size projection.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

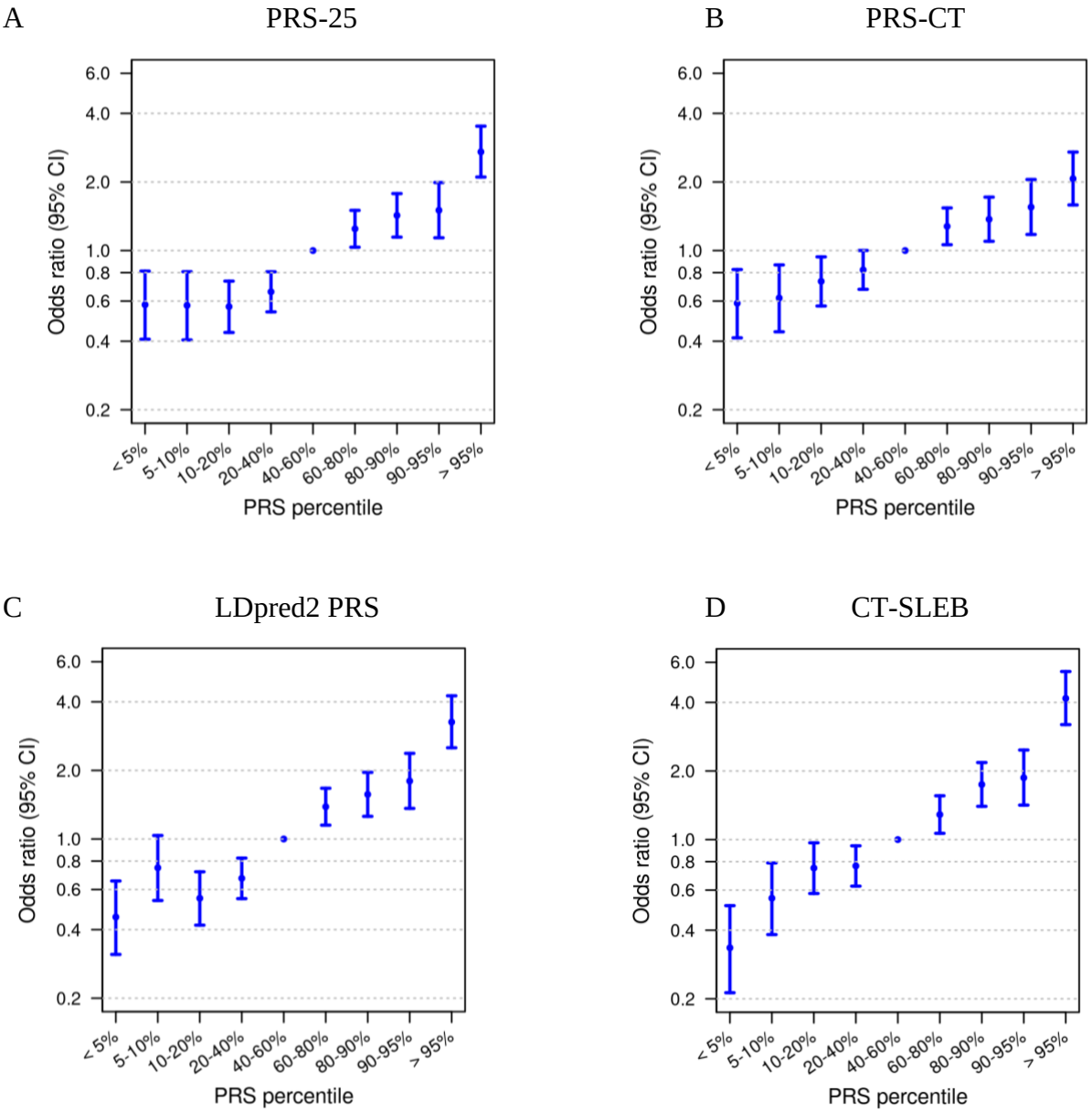
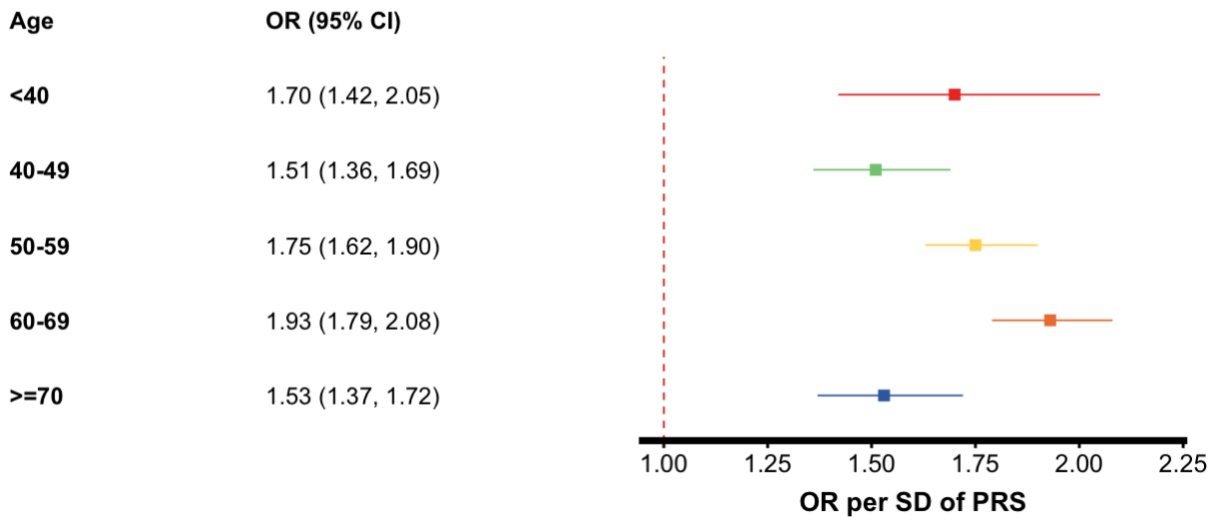


