Supplementary Methods

1. Study population

Participants in the Seoul National University Bundang Hospital (SNUBH) cohort underwent standard esophagogastroduodenoscopy at SNUBH from 2006 to 2018. The GENIE cohort followed participants who underwent health check-ups at the Seoul National University Hospital Healthcare System Gangnam Center from 2003 to 2017. All participants provided informed consent and completed the questionnaire under the supervision of well-trained interviewers. The questionnaire collected data on age, sex, smoking status, alcohol consumption, salt preference, family history of gastric cancer (GC), and history of *Helicobacter pylori* eradication therapy. *H. pylori* infection status was determined comprehensively using a rapid urease test (CLOtest, Delta West) or a ^{13}C urea breath test (UBiTkit, Otsuka Pharmaceutical Co. Ltd.) or *H. pylori* serology (serum anti-*H. pylori* IgG antibody) testing [1-3].

The test set was derived from the Korean Genome and Epidemiology Study (KoGES) [4], a consortium of three population-based cohorts: the Ansan and Ansung, the Cardiovascular Disease Association Study, and the Health Examinee cohorts. The study participants were aged ≥ 40 years at the baseline and answered questionnaires during the periodic follow-up assessments. Further details on the study are described elsewhere [4].

2. Genotyping, quality control and imputation

DNA was extracted from the blood samples of all participants. A total of 10,187 participants from GENIE and 72,297 from KoGES were genotyped using the Korea Biobank Array (KoreanChip) [5] version 1.0. A total of 1,235 participants from the SNUBH were genotyped using KoreanChip version 1.1. Each variant was called using K-medoids to minimize batch heterogeneity and quality control (QC) was subsequently performed for the genotypes in each cluster [6]. Variants were excluded if their missing rate was > 0.05 or the *p*-value for the Hardy-Weinberg equilibrium (HWE) test was $< 1\times10^{-5}$. Participants with a missing genotype rate > 0.05, a single nucleotide polymorphism (SNP) heterozygosity rate greater than three standard deviations from the mean of all participants, or $0.2 < X$ chromosome homozygosity < 0.8 were also excluded. After merging the clusters, the QC steps were repeated. In addition, variants that differed significantly among clusters for HWE, missing rate, or minor allele frequency (MAF) at $\alpha=10^{-5}$ were further excluded to increase the quality of the data. After QC, imputation was performed using the Northeast Asian Reference Database imputation server with a reference

panel of 1,779 Northeast Asians [7]. Eagle v2.4 [8] and Minimac4 [9] were used for pre-phasing and imputation, respectively. After imputation, imputed SNPs with $r^2 < 0.3$ and multiallelic SNPs were removed. SNUBH and GENIE data were merged into a validation dataset, and SNPs with a missing rate > 0.05, HWE p-value < 1×10^{-5}, or MAF < 0.01 were excluded. Additionally, SNPs without a reference SNP identifier (rsID) or those not found in the genome-wide association study (GWAS) summary statistics of the discovery set were excluded. Participants whose SNP heterozygosity was > 5×the interquartile range (IQR) or whose principal component (PC) score was outside 5×IQR were excluded. The final validation set was comprised of 10,325 participants. To calculate the polygenic risk score (PRS) for GC (PRS-GC), 3,573,357 SNPs were identified. Similarly, for the PRS for alcohol consumption (PRS-Alcohol), 3,374,225 SNPs were accessible. The same QC procedures was used for the test set. The final test set included 71,519 participants, and the PRS-GC and PRS-Alcohol were calculated using 3,414,662 and 3,076,703 SNPs, respectively. All genetic data were analyzed using PLINK [10], ