Genomic Insights for Personalized Care: Motivating At-Risk Individuals Toward Evidence-Based Health Practices

Tony Chen[1,*], Giang Pham[2], Louis Fox[2], Jingning Zhang[3], Jinyoung Byun[4], Younghun Han[4], Gretchen R.B. Saunders[5], Dajiang Liu[6], Michael J. Bray[7,8], Alex T. Ramsey[2], James McKay[9], Laura Bierut[2], Christopher I. Amos[4,10], Rayjean J. Hung[11], Xihong Lin[1,12], Haoyu Zhang[13,†,*], Li-Shiun Chen[2,†,*]

[1] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, USA
[2] Department of Psychiatry, Washington University School of Medicine, St. Louis, USA
[3] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA.
[4] Department of Medicine, Section of Epidemiology and Population Science, Institute for Clinical and Translational Research, Houston, TX, USA
[5] Department of Psychology, University of Minnesota, Minneapolis, MN, USA
[6] Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA, USA
[7] Department of Genetic Counseling, Bay Path University, Longmeadow, MA, USA
[8] ThinkGenetics, Inc, USA
[9] International Agency for Research on Cancer, World Health Organization, Lyon, France
[10] Dan L Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX, USA
[11] Prosserman Centre for Population Health Research, Lunenfeld-Tanenbaum Research Institute, Sinai Health, and University of Toronto, Toronto, Canada
[12] Department of Statistics, Harvard University, Cambridge, USA
[13] Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA

*Correspondence to: Tony Chen (tonychen@g.harvard.edu), Haoyu Zhang (haoyu.zhang2@nih.gov) and Li-Shiun Chen (li-shiun@wustl.edu).
†These authors jointly supervised this work: Haoyu Zhang and Li-Shiun Chen.

Abstract

Lung cancer and tobacco use pose significant global health challenges and require a comprehensive translational roadmap for improved prevention strategies. We propose the GREAT care paradigm (Genomic Informed Care for Motivating High Risk Individuals Eligible for Evidence-based Prevention), which employs polygenic risk scores (PRSs) to stratify disease risk and personalize interventions, such as lung cancer screening and tobacco treatment. We developed PRSs using large-scale multi-ancestry genome-wide association studies and adjusted for genetic ancestry for standardized risk stratification across diverse populations. We applied our PRSs to over 340,000 individuals of diverse ethnic background and found significant odds ratios for lung cancer and difficulty quitting smoking. These findings enable the evaluation of PRS-based interventions in ongoing trials aimed at motivating health behavior changes in high-risk patients. This pioneering approach enhances primary care with genomic insights, promising improved outcomes in cancer prevention and tobacco treatment, and is currently under assessment in clinical trials.

Introduction

The worldwide burden of lung cancer and tobacco smoking presents major challenges to global health[1]. Evidence-based practices to reduce their risk such as cancer screening and tobacco treatment (e.g. smoking cessation medication) have long existed but are infrequently used in most primary care practices. Communication of the precision risk of lung cancer and precision benefit of smoking cessation is a promising but untested strategy to promote health behavior changes to reduce cancer risk. To address this gap, polygenic risk scores (PRSs) emerge as a valuable approach to assess disease susceptibility among populations and pinpoint individuals at higher risk[2–4]. PRSs can be derived from large-scale genome-wide association studies (GWAS) to estimate individual disease risk and have shown promise in predicting health outcomes and promoting preventive healthcare[5–9]. Despite their potential, PRSs' implementation in primary care is limited, especially in diverse populations. Implementing a PRS-based precision intervention is crucial in order to address the multifaceted needs of different communities and individuals. Harnessing PRSs effectively can make significant progress in mitigating lung cancer's public health impact.

Ongoing studies like eMERGE[10], GenoVA[11], and WISDOM[12] are leading the implementation of PRS into genetic risk reports (**Table 1**). They aim to personalize medical reports and understand the impact of PRS on screening, diagnostic procedures, and patient behavior. Notably, a gap persists as these initiatives have not yet formulated a PRS specifically for lung cancer. A likely reason is that the global burden of lung cancer is primarily driven by tobacco smoking rather than genetics[13,14]. However, accounting for the genetic basis of lung cancer may provide patients and clinicians with additional actionable information. The unique value proposition of a lung cancer-specific PRS lies in leveraging established and clear guideline-based prevention strategies, including smoking cessation treatment and lung cancer screening[15]. By incorporating PRSs for lung cancer and difficulty quitting smoking without treatment, there is an opportunity to revitalize and enhance these often under-utilized prevention practices.

We introduce the Genomic Informed Care for Motivating High Risk Individuals Eligible for Evidence-based Prevention (GREAT) framework as a novel approach to incorporate PRS-enabled interventions in clinical settings (**Figure 1**). The core of GREAT

is the use of PRSs that offer precise risk estimates for lung cancer and difficulty quitting. By providing patients with personalized risk information, we aim to activate behavior change mechanisms that promote preventive actions. The primary targets of this intervention are high-risk individuals eligible for evidence-based prevention practices, such as lung cancer screening and smoking cessation. By integrating precision risk information with the benefits of timely interventions, GREAT empowers patients to make informed decisions about their health and motivates them to take proactive steps towards prevention. Ultimately, our objective is not only to motivate, but to significantly reduce lung cancer morbidity and mortality through this innovative care paradigm.

Effectively translating PRSs into clinical practice requires a comprehensive and pragmatic translational roadmap for equitable and effective implementation (**Figure 1**). First, to address ancestry diversity, we take a two-step approach of (1) constructing PRS based on large-scale multi-ancestry GWAS, and (2) standardizing PRS distributions across the continuum of genetic ancestry[16,17] by leveraging reference data from the 1000 Genomes Project Phase 3 (1000G)[18]. Given the variations in allele frequencies (AF) and linkage disequilibrium (LD) across ancestries, this step is critical to ensure accurate risk stratification based on PRS distributions for all patients[10,19]. Second, to document accuracy and transportability of the PRSs to diverse populations, we perform large-scale validation using data from individuals of diverse self-reported ethnicities in the UK Biobank (UKBB) and Genetic Informed Smoking Cessation (GISC) trial. Third, we translate risk into actionable categories by setting appropriate thresholds for the PRS. The alignment of these thresholds with clinical significance involves many considerations for meaningful risk stratification. Fourth, we propose clear and patient-friendly communication strategies, including visual aids and educational materials, to facilitate understanding and meaningful interactions between patients and healthcare providers. Effectively communicating both the risk and precision of the PRS results is challenging but essential to empower patients to make informed decisions about their health. Moreover, it is crucial to consider patient perceived risk, perceived benefit, and personal relevance when discussing PRS results with patients. Patients' understanding and interpretation of PRS may vary, leading to differing levels of engagement in preventive

actions. Hence, comprehensive patient education programs can enhance awareness and knowledge about PRS, its implications, and available preventive measures.

In this paper, we introduce the design for two cluster randomized clinical trials (RCT): (1) PRECISE, which evaluates the effectiveness of a multilevel intervention, *RiskProfile*, on increasing lung cancer screening and tobacco treatment utilization in primary care (NIDA Grant 5R01CA268030-02); and (2) MOTIVATE, which evaluates the effect of PrecisionTx, a multilevel intervention to promote precision tobacco treatment in primary care (NIDA Grant 5R01DA056050-02). Through the innovative use of PRS, our aim is to motivate lung cancer screening and tobacco treatment among high-risk patients. We present a new care paradigm (**Figure 1**) and outline a translational roadmap (**Figure 2**) that discusses potential barriers and solutions for implementation. By incorporating personalized risk assessments, such PRS-enabled interventions have the potential to significantly improve lung cancer prevention strategies and patient outcomes.

Results

Sample Characteristics

For primary validation, we used data from 340,154 unrelated individuals in the UK Biobank (UKBB)[20], given its large sample size, rich clinical data, and inclusion of individuals from diverse ancestry backgrounds (**Methods**). Lung cancer validation involved 1,830 cases and 338,324 controls across five self-reported ethnic backgrounds: European (EUR, N = 318,043, N_{case} = 1,762), African (AFR, N = 6,409, N_{case} = 19), East Asian (EAS, N = 599, N_{case} = 2), and South Asian (SAS, N = 7,520, N_{case} = 10), with 7,583 (N_{case} = 37) with “Other” self-reported ethnic backgrounds such as mixed or unknown (**Supplementary Table 1a**). Lung cancer occurrence was slightly higher among men (53.3% of cases) compared to women (46.7% of cases).

For difficulty quitting, the cohort comprised 34,923 current smokers and 117,483 individuals who had previously smoked. The breakdown by self-reported ethnicity for this analysis was as follows: EUR (N = 145,483, 95.4%), AFR (N = 1,874, 1.2%), EAS (N = 131, 0.1%), and SAS (N = 1,733, 2.2%) (**Supplementary Table 1b**). Among those who had quit smoking, 50.3% were male among Europeans, while quitting in non-European males varied between 51.2% and 81.5%.

In addition, we validated the PRSs using data from the Genetically Informed Smoking Cessation (GISC) trial, which more accurately reflects the patient demographics anticipated in the PRECISE and MOTIVATE trials. The difficulty quitting analysis encompassed 647 current smokers and 149 former smokers, with the ancestry distribution as follows: 503 of European descent, 257 of African descent, and 36 of other ancestries (**Supplementary Table 1c**).

Harmonization PRS distributions across ancestry

To harmonize the polygenic risk scores (PRS) across diverse ancestries, we first projected genotypes of individuals from the UKBB and GISC onto a principal components analysis (PCA) space using PC loadings derived from 55,248 variants within 1000G dataset (**Methods**). The resulting PC scores aligned closely with the continental ancestries represented in the 1000G, confirming that projecting genotypes to an externally-defined PC-space still maintains similar clustering by ethnicity (**Figure 3**). The SNPs used and their corresponding loadings for the top five PCs are detailed in **Supplementary Table 4**.

We constructed PRSs for lung cancer and smoking cessation using large multi-ancestry GWAS with publicly accessible summary statistics (**Methods, Supplementary Tables 2-3**). Variations in the raw PRS across ancestries were notable (**Figure 4**). For instance, only 6% of AFR individuals in the UKBB cohort had a raw lung cancer PRS above the 80th percentile when benchmarked against the 1000G distribution. Conversely, 57% of EAS individuals ranked below the 33rd percentile for the difficulty quitting PRS (**Supplementary Table 6a-b**). Such differences indicate that applying a universal cutoff for PRS without ancestry adjustment could lead to skewed risk profiling and inaccurate clinical recommendations. Even with the smaller sample size of the GISC dataset, there were noticeable difference in PRS distribution similar to those in the UKBB data (**Supplementary Tables 6c-d**).

To address this, we use a two-step ancestry adjustment procedure that regresses out ancestry PCs from the raw PRS such that the mean and variance of the PRS distribution are consistent across all populations (**Methods**, **Figure 4**, **Supplementary Table 5**). This adjustment step places individuals from different ancestries on a

standardized scale, enabling the use of a single risk stratification cutoff irrespective of an individual's ancestral background. After ancestry adjustment, the corresponding proportions of the UKBB individuals within each risk category closely match 20%-60%-20% for lung cancer, and 33.3%-33.3%-33.3% for difficulty quitting, so that patients of any background can be appropriately compared against a single reference distribution for each outcome (**Supplementary Table 6a-b**).

Risk stratification for lung cancer and smoking cessation

All participants receiving interventions in our two ongoing trials are high-risk primary care patients who meet the criteria for lung cancer screening with elevated risks of lung cancer. Thus, we will assign patients to one of three PRS-based risk categories – "at risk", "high risk", and "very high risk" – using percentile cutoffs of ancestry-adjusted PRS distributions based on 1000G. To quantify patient risk, we calculated odds ratios (ORs) relative to the "at risk" group using 350,154 UKBB participants for lung cancer and 152,406 for difficulty quitting (**Figure 5**, **Supplementary Tables 4-5**). For lung cancer, individuals within the 0-20th percentiles of the adjusted PRS distribution were categorized as "at risk", those in the 20-80th percentiles as "high risk", and the 80-100th percentiles as "very high risk". These percentiles yielded overall ORs of 1.42 (95% CI: 1.24 – 1.65) for "high risk" and 1.85 (95% CI: 1.58 – 2.18) for "very high risk" group compared to the "at risk" group (**Supplementary Table 7a**). Notably, the ORs derived from our ancestry-adjusted PRS were nearly identical to those obtained by matching individuals' raw PRS values with ancestry-specific distributions in 1000G (**Supplementary Table 7b**). This indicates that ancestry adjustment not only preserves similar results as the ancestry-matched approach but also crucially supports the inclusion and robust risk stratification of individuals with mixed or unknown ethnic backgrounds.

However, differences in ORs were still observed between EUR and non-EUR participants, potentially due to the limited number of non-EUR cases (67 out of 1,830 total lung cancer cases) in the UKBB cohort, which is predominantly EUR. The OR for the EUR "very high risk" group was 1.85 (95% CI: 1.57 – 2.19), which is consistent with the combined odds ratios across all ancestries (**Supplementary Table 7a**). The OR for all non-EUR samples in the "very high risk" category was modestly reduced to 1.63 (95% CI:

0.78 – 3.51) (**Supplementary Table 7b**). Given the small number of non-European cases, enhanced diversity in biobank-scale validation data should better illustrate the difference in risk stratification between raw and ancestry-adjusted PRS.

Since difficulty quitting is a behavior trait with no established absolute risk rates like cancer, we use terciles (0-33rd, 33rd-67th, and 67-100th percentiles) of the ancestry-adjusted PRS distribution to provide provides with slightly more agnostic risk information. The resulting ORs among UKBB participants were 1.19 (95% CI: 1.15 – 1.22) for "high risk" and 1.36 (95% CI: 1.32 – 1.41) for "very high risk" relative to "at risk" (**Supplementary Table 8a-b**). Further validation using smoking status outcomes from the GISC trial assessed showed higher overall ORs in both risk categories, but risk stratification using the ancestry-adjusted PRS distribution still reflected similar odds ratios as ancestry-matched PRS distributions (**Supplementary Table 9a-b**). However, since the GISC trial have much smaller sample size compared to UKBB, the confidence intervals were notably wider. Similar to the lung cancer analysis, the ORs using our ancestry-adjusted PRS aligned closely those derived from ancestry-matched raw PRS.

Using ancestry-adjusted PRS ensures accurate risk stratification across all ethnic backgrounds, a critical consideration given the substantial variability in raw PRS distributions across diverse populations (**Figure 4**). The outcome-based validation in UKBB and GISC further verify that the ancestry-adjusted PRS yields valid risk stratification. These findings collectively facilitate more robust and standardized application of PRS in clinical reporting.

Translating polygenic risk scores into clinical reports

We highlight two example trials: PRECISE (NIDA Grant 5R01CA268030-02) and MOTIVATE (NIDA Grant 5R01DA056050-02), designed to promote health behavior change using genetically-informed interventions, RiskProfile and PrecisionTx, respectively. These interventions incorporate PRS in communicating precision risk of lung cancer and precision benefits of smoking cessation to promote evidence-based practices such as cancer screening and tobacco treatment in high-risk individuals who smoke and/or are eligible for lung cancer screening. PRS risk stratification from either RiskProfile or PrecisionTx and clinical information are delivered within a comprehensive report, along

with actionable recommendations to reduce lifetime risk (**Figure 6**). Access to 23andme genotypes and expanded health information has been a motivating component for the research participants. Both PRECISE and MOTIVATE are currently in the preliminary phases of recruiting primary care providers and patients. The recruitment strategy aims to engage over 100 physicians and 1600 patients in these trials.

Discussion

In this study, we introduce the GREAT framework in primary care. The application of PRSs in the two example trials offers precise risk estimates for lung cancer and difficulty quitting to high-risk individuals to activate behavior change mechanisms that promote health. To enable the interventions, we present a feasible translational roadmap to transform genetic data, implemented in two example PRS-enabled interventions designed to promote health behaviors. These behavior-change tools will be evaluated for implementation and effectiveness in motivating patients at high risk to reduce their risk by increased cancer screening and smoking cessation.

Importantly, we have accomplished our goals of generating behavior-change interventions by a) framing our translational message specifically for high-risk patients who have not received guideline-recommended cancer screening or tobacco treatment[21–24], b) translating risk categories into precision risk and benefit that are designed to motivate health behavior changes, and c) ensuring inclusion of diverse ancestry with PC-regression-based PRS adjustment.

Our goal is to enable a robust implementation of PRS in currently funded clinical trials to evaluate the efficacy of these relatively novel interventions. The GREAT framework guides the implementation of PRS-enabled interventions in primary care settings. Critical questions such as timing, methodology, and location of these interventions' delivery to patients and providers are addressed to optimize its acceptability, understanding, and potential impact.

Our approach to ancestry adjustment of PRS employs widely accessible data from the 1000G dataset, as an alternative to methods in the GenoVA[11] and eMERGE[10] studies that use data from the Mass-General Brigham Biobank and All of Us, respectively. We have validated the transferability of our 1000G-based standardization in external datasets

from the UKBB and GISC, allowing future trials to adopt a similar methodology irrespective of their specific genetic data. By utilizing our provided PC loadings and PRS standardization formula for lung cancer and difficulty quitting based on 1000G data, new patients in these trials can receive accurate risk categorization reports, bypassing the potential inaccuracies of self-reported ethnicity and the need for re-training PCA models.