Figure 2. Relative risk estimated for quantiles of each polygenic risk score (PRS) and lung adenocarcinoma in the validation dataset of women with East Asian ancestry, treating the 40th to 60th percentile as the references. Odds ratios of PRS per standard deviation (SD) and 95% confidence intervals are shown for (A) the single-ancestry 25 SNP polygenic risk score, PRS-25 (A), (B) Clumping and thresholding method, PRS-CT, (C) Bayesian-based genome-wide approach, LDpred2 PRS, and (D) multi-ancestry approach, CT-SLEB.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

A



B

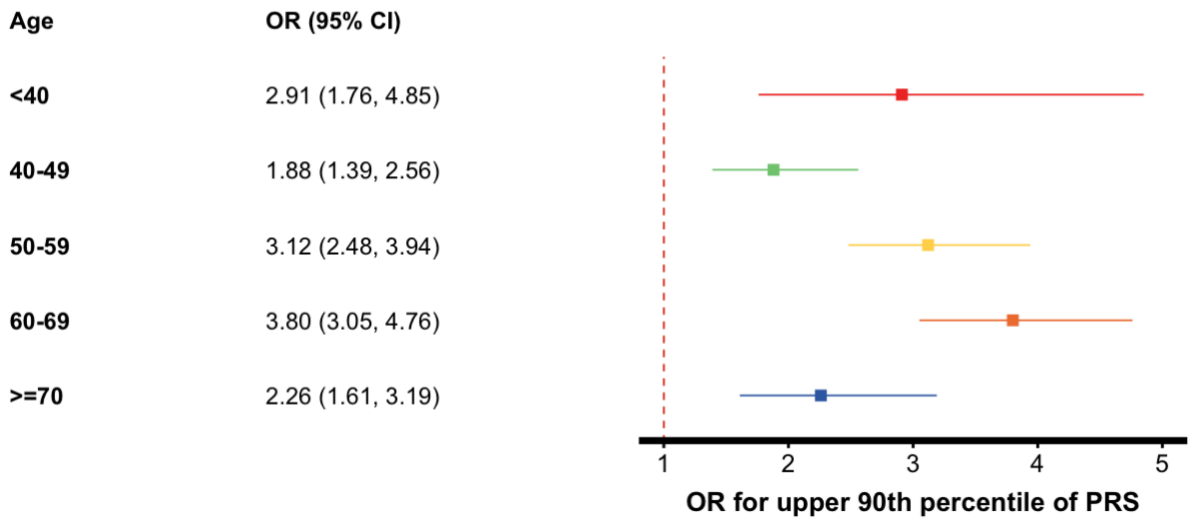


Figure 3. Association between polygenic risk score (PRS) and lung adenocarcinoma by age groups. (A) Odds ratios (ORs) per standard deviation (SD) of the PRS and 95% confidence intervals, and (B) ORs for individuals in the upper 90th percentile of the PRS.

A B

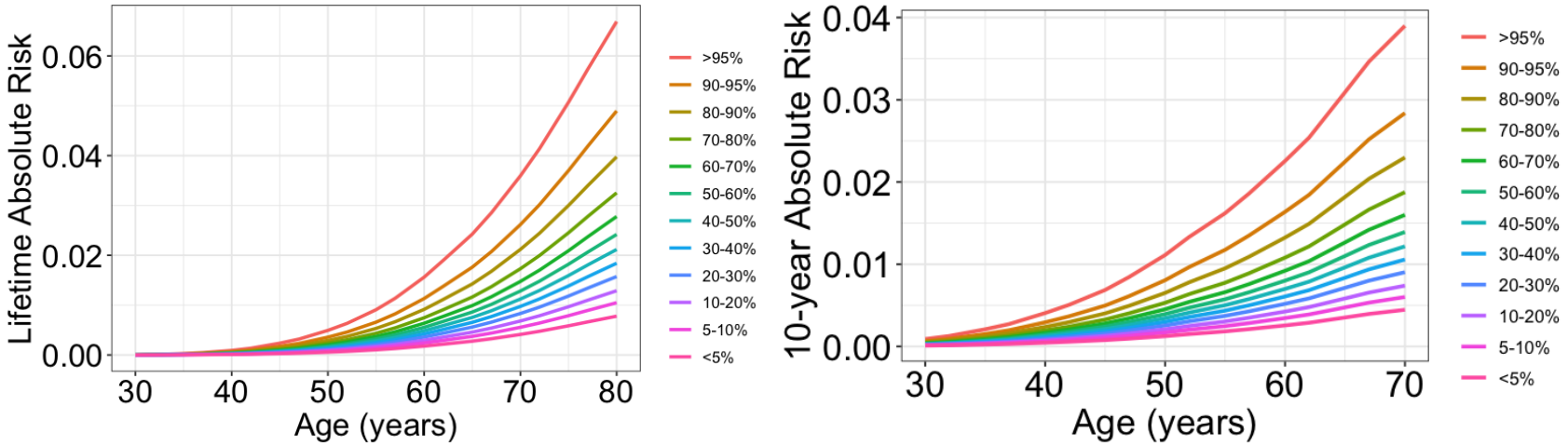


Figure 4. Lifetime cumulative and 10-year absolute risk of developing lung adenocarcinoma. (A) Lifetime (age 30-80) cumulative risk and (B) 10-year absolute risk of developing lung adenocarcinoma in never-smoking women in East Asia by percentiles of the CT-SLEB polygenic risk score (PRS). Absolute risks were calculated using the iCARE package[38], based on Taiwan's age-specific incidence and mortality data, and the PRS relative risks, as described in the Methods section.