A notable gap in current practice is the absence of genetic information in electronic health records (EHRs) for decision support and the lack of PRS generation in clinical labs. Implementing multilevel precision interventions in primary care necessitates a workflow that incorporates the use of EHRs for recruitment, protocols of biomarker testing, and a standardized process to generate the personalized intervention reports[25]. This requires collaborations with primary care stakeholders, community advisory boards, genetic counseling, and health communication to improve the messaging and visualization for intervention clarity, accuracy, and impact[26–30]. Patients expressed a notable interest in receiving personalized interventions. In our previous study, 85% of smoking patients reported a high interest in receiving genetically tailored tobacco treatment[31]. Further, a substantial majority (95%) of individuals who smoke endorsed the importance of receiving genetic results, in particular to guide their treatment[27]. Following receipt of a personalized genetic risk profile for smoking cessation, 91% of participants who smoke found the tool to be highly useful, most notably to better understand their health, cope with health risks, and feel more in control of their health[28]. Such pronounced interest and the perceived significance of genetic data highlight the growing demand for personalized interventions among patients who smoke. Personalized interventions may further increase patient compliance**.** For example, our study found that patients reported higher interest in taking medication (97.5% vs. 61.0%, p<.0001) when medication was personalized based on their genetics[31]. These data demonstrate the translation potential of personalized genetics in motivating patients for positive behavior change.

Unlike most current research that evaluates PRS-enabled interventions in general patient populations, our work provides a unique aspect by designing and evaluating these interventions specifically among high-risk patients who will benefit tremendously from the recommended health behaviors (lung cancer screening and smoking cessation) when general medical advice is not enough to motivate such behaviors.

Here we share three key design considerations for best practices. First, for equitable implementation of precision health interventions, tools must be designed with racial/ethnic minority communities engaged in the development process at the outset, rather than solely examining whether these interventions work for these communities post hoc. We engage in ongoing participatory sessions with racial and ethnic minority communities and advisory boards across all phases through iterative cycles of intervention development, feedback, and testing so that innovative genomics-informed tools are designed for use and benefit across diverse populations.

Second, we have chosen clinically meaningful thresholding for PRS risk categories in communicating personalized risks and benefits with patients in our research. The categories were selected because all participants receiving interventions in our two ongoing trials are high-risk primary care patients eligible for lung cancer screening, who are current or previous heavy smokers. We defined lung cancer risk by the bottom 20%, middle 60% and 20%, and difficulty quitting smoking by slightly more agnostic tertiles, to motivate positive behavior change. We aim to follow best practices of communicating uncertainties. Furthermore, we need to make decision on options of risk presentation such as a continuous or categorical assignment. Importantly, we strive to be transparent about the imprecision in both risk estimates and action thresholds for PRS.

Third, we expect to update our workflow to adapt to new GWAS and evolving methodologies. As scientific knowledge rapidly progresses, outdated or inaccurate PRS predictions can hinder effective implementation. To address this challenge, a dynamic PRS framework that allows for regular updates based on new scientific discoveries is necessary. This will ensure that the PRS-based intervention remains current and aligned with the latest advancements, ultimately shortening the implementation gap and maximizing its impact on preventive healthcare outcomes. Leveraging current recommendations on genetic counseling, we have established a process and threshold to incorporate new evidence into our intervention regarding smoking cessation and lung cancer risk. This process will a) adjudicate population specific new evidence regarding genetics and biomarkers, b) evaluate its impact on changes in risk levels at personal and population level, and c) develop effective communication regarding the dynamic nature of genetic evidence with patients and providers.

There are several limitations in our work as we hope to share our experiences to help inform the knowledge pool for the best practices in creating PRS-enabled interventions that may be disease-, population-, or context-specific. We have tailored our approach to our unique outcomes (lung cancer and difficulty quitting smoking), population (patients eligible for lung cancer screening and tobacco treatment), and context (primary care settings) to optimize the potential health impact of our intervention tools. Another notable limitation is the underrepresentation of non-European populations in the multi-ancestry GWAS employed to derive the PRS weights, as well as in the UKBB used to evaluate the PRSs. These factors may reduce the predictive power of the PRS in non-European populations. A dominance of European populations persists in most existing GWAS[32–35], not limited to lung cancer and smoking cessation. With the burgeoning emphasis on incorporating minority populations in GWAS analyses and the ongoing development of new PRS approaches[7,8,36–39] that focus on enhancing predictive power in diverse populations, we can iteratively refine the PRS implementation in our trial to synchronize with the most contemporary advancements.

Many questions need to be answered in the near future. First, <u>how can we reduce the time lag from evidence to implementation?</u> One challenge is the constant evolution of evidence that identifies new biomarkers for treatment. For instance, our recent work highlights the potential of polygenic risk scores in guiding future treatment approaches[40]. However, despite the presence of actionable precision treatment findings, the ever-changing evidence base and the perception that even better data are on the horizon have hindered effective implementation[29,41]. In this proposal, we seek to overcome this challenge by utilizing cutting-edge, biology-based metabolic and genetic markers that offer robust evidence for precision treatment. The motivation behind this approach is to reduce the time lag from evidence generation to practical implementation, particularly in the context of precision medicine, where the evidence base is continuously evolving and dynamic. This affords the unique opportunity to measure and report on the time from landmark publications to implementation of key findings, an approach that is being increasingly called for in translational science[42]. By leveraging state-of-the-art markers, we aim to enhance the efficiency and effectiveness of precision treatment and ensure that patients can benefit from the latest and most accurate recommendations.

Second, can we truly evaluate the effect of precision interventions? We expect that precision interventions may activate multiple mechanistic pathways to the uptake and efficacy of lung cancer screening or tobacco treatment. Understanding potential plausible mechanisms is needed to improve or refine the intervention for intended outcomes and contexts. Third, can these precision interventions be scaled in the real-world clinics? Evidence has shown that physicians are highly receptive to guidance on medication recommendations based on biomarkers[43]. To reduce burden, we need to leverage existing EHR tools (e.g. Best Practice Advisories) to efficiently facilitate physician prescribing[44,45]. Understanding of mechanistic and implementation outcomes ill guide scalable, efficient delivery components for integration into clinic workflows[25], trained embedded staff, and digital therapeutic tools to enable these PRS-informed behavioral interventions[46].

In conclusion, a well-designed roadmap that validates the PRS, creates it using TE weights, translates risk into actionable categories, communicates effectively, considers patient perspectives, and accommodates evolving science is essential for the equitable and pragmatic translation of PRS into clinical care. By addressing the barriers and implementing potential solutions at each stage, we can harness the power of PRS to improve preventive healthcare and make a meaningful difference in reducing the burden of diseases like lung cancer.

Acknowledgements

We would like to thank Reeya Joseph for her editorial support with the introduction, Peter Kraft for his advice on our manuscript, and Scott Vrieze for his assistance with summary statistics for difficulty quitting. We would also like to thank and acknowledge the participants enrolled in the UK Biobank (obtained under UK Biobank resource application 52008) and GISC trial for contributing vital data to this work.

This research was supported by NIH Training Grant T32GM135117 and NSF Graduate Research Fellowship DGE-2140743 (T.C.), R35-3CA197449, R01-HL163560, U01-HG009088, and U01-HG012064 (X.L.), NIH Intramural Research Program (H.Z.), NIH 5T32-HL007776-25, R01-DA056050, R01-CA268030, P30-CA091842-19S5, P30-CA091842-16S2 and P50-CA244431 (L.C.) and U19-CA203654.

Author Contributions

T.C., H.Z., and L.C. conceived the project, T.C., G.P., and L.F. carried out all data analyses under the supervision of H.Z. and L.C. G.S. and D.J. assisted with obtaining and processing GWAS summary statistics for difficulty quitting. T.C., G.P., H.Z., and L.C. drafted the manuscript. All co-authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

Laura J. Bierut is listed as an inventor on Issued U.S. Patent 8,080,371, “Markers for Addiction” covering the use of certain SNPs in determining the diagnosis, prognosis, and treatment of addiction. Michael J. Bray is an employee at ThinkGenetic, Inc. Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy, or views of the International Agency for Research on Cancer/World Health Organization. All other authors have no conflict of interests to report.

Data and Code Availability

R code and plink commands, as well as accompanying data, used for analysis are provided in a walkthrough available on GitHub at https://github.com/chen-tony/GREAT.