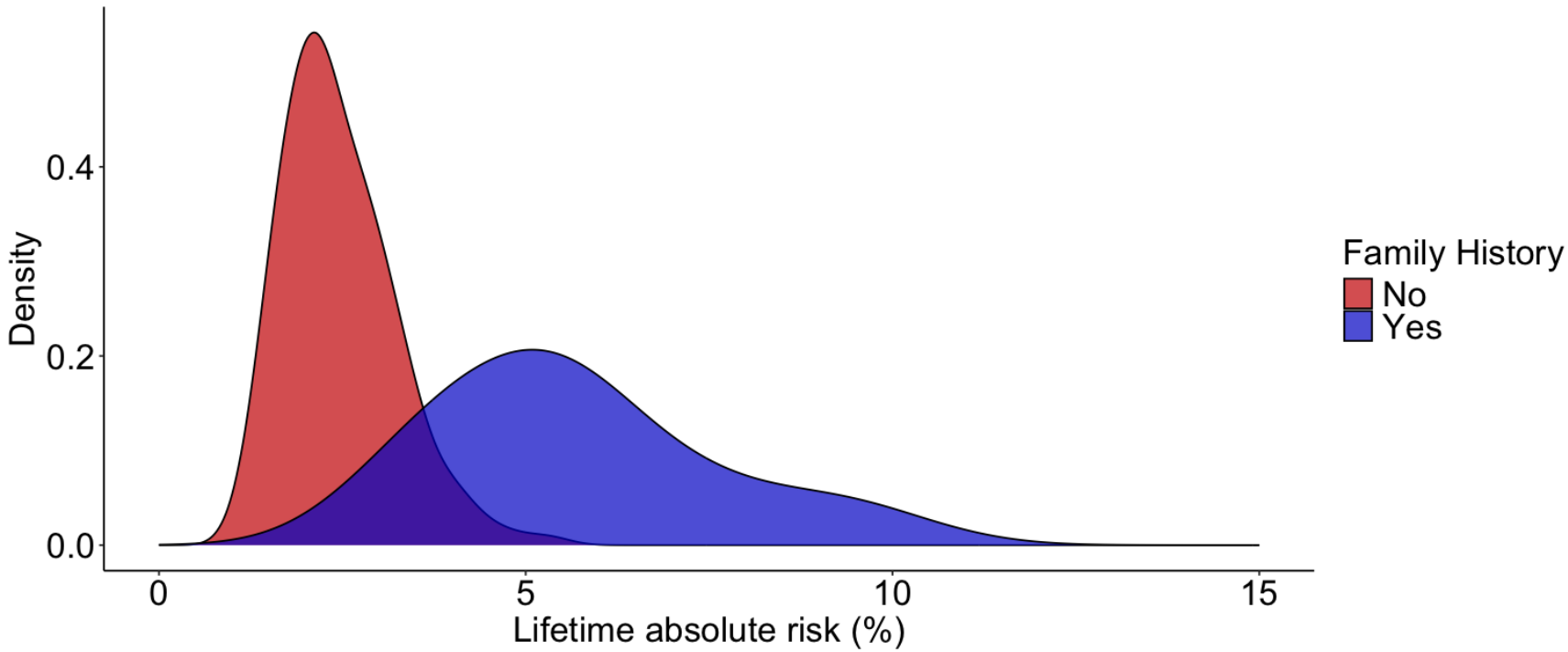
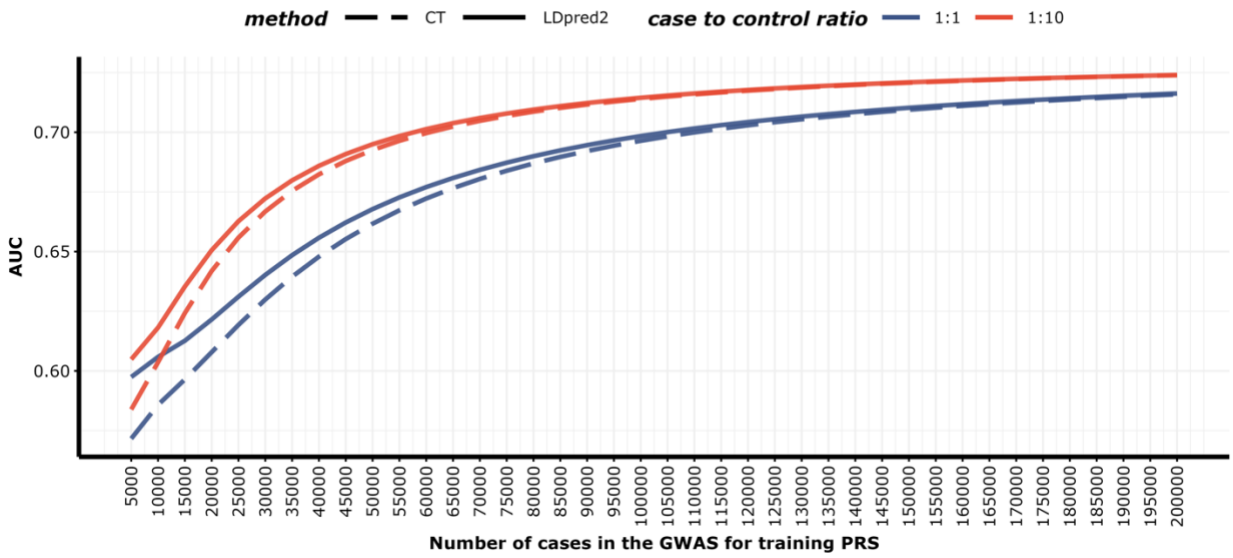


Figure 5. Lifetime cumulative absolute risk of developing lung adenocarcinoma by family history. Lifetime (age 30-80) absolute risk of developing lung adenocarcinoma in never-smoking East Asian women stratified by first-degree family history of lung cancer. Absolute risks were calculated using iCARE package[38], based on Taiwan's age specific incidence and mortality data, and the relative risks of the PRS and family history, as described in the Methods section.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

A



B

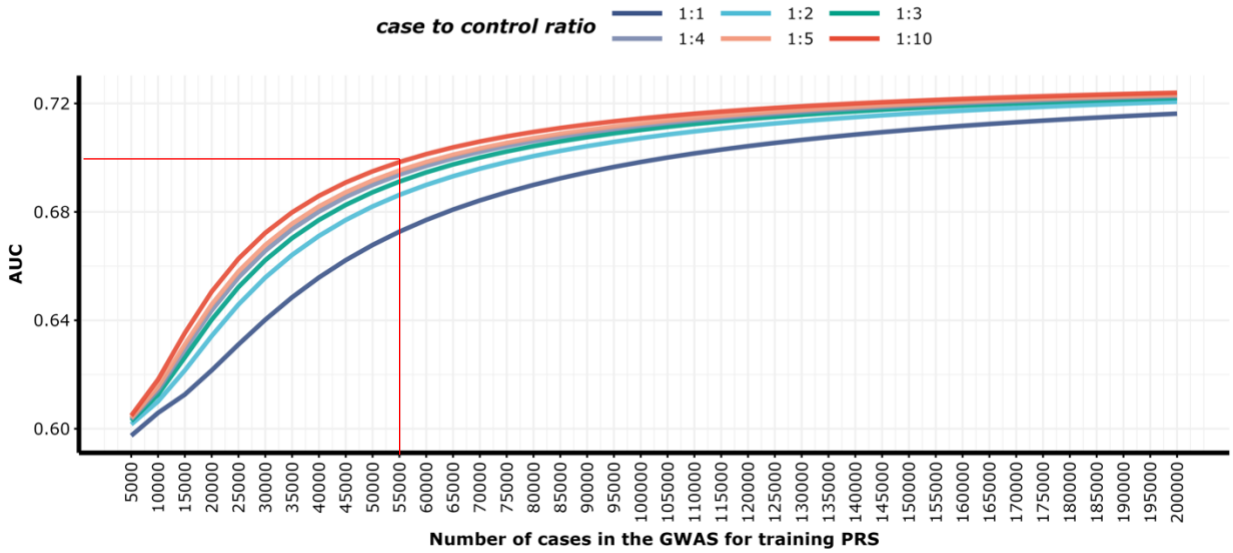


Figure 6. Projected area under the receiver operating characteristic curve (AUC) of polygenic risk scores (PRS) built using genome-wide association studies (GWAS) with varying sample sizes for lung adenocarcinoma in never-smoking East Asian women. (A)

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

AUC values for PRS-CT and LDpred2 PRS with case-to-control ratios of 1:1 and 1:10. (B) AUC values for LDpred2 PRS across case-to-control ratios of 1:1, 1:2, 1:3, 1:4, 1:5, and 1:10.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