Genotype data from the 1000 Genomes Project (Phase 3) were acquired via the plink website https://www.cog-genomics.org/plink/2.0/resources. Population information were obtained from public resources: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ and https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/1000G_698_related_high_coverage.sequence.index.

Figures

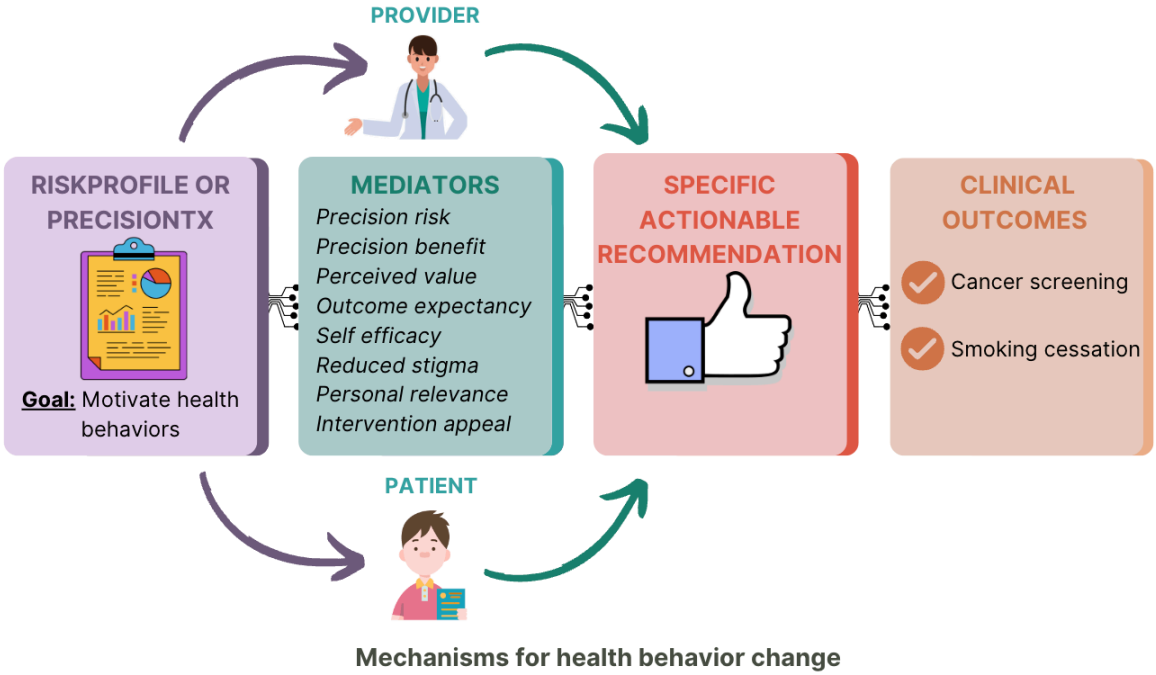

Figure 1. Care Paradigm: Genomic Informed Care for Motivating High Risk Individuals Eligible for Evidence-based Prevention (GREAT). The GREAT framework is a primary care paradigm that integrates genetic and clinical risk in precision health. Individuals and their providers in two upcoming trials (PRECISE and MOTIVATE) are enrolled and provided with multilevel interventions (e.g. RiskProfile and PrecisionTx) to promote clinical outcomes of lung cancer screening, tobacco treatment, and successful smoking cessation in primary care settings. Mechanisms of health behavior changes (e.g., perceived benefit, self-efficacy, and outcome expectancy) will be evaluated. During the specific actionable recommendations phase, personalized shared decision-making will be facilitated by multilevel actions between patients and clinicians for better clinical outcomes.

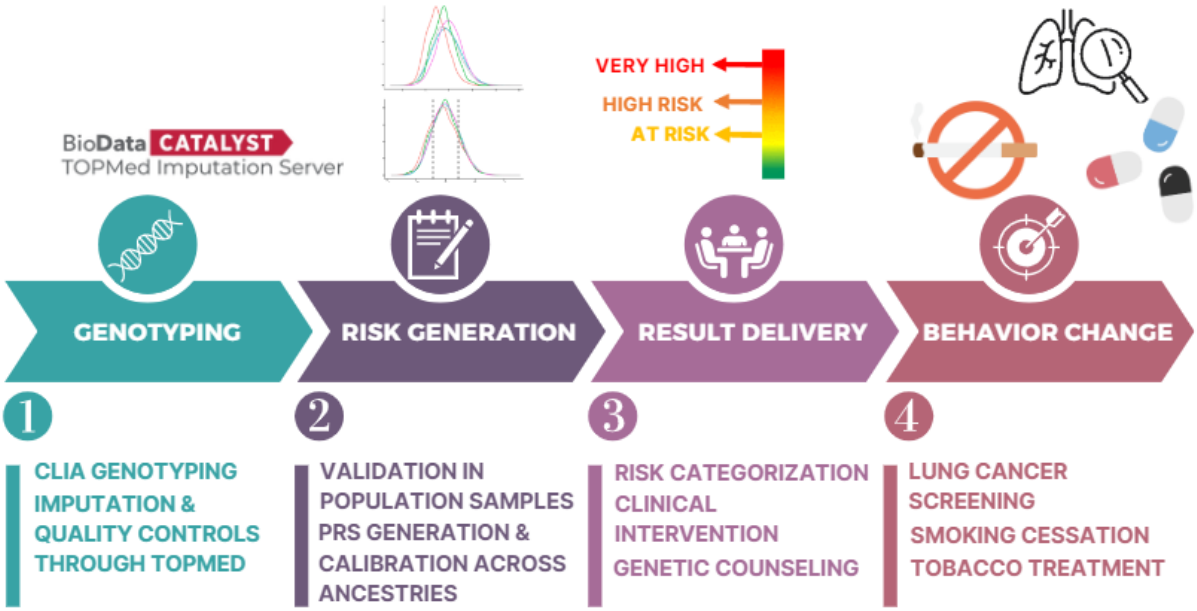

Figure 2 Roadmap for translating genetic data to a genetic risk profile as a multilevel intervention in primary care. In step 1, enrolled participants' genetic data are analyzed by 23andMe's Clinical Laboratory Improvement Amendments (CLIA) certified genotyping process. Imputation and quality controls are conducted through the Trans-Omics for Precision Medicine (TOPMed) server to ensure the integrity and reliability of the genetic data, as well as to impute the GWAS variants. Step 2 involves identifying available GWAS variants and weights to create the raw Polygenic Risk Scores (PRS). The PRS is adjusted for genetic ancestry using reference data such as the 1000 Genomes Project Phase 3 and applied to validation data such as the UK Biobank to establish risk categories and compute odds ratios. In step 3, these scores are converted into 3 risk levels based on the established thresholds. In step 4, a report with precision treatment is created and communicated to both the participant and the provider to make informed and educated decisions. Behavioral interventionists offer personalized guidance on behavior change, leveraging the updated genetic insights. The outcome aims to increase lung cancer screening orders, improve participant adherence, promote smoking cessation, and highlight the benefits of tobacco treatment.

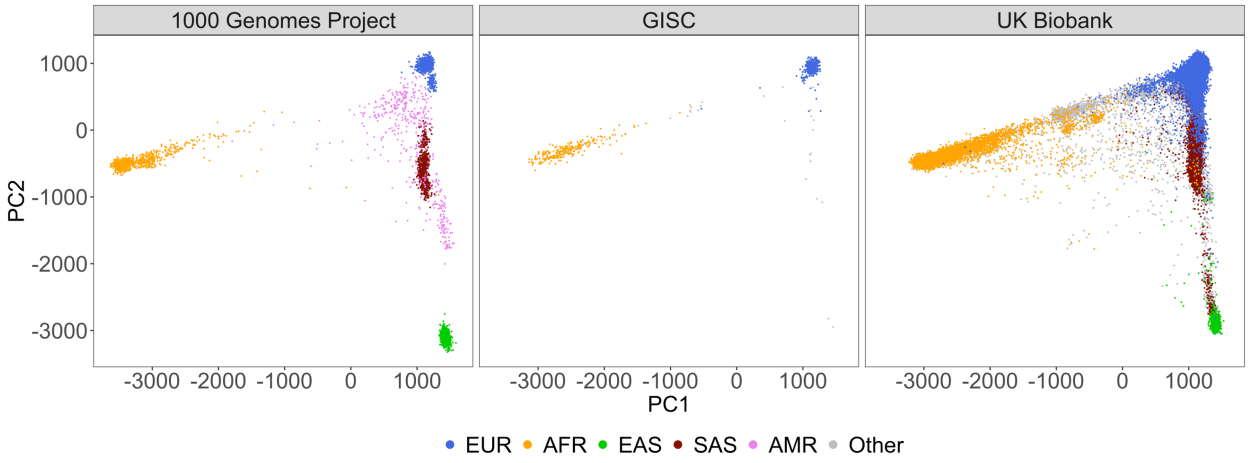

Figure 3. Cross-dataset discrimination of self-reported ethnicity via PCA Projections in 1000G, UKBB and GISC. This figure illustrates the utility of principal components analysis (PCA) loadings obtained from the 1000 Genomes Project Phase 3 (1000G) in discriminating ancestries within external datasets, specifically the UK Biobank (UKBB) and the Genetically Informed Smoking Cessation (GISC) trial. PCA was initially conducted on the globally diverse genotype data of 1000G. The resultant PCA-space was then used to project genotype data from both the UKBB and GISC. The scatter plot displays the first and second PCs for each individual in these datasets, with points distinctly marked by self-reported ethnicity.