References

1. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **68**, 394–424 (2018).
2. Sun, S., Schiller, J. H. & Gazdar, A. F. Lung cancer in never smokers--a different disease. *Nat. Rev. Cancer* **7**, 778–790 (2007).
3. Cheng, T. Y. D. *et al.* The International Epidemiology of Lung Cancer: Latest Trends, Disparities, and Tumor Characteristics. *J. Thorac. Oncol.* **11**, 1653–1671 (2016).
4. Byun, J. *et al.* 'Cross-ancestry genome-wide meta-analysis of 61,047 cases and 947,237 controls identifies new susceptibility loci contributing to lung cancer'. *Nat Genet* **54**, 1167 (2022).
5. Landi, M. T. *et al.* Tracing lung cancer risk factors through mutational signatures in never-smokers. *Am. J. Epidemiol.* **190**, 962–976 (2021).
6. Corrales, L. *et al.* Lung cancer in never smokers: The role of different risk factors other than tobacco smoking. *Crit. Rev. Onco. Hematol.* **148**, (2020).
7. Sisti, J. & Boffetta, P. What proportion of lung cancer in never-smokers can be attributed to known risk factors? *Int. J. Cancer* **131**, 265–275 (2012).
8. Shi, J. *et al.* Genome-wide association study of lung adenocarcinoma in East Asia and comparison with a European population. *Nature Commun.* **14**, (2023).
9. Dai, J. *et al.* Identification of risk loci and a polygenic risk score for lung cancer: a large-scale prospective cohort study in Chinese populations. *Lanet. Respir. Med.* **7**, 881–891 (2019).
10. Seow, W. J. *et al.* 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**, 454–465 (2017).
11. Lan, Q. *et al.* Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. *Nat. Genet.* **44**, 1330–1335 (2012).
12. Li, Y. *et al.* Genetic variants and risk of lung cancer in never smokers: a genome-wide association study. *Lanet Oncol.* **11**, 321–330 (2010).
13. Hung, R. J. *et al.* Lung cancer risk in never-smokers of European descent is associated with genetic variation in the 5p15.33 TERT-CLPTM1Ll region. *J. Thorac Oncol* . **14**, 1360–1369 (2019).
14. Lebrett, M. B. *et al.* Validation of lung cancer polygenic risk scores in a high-risk case-control cohort. *Genet. Med.* **25**, (2023).
15. Hung, R. J. *et al.* Assessing Lung Cancer Absolute Risk Trajectory Based on a Polygenic Risk Model. *Cancer Res.* **81**, 1607–1615 (2021).
16. Linder, J. E. *et al.* Returning integrated genomic risk and clinical recommendations: The eMERGE study. *Genet. Med.* **25**, (2023).
17. Hao, L. *et al.* Development of a clinical polygenic risk score assay and reporting workflow. *Nat. Med.* **28**, 1006–1013 (2022).
18. Shieh, Y. *et al.* Breast Cancer Screening in the Precision Medicine Era: Risk-Based Screening in a Population-Based Trial. *J. Natl. Cancer Inst.* **109**, (2017).
19. Kachuri, L. *et al.* Principles and methods for transferring polygenic risk scores across global populations. *Nat. Rev. Genet.* **25**, 8–25 (2024).
20. Jin, J. *et al.* MUSSEL: Enhanced Bayesian polygenic risk prediction leveraging information across multiple ancestry groups. *Cell Genom.* **4**, (2024).

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

21. Zhang, J. *et al.* An ensemble penalized regression method for multi-ancestry polygenic risk prediction. *Nat. Commun.* **15**, 1–14 (2024).
22. Hou, K. *et al.* Admix-kit: an integrated toolkit and pipeline for genetic analyses of admixed populations. *Bioinformatics* **40**, (2024).
23. Hoggart, C. J. *et al.* BridgePRS leverages shared genetic effects across ancestries to increase polygenic risk score portability. *Nat. Genet.* **56**, 180–186 (2024).
24. Zhang, H. *et al.* A new method for multiancestry polygenic prediction improves performance across diverse populations. *Nat. Genet.* **55**, 1757–1768 (2023).
25. Zhou, G., Chen, T. & Zhao, H. SDPRX: A statistical method for cross-population prediction of complex traits. *Am. J. Hum. Genet* . **110**, 13–22 (2023).
26. Ruan, Y. *et al.* Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet.* **54**, 573–580 (2022).
27. Shi, J. *et al.* Genome-wide association study of lung adenocarcinoma in East Asia and comparison with a European population. *Nature Communications 2023 14:1* **14**, 1–17 (2023).
28. Hu, Z. *et al.* A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. *Nat. Genet.* **43**, 792–796 (2011).
29. Wang, L. *et al.* Genetically determined height was associated with lung cancer risk in East Asian population. *Cancer Med.* **7**, 3445–3452 (2018).
30. Dong, J. *et al.* Association analyses identify multiple new lung cancer susceptibility loci and their interactions with smoking in the Chinese population. *Nat. Genet.* **44**, 895–898 (2012).
31. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
32. Chang, C. C. *et al.* Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
33. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424–5431 (2021).
34. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
35. Hung, R. J. *et al.* Assessing Lung Cancer Absolute Risk Trajectory based on a Polygenic Risk Model. *Cancer Res* **81**, 1607 (2021).
36. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
37. Foucher, Y. *et al.* RISCA: Causal Inference and Prediction in Cohort-Based Analyses. *R package version 1.01* Preprint at https://cran.r-project.org/package=RISCA (2020).
38. Pal Choudhury, P. *et al.* iCARE: An R package to build, validate and apply absolute risk models. *PLoS One* **15**, e0228198 (2020).
39. Chien, L. H. *et al.* Predicting lung cancer occurrence in never-smoking females in Asia: TNSF-SQ, a prediction model. *Cancer Epidemiology Biomarkers and Prevention* **29**, 452–459 (2020).
40. Zhang, Y., Qi, G., Park, J. H. & Chatterjee, N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat. Genet.* **50**, 1318–1326 (2018).

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

41. Choudhury, P. P. *et al.* iCARE: An R package to build, validate and apply absolute risk models. *PLoS One* **15**, (2020).
42. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
43. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, (2019).
44. Kullo, I. J. *et al.* Polygenic scores in biomedical research. *Nat. Rev. Genet.* **23**, 524–532 (2022).
45. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
46. Chang, G. C. *et al.* Low-dose CT screening among never-smokers with or without a family history of lung cancer in Taiwan: a prospective cohort study. *Lancet Respir Med* **12**, 141–152 (2024).
47. Wei, X. *et al.* Development and evaluation of a polygenic risk score for lung cancer in never-smoking women: A large-scale prospective Chinese cohort study. *Int. J. Cancer* **154**, 807–815 (2024).
48. Schaid, D. J., Sinnwell, J. P., Batzler, A. & McDonnell, S. K. Polygenic risk for prostate cancer: Decreasing relative risk with age but little impact on absolute risk. *The American Journal of Human Genetics* **109**, 900–908 (2022).
49. Krist, A. H. *et al.* Screening for Lung Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA - Journal of the American Medical Association* **325**, 962–970 (2021).
50. Blechter, B. *et al.* Polygenic risk score, environmental tobacco smoke, and risk of lung adenocarcinoma in never-smoking women in Taiwan. *JAMA Netw. Open* **6**, E2339254 (2023).
51. Katki, H. A. *et al.* Increase in power by obtaining 10 or more controls per case when type-1 error is small in large-scale association studies. *BMC Med. Res. Methodol.* **23**, (2023).
52. Chen, T. *et al.* Genomic Insights for Personalized Care: Motivating At-Risk Individuals Toward Evidence-Based Health Practices. *medRxiv* 2024.03.19.24304556 (2024) doi:10.1101/2024.03.19.24304556.

Data Availability

All data supporting the findings described in this paper are available in the paper and in the Supplementary Information and from the corresponding author or as otherwise indicated upon request. The individual genotype data for the FLCCA data are in dbGaP phs000716.v1.p1 (Genome-Wide Association Study of Lung Cancer Susceptibility in Never-Smoking Women in Asia). For the NCC and ACC studies, please contact Kouya Shiraishi at kshirais@ncc.go.jp or Takashi Kohno at tkkohno@ncc.go.jp for summary statistics. The GWAS data for the European populations contributing to this study are available at dbGap under accession phs000877.v1.p1

medRxiv preprint doi: https://doi.org/10.1101/2024.06.26.24309127; this version posted November 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

(Transdisciplinary Research Into Cancer of the Lung (TRICL), https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000876.v2.p1), phs001273.v3.p2 (Oncoarray Consortium, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001273.v3.p2). To gain access to all data in dbGaP cited in this paper, please apply for dbGaP Authorized Access. GWAS data for the European populations contributing to this study are available under accession phs000877.v1.p1 (ILLCO), phs001273.v1.p1 (OncoArray).

Code Availability

Methods implemented and their corresponding repositories include: SCT and LDpred2 at https://github.com/privefl/bigsnpr, PRS-CSx at https://github.com/getian107/PRScsx. CT-SLEB at https://github.com/andrewhaoyu/CTSLEB, and GENESIS at https://github.com/yandorazhang/GENESIS.

PLINK: https://www.cog-genomics.org/plink/1.9.

Most of our statistical analyses were performed using the following R packages: ggplot2 v.3.3.3, dplyr v.1.0.4, data.table v.1.13.6, iCARE v.1.30.0

https://github.com/KevinWFred/PRS_EASLC