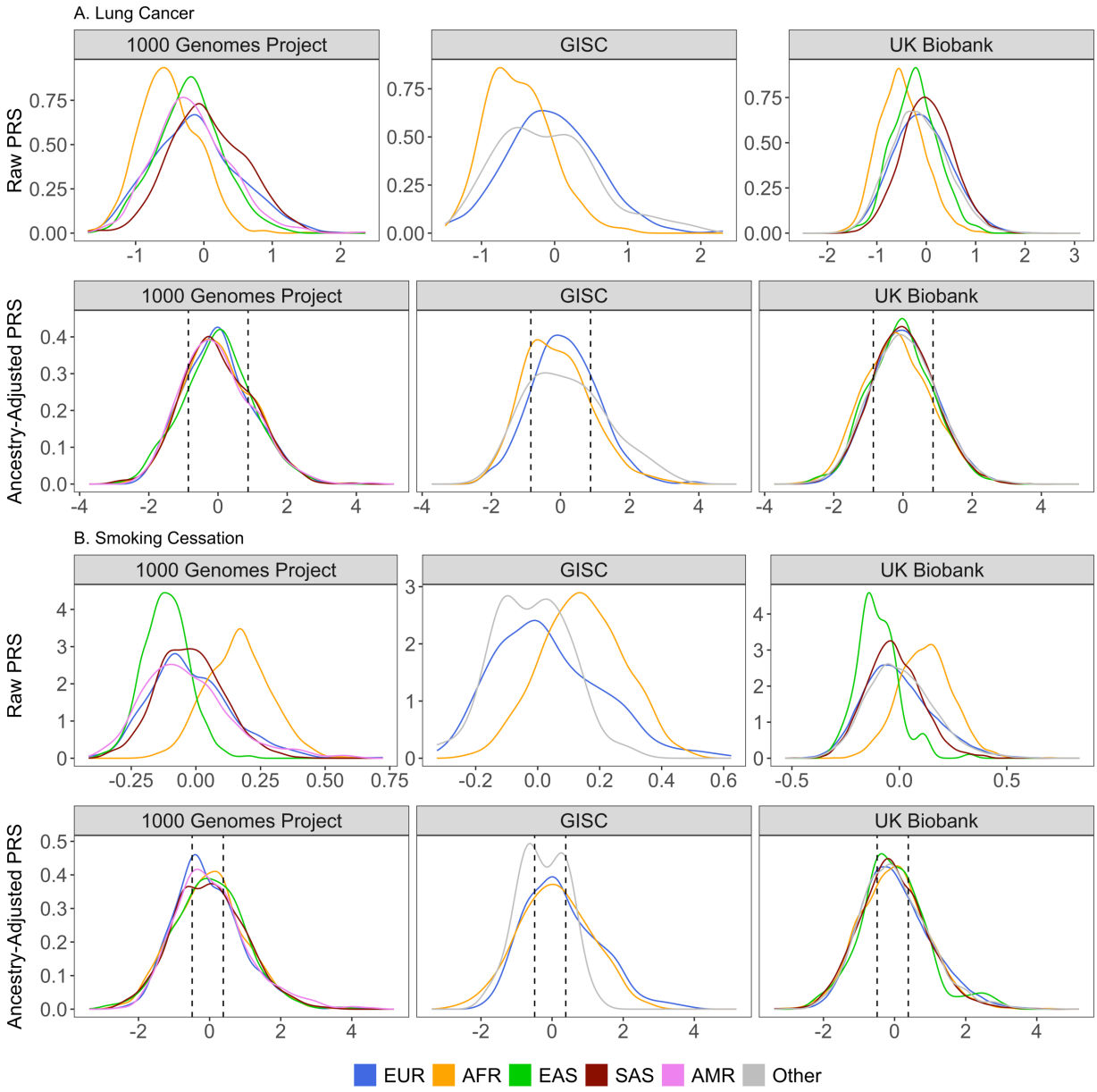

Figure 4 Ancestry adjustment of PRS for lung cancer and quit difficulty PRS across ancestral populations. We showcase the adjustment process for polygenic risk scores (PRS) for lung cancer (Panel A) and difficulty quitting smoking (Panel B) within the 1000 Genomes Project (1000G) and UK Biobank datasets. It displays both raw and ancestry-adjusted PRS, with data points color-coded according to self-reported continental ancestries. Ancestry adjustment effectively centers the PRS for different ancestries, mitigating the risk of incorrect stratification due to ancestry-related biases. Dotted vertical lines correspond to the 20th and 80th percentiles for lung cancer PRS distribution and 33rd and 67th percentiles for difficulty quitting PRS among all 3,202 samples in the 1000 Genomes Project.

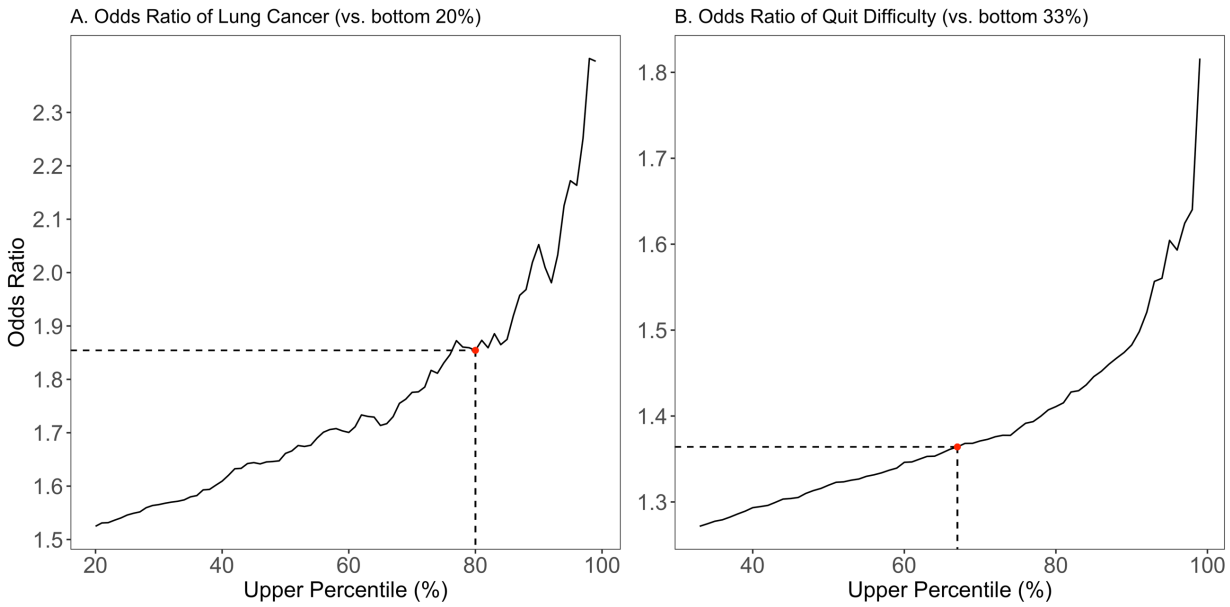

Figure 5. Risk stratification through ancestry-adjusted PRS percentiles and associated odds ratios. This figure illustrates the odds ratios (ORs) calculated for lung cancer (Panel A) and difficulty quitting smoking (Panel B) among UK Biobank participants (N=340,154 for lung cancer and N=152,406 for smoking), based on selected cut points within the ancestry-adjusted PRS distribution. The dashed lines mark the upper percentiles used for defining risk categories in our study: the 80th percentile for lung cancer, correlating with an OR of 1.85 (95% CI: 1.58 – 2.18) and the 67th percentile for quit difficulty, corresponding to an OR of 1.36 (95% CI: 1.32-1.41).

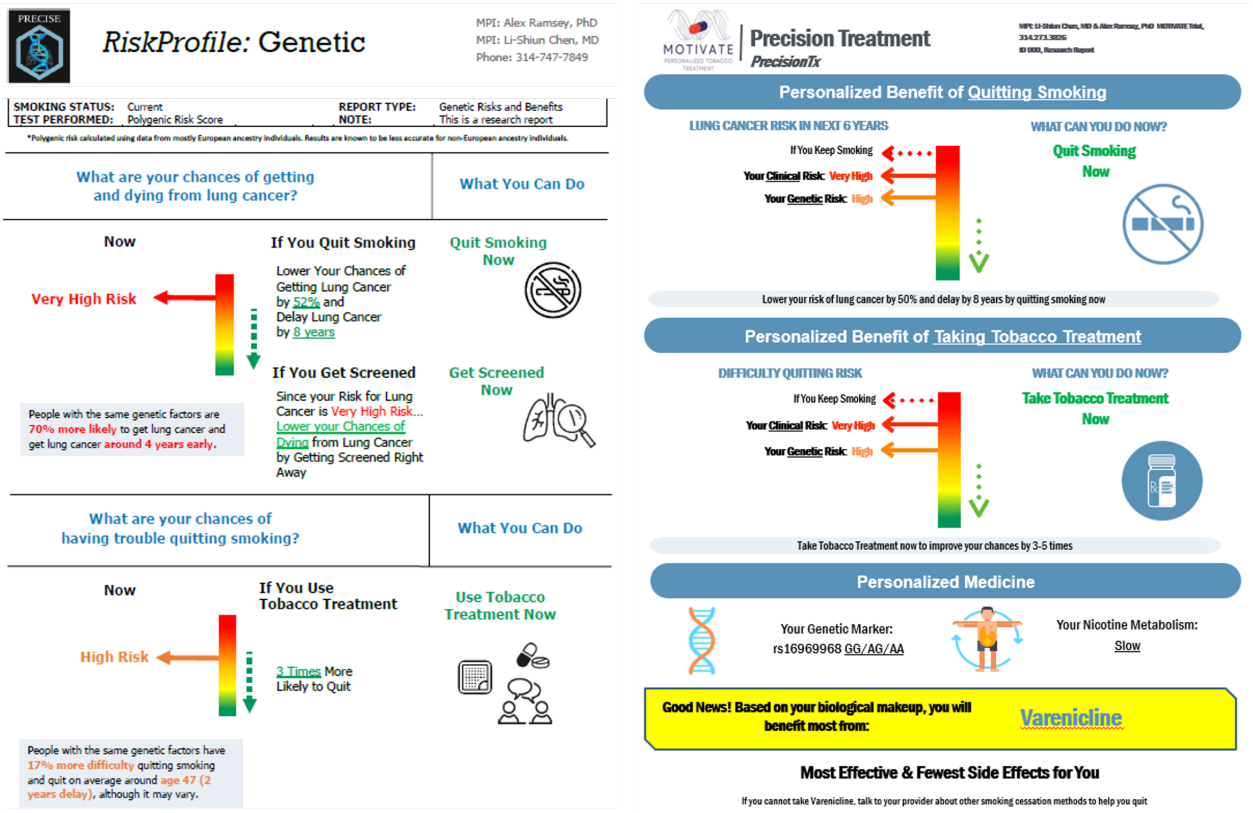

Figure 6. Example clinical reports for lung cancer (left) and smoking (right). We present two genomically-informed interventions using the GREAT framework. RiskProfile (left) is designed to motivate lung cancer screening and tobacco treatment among eligible patients. PrecisionTx (right) is designed to motivate precision tobacco treatment and smoking cessation. Both tools utilize ancestry-adjusted PRS to stratify patients into "at risk" (yellow), "high risk" (orange), and "very high risk" (red) genetic risk categories. While RiskProfile focuses more on prevention and PrecisionTx focuses more on treatment, both interventions expand beyond personalized risk to also provide personalized benefit of cancer screening and personalized medication recommendation, and use a multilevel intervention design directed to both physicians and patients in clinical settings.

Tables

	GenoVA	eMERGE	Wisdom	PRECISE / MOTIVATE
Goal	Motivate early diagnosis	Increasing clinical actions to mitigate risk of future disease	Personalized screening frequency	Motivate lung cancer screening (LCS) / tobacco treatment
Targets				
Target population	All general Veterans' Affairs PCP* and patients	Patients in the large healthcare system	All women can sign up online	**High risk** PC patients (LCS eligible / smokers) and their PCPs
Condition(s)**	BrCa, PrCa, CRCa, Afib, CAD, T2D	BrCa, PrCa, CRCa, Afib, CAD, T1D, T2D, BMI, Asthma, CKD, HCL	BrCA	**LC, smoking cessation**
Target outcome	New diagnoses	Change screening practice or lifestyle	Screening compliance and cancer detection	**LCS, tobacco tx, and smoking cessation**
Polygenic Risk Scores				
Genotyping	Study team		Color genomics	23andMe
Imputation	1000 Genomes	All of Us		TOPMed
Create PRS	Top or many variants European weights Ancestry calibration	Terra cloud platform Ancestry calibration	Color genomics	Top independent variants Multi-ancestry weights Ancestry calibration
Risk Stratification	2 levels (OR of 2)	OR of 5-10		3 levels (top 20%, middle, bottom 20%) or tertiles
Intervention Design				
Risk representation	Only genetic Format of relative risk	Combined clinical + genetic Format of absolute risk	Combined clinical + genetic Format of absolute risk	Separate genetic + clinical Format of relative risk
Messaging	Average vs. elevated risk	Average vs. elevated risk	No discussion of risk Only recommend screening schedule	At risk, at high risk, or at very high risk
Behavior mechanisms targeted				**Perceived risk/benefit, Self efficacy, Personal relevance**

* PCP=primary care provider

** BrCa=breast cancer, PrCa=prostate cancer, CRCa=colorectal cancer, Afib=atrial fibrillation, CAD=coronary artery/heart disease, T2D=type 2 diabetes, T1D=type 1 diabetes, BMI=body mass index/obesity, CKD=chronic kidney disease, HCL=hypercholesterolemia, lung cancer=lung cancer

Table 1. **Research on PRS use in clinical settings**. We compare the PRECISE and MOTIVATE trials, part of our GREAT framework, with existing PRS-informed trials: GenoVA, eMERGE, and WISDOM. Bolded text in the PRECISE / MOTIVATE column highlight the points where our trials differ from the current trials. Namely, the PRECISE and MOTIVATE trials investigate lung cancer and smoking and will focus on high risk patients who are smokers or eligible for lung cancer screening. We also look at lung cancer screening, tobacco treatment, and smoking cessation as unique target outcomes. Finally, in addition to genetic and clinical risk messaging, the two trials have a unique emphasis on behavior mechanisms around lung cancer and smoking.

References

1. Kocarnik, J. M. *et al.* Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life Years for 29 Cancer Groups From 2010 to 2019. *JAMA Oncol* **8**, 420 (2022).
2. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* vol. 19 581–590 Preprint at https://doi.org/10.1038/s41576-018-0018-x (2018).
3. Lewis, C. M. & Vassos, E. Polygenic risk scores: From research tools to clinical instruments. *Genome Medicine* vol. 12 Preprint at https://doi.org/10.1186/s13073-020-00742-5 (2020).
4. Adeyemo, A. *et al.* Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nature Medicine* vol. 27 1876–1884 Preprint at https://doi.org/10.1038/s41591-021-01549-6 (2021).
5. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet* **97**, 576–592 (2015).
6. Wray, N. R., Goddard, M. E. & Visscher, P. M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res* **17**, 1520–1528 (2007).
7. Zhang, H. *et al.* A new method for multiancestry polygenic prediction improves performance across diverse populations. *Nat Genet* (2023).
8. Ruan, Y. *et al.* Improving polygenic prediction in ancestrally diverse populations. *Nat Genet* **54**, 573–580 (2022).
9. Chen, T., Zhang, H., Mazumder, R. & Lin, X. Ensembled best subset selection using summary statistics for polygenic risk prediction. *bioRxiv* (2023).
10. Linder, J. E. *et al.* Returning integrated genomic risk and clinical recommendations: The eMERGE study. *Genetics in Medicine* **25**, 100006 (2023).
11. Hao, L. *et al.* Development of a clinical polygenic risk score assay and reporting workflow. *Nat Med* **28**, 1006–1013 (2022).
12. 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**, djw290 (2017).
13. Zhang, P. *et al.* Association of smoking and polygenic risk with the incidence of lung cancer: a prospective cohort study. *Br J Cancer* **126**, 1637–1646 (2022).
14. Kanwal, M., Ding, X.-J. & Cao, Y. Familial risk for lung cancer. *Oncol Lett* **13**, 535–542 (2017).
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. Lewis, A. C. F. *et al.* Getting genetic ancestry right for science and society. *Science (1979)* **376**, 250–252 (2022).

17. Ding, Y. *et al.* Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* (2023).
18. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
19. Ge, T. *et al.* Development and validation of a trans-ancestry polygenic risk score for type 2 diabetes in diverse populations. *Genome Med* **14**, 70 (2022).
20. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med* **12**, e1001779 (2015).
21. US Preventive Services Task Force. Lung Cancer: Screening. *https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/lung-cancer-screening* (2021).
22. Agency for Healthcare Research and Quality. Clinical Guidelines and Recommendations. *https://www.ahrq.gov/prevention/guidelines/index.html* (2021).
23. Tobacco Use and Dependence Guideline Panel. *US Department of Health and Human Services. (2008). Tobacco Use and Dependence Guideline Panel. Treating Tobacco Use and Dependence: 2008 Update*. (US Department of Health and Human Services, Rockville, MD, 2008).
24. Krist, A. H. *et al.* Screening for Lung Cancer. *JAMA* **325**, 962 (2021).
25. Ayatollahi, H., Hosseini, S. F. & Hemmat, M. Integrating Genetic Data into Electronic Health Records: Medical Geneticists' Perspectives. *Healthc Inform Res* **25**, 289 (2019).
26. Bourdon, J. L. *et al.* In-vivo design feedback and perceived utility of a genetically-informed smoking risk tool among current smokers in the community. *BMC Med Genomics* **14**, 139 (2021).
27. Ramsey, A. T. *et al.* Participatory Design of a Personalized Genetic Risk Tool to Promote Behavioral Health. *Cancer Prevention Research* **13**, 583–592 (2020).
28. Ramsey, A. T. *et al.* Proof of Concept of a Personalized Genetic Risk Tool to Promote Smoking Cessation: High Acceptability and Reduced Cigarette Smoking. *Cancer Prevention Research* **14**, 253–262 (2021).
29. Ramsey, A. T. *et al.* Toward the implementation of genomic applications for smoking cessation and smoking-related diseases. *Transl Behav Med* **8**, 7–17 (2018).
30. Chen, L.-S., Baker, T. B., Ramsey, A., Amos, C. I. & Bierut, L. J. Genomic medicine to reduce tobacco and related disorders: Translation to precision prevention and treatment. *Addiction Neuroscience* **7**, 100083 (2023).
31. Chiu, A. *et al.* Most Current Smokers Desire Genetic Susceptibility Testing and Genetically-Efficacious Medication. *Journal of Neuroimmune Pharmacology* **13**, 430–437 (2018).

32. Peterson, R. E. *et al.* Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell* **179**, 589–603 (2019).
33. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
34. Need, A. C. & Goldstein, D. B. Next generation disparities in human genomics: concerns and remedies. *Trends in Genetics* **25**, 489–494 (2009).
35. Fitipaldi, H. & Franks, P. W. Ethnic, gender and other sociodemographic biases in genome-wide association studies for the most burdensome non-communicable diseases: 2005–2022. *Hum Mol Genet* **32**, 520–532 (2023).
36. Kachuri, L. *et al.* Principles and methods for transferring polygenic risk scores across global populations. *Nat Rev Genet* (2023).
37. Jin, J. *et al.* MUSSEL: Enhanced Bayesian Polygenic Risk Prediction Leveraging Information across Multiple Ancestry Groups. *bioRxiv* (2023).
38. Zhou, G., Chen, T. & Zhao, H. SDPRX: A statistical method for cross-population prediction of complex traits. *The American Journal of Human Genetics* **110**, 13–22 (2023).
39. Zhang, J. *et al.* An Ensemble Penalized Regression Method for Multi-ancestry Polygenic Risk Prediction. *bioRxiv* (2023).
40. Bray, M. *et al.* The Promise of Polygenic Risk Prediction in Smoking Cessation: Evidence From Two Treatment Trials. *Nicotine & Tobacco Research* **24**, 1573–1580 (2022).
41. Ramsey, A. T. *et al.* Designing for Accelerated Translation (DART) of emerging innovations in health. *J Clin Transl Sci* **3**, 53–58 (2019).
42. Proctor, E. *et al.* FAST: A Framework to Assess Speed of Translation of Health Innovations to Practice and Policy. *Global Implementation Research and Applications* **2**, 107–119 (2022).
43. Scoville, E. A. *et al.* Precision nicotine metabolism-informed care for smoking cessation in Crohn's disease: A pilot study. *PLoS One* **15**, e0230656 (2020).
44. Chen, L.-S. *et al.* Low-Burden Strategies to Promote Smoking Cessation Treatment Among Patients With Serious Mental Illness. *Psychiatric Services* **69**, 849–851 (2018).
45. Ramsey, A. T. *et al.* Care-paradigm shift promoting smoking cessation treatment among cancer center patients via a low-burden strategy, Electronic Health Record-Enabled Evidence-Based Smoking Cessation Treatment. *Transl Behav Med* (2019).
46. Kaphingst, K. A. *et al.* Comparing models of delivery for cancer genetics services among patients receiving primary care who meet criteria for genetic evaluation in two healthcare systems: BRIDGE randomized controlled trial. *BMC Health Serv Res* **21**, 542 (2021).

47. 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**, 1167–1177 (2022).
48. Saunders, G. R. B. *et al.* Genetic diversity fuels gene discovery for tobacco and alcohol use. *Nature* **612**, 720–724 (2022).
49. Chen, S. *et al.* A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. *bioRxiv* 2022.03.20.485034 (2022).
50. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426-3440.e19 (2022).
51. Chen, L. *et al.* Genetic Variant in *CHRNA5* and Response to Varenicline and Combination Nicotine Replacement in a Randomized Placebo-Controlled Trial. *Clin Pharmacol Ther* **108**, 1315–1325 (2020).
52. Dey, R. *et al.* Efficient and accurate frailty model approach for genome-wide survival association analysis in large-scale biobanks. *Nat Commun* **13**, 5437 (2022).
53. Hinrichs, A. S. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34**, D590–D598 (2006).
54. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *The American Journal of Human Genetics* **100**, 635–649 (2017).
55. Manrai, A. K. *et al.* Genetic Misdiagnoses and the Potential for Health Disparities. *New England Journal of Medicine* **375**, 655–665 (2016).
56. Ochoa, A. genio: Genetics Input/Output Functions. Preprint at https://CRAN.R-project.org/package=genio (2023).
57. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904–909 (2006).
58. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
59. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* **50**, 1219–1224 (2018).
60. Krainc, T. & Fuentes, A. Genetic ancestry in precision medicine is reshaping the race debate. *Proceedings of the National Academy of Sciences* **119**, (2022).
61. Malina, D. *et al. Race and Genetic Ancestry in Medicine-A Time for Reckoning with Racism*. *n engl j med* vol. 384 (2021).

Methods

Reference and validation data

We conducted our analyses using data from three prominent datasets: UK Biobank (UKBB)[20], 1000 Genomes Project (1000G)[50], and Genetically Informed Smoking Cessation Trial (GISC)[51]. The UKBB, a widely recognized dataset, encompasses rich genetic and clinical data from approximately 500,000 British individuals. Our study specifically used data from 340,154 unrelated multi-ancestry individuals, up to third-degree relatives[52], who had consented as of September 5, 2023 (**Supplementary Table 1**). The 1000G dataset provides a globally diverse genetic reference of 3,202 individuals, with 633 Europeans (EUR), 893 Africans (AFR), 585 East Asians (EAS), 601 South Asians (SAS), and 490 Admixed Americans (AMR). We synchronized our data by using the latest data release on hg38 reference genome and using liftOver[53] to convert to hg37 and align with our UKBB genotype data.

UKBB was used for the primary validation of our PRS due to its considerable sample size and inclusion of non-European ethnicity. We used self-reported ethnicity using UKBB Field 21000, defining European (EUR) as White, British, Irish, or any other white background; African (AFR) as Black, Caribbean, African, Black or Black British, or any other black background; East Asian (EAS) as Chinese; and South Asian (SAS) as Indian, Pakistani, Bangladeshi, Asian or Asian British, or any other Asian background. From the 340,154 participants, the breakdown was 318,043 EUR, 6,409 AFR, 599 EAS, 7,520 SAS, and 7,583 with other self-reported ethnicity. Participants' mean age was 56.6 years (SD 8.2; range 38-81 years), and the cohort was 54.1% female (183,969 individuals). Our lung cancer analyses included 1,830 lung cancer cases in the UKBB, defined by whether a patient had at least one ICD10 code in C34.0-CD34.9 under Field 40006, and 338,334 controls with no ICD10 codes recorded (**Supplementary Table 1**). The smoking cessation analysis involved 152,406 "ever-smokers", including 117,483 former and 34,923 current smokers, with the latter defined as having difficulty quitting based on Field 20116. We excluded 186,040 'never smokers' and 1,312 participants who opted for 'prefer not to answer'.

The Genetically Informed Smoking Cessation (GISC) trial is a prospective, randomized, placebo-controlled smoking cessation trial conducted at Washington University in St Louis[51]. This study includes 822 total individuals, all of whom are smokers. We focused on 796 individuals with genetic information, including 503 of European self-reported ethnicity, 257 African self-reported ethnicity, and 36 self-reported as "Other". GISC data was used for secondary validation, as the patient population more closely resembles the expected patients enrolled in our PRECISE and MOTIVATE trials.

Construction of polygenic risk scores

Our trial incorporates the latest findings by utilizing recently genome-wide association study (GWAS) summary statistics for lung cancer[47] and difficulty quitting[48]

which exclude samples from the UK Biobank to avoid overlapping with our validation data. While these meta-analyses predominantly consist of individuals of European ancestry, they also include a substantial proportion of non-European ethnic background—about 26% for lung cancer and 21% for difficulty quitting—which enhances the generalizability of the findings[54,55].

For lung cancer risk, we started with a set of 128 published SNPs found to be predictive of 5-year and lifetime cumulative risk for lung cancer[15]. Out of these, 101 SNPs overlapped with the published summary statistics, reference, and validation data (1000G, UKBB, and GISC), and the 23andMe genotyping array used for the trial (**Supplementary Figure 1**). These SNPs were then assigned effect sizes from the fixed-effect meta-analyses estimates in the most recent lung cancer GWAS that includes EUR, AFR, and EAS ancestry[47]. Use of 23andme has been an incentive for patient participation.

For difficulty quitting, we started with 206 SNPs and SNP effects identified as predictive of smoking cessation[48]. Among these 206, we identified 177 SNPs following the same filtering procedure for lung cancer and used a final list of 175 SNPs after removing 2 multiallelic SNPs.

The PRS construction began with the alignment of genotype data to the summary statistics, ensuring consistent PRS regardless of initial reference and alternative allele coding. Specifically, for any SNP G with reversed alleles, we recoded it as $2 - G$ to avoid discrepancies. If we were to instead change the sign of the corresponding effect size β, there would be an added constant of 2β within the PRS, which can alter the overall PRS distribution and subsequent risk stratification if patient genotypes are coded differently. Once flipped SNPs were recoded, we generated the PRS as a weighted sum of SNP dosages, utilizing the effect sizes β from the published summary statistics. The raw PRS for an individual patient i with M SNPs was computed as

$$PRS_i = \beta_1 G_{i1} + \beta_2 G_{i2} + \cdots + \beta_M G_{iM}.$$

PRS calculations were performed using R, with genotype data input via the genio package[56]. PRS SNPs and weights for lung cancer and difficulty quitting are provided in **Supplementary Tables 2-3**, respectively.

Principal components analysis of the 1000 Genomes Project

To ensure our PRS can be applied universally across ancestries, we conducted ancestry adjustment using the 1000G dataset, which provides a representative cross-section of the five major global superpopulations: AFR, AMR, EAS, EUR and SAS[18]. Principal components analysis (PCA) is popular tool in ancestry inference, as it can capture continental genetic diversity and provide a continuous, label-free quantification of genetic ancestry[16,17,57]. We performed PCA on all 3,202 1000G samples, using 55,248 SNPs that are shared among the recommended SNPs set by gnomAD[49], 1000G reference data, UKBB and GISC validation data, and the 23andMe genotyping array used for the trial (Error! Reference source not found.). PCA was performed using plink 2.0[58] with the f

ollowing command to generate the top five PCs: "--pca allele-wts 5". This process resulted in a set of loadings or weights for each of the 55,248 SNPs corresponding to each principal component. We then applied these loadings to genotype data from 1000G, UKBB, and GISC within plink 2.0, employing the command: "--score [i] [j] header cols=+scoresums,-scoreavgs,-dosagesum,-nallele --score-col-nums [k1]-[k2]" to generate the PC scores.

Standardizing PRS distributions across the continuum of genetic ancestry

We standardized the PRS distributions for lung cancer and difficulty quitting using data from the 1000G dataset, employing a regression-based method to adjust for distributional differences across ancestries[11,19,59]. This adjustment process involves two key steps:

First, we conducted a linear regression of the raw PRS against the top five PCs derived from the PCA, such that the PRS of individual i is a linear model of their PCs with random noise:

$$PRS_i = \alpha_0 + \alpha_1 PC_{i1} + \alpha_2 PC_{i2} + \cdots + \alpha_5 PC_{i5} + \epsilon_i^{mean}.$$

We obtained an estimated intercept $\widehat{\alpha_0}$, and weights $(\widehat{\alpha_1}, \cdots, \widehat{\alpha_5})$ for each PC. The residuals of the raw PRS, accounting for the linear effects of the PCs, were calculated as:

$$R_i = PRS_i - \widehat{\alpha_1} PC_{i1} - \cdots - \widehat{\alpha_5} PC_{i5}.$$

This first step is designed to remove mean differences in the PRS distribution across ancestries. Subsequently, we used the square residuals R_i^2 as a measure for PRS variance, and ran a secondary linear regression model with:

$$R_i^2 = \gamma_0 + \gamma_1 PC_{i1} + \gamma_2 PC_{i2} + \cdots + \gamma_5 PC_{i5} + \epsilon_i^{var}.$$

From this, we derived a second estimated intercept ($\widehat{\gamma_0}$) and a new set of weights $(\widehat{\gamma_1}, \cdots, \widehat{\gamma_5})$ for each PC's effect on the variance of the PRS. The final ancestry-adjusted PRS for each individual i was then computed as:

$$PRS_i^{cal} = \frac{PRS_i - \widehat{\alpha_0} - \widehat{\alpha_1} PC_{i1} - \cdots - \widehat{\alpha_5} PC_{i5}}{\sqrt{\widehat{\gamma_0} + \widehat{\gamma_1} PC_{i1} + \widehat{\gamma_2} PC_{i2} + \cdots + \widehat{\gamma_5} PC_{i5}}}$$

By scaling the residuals with the fitted values from the second regression, we standardized the variance of the PRS distribution across ancestries to mean 0 and variance 1.

We validated this ancestry adjustment procedure in the UKBB and GISC datasets by applying the PC coefficients $(\widehat{\alpha_1}, \cdots, \widehat{\alpha_5}, \widehat{\gamma_1}, \cdots, \widehat{\gamma_5})$ to the raw PRS. This ancestry-adjusted PRS accurately reflects an individual's genetic risk independent of their ancestry, facilitating a unified risk stratification methodology. This is especially crucial for individuals with admixed or unknown ancestry, for whom discrete ancestry-specific prediction models may be unsuitable or invalid[4,16,36,60,61].

Risk categories determination

To stratify patients by genetic risk for lung cancer and difficulty quitting, we calculated odds ratios (OR). These ratios compare the probability of an outcome occurring in individuals within a percentile range p (i.e. 80-100%) of the ancestry-adjusted PRS distribution with the probability of the same outcome occurring in individuals within another percentile range q (i.e. 0-20%).

$$OR_{pq} = \frac{P(Y = 1|PRS \in p)/P(Y = 0|PRS \in p)}{P(Y = 1|PRS \in q)/P(Y = 0|PRS \in q)}$$

Following the determination of the desired OR for each health outcome, we established cut points within the PRS distribution to categorize individuals into three distinct risk groups: “at risk”, “high risk”, and “very high risk”.

We chose clinically meaningful thresholds to define three risk categories – “at risk”, “high risk”, and “very high risk” – in communicating personalized risks and benefits with patients in our research. We use these category names because all participants in these two ongoing trials are high-risk patients eligible for lung cancer screening with active heavy smoking. For lung cancer, we categorize patients by the bottom 20%, middle 60%, and top 20%. For difficulty quitting, we divide the PRS distribution into thirds – using the bottom, middle, and top 33%. Since difficulty quitting is a behavior trait with no established absolute risk rates like cancer, we use these percentiles to provide slightly more agnostic risk information.

For our ancestry-adjusted PRS, we use the distribution among all 1000G samples to set percentile ranges and evaluate corresponding odds ratios among UKBB participants. For comparison, we also evaluate odds ratios using ancestry-matched raw PRS distributions, i.e. European-only 1000G PRS distribution for self-reported European UKBB participants. For UKBB participants with “Other” ethnic background, we use the raw PRS distribution among all 1000G samples, rather than matching to a specific group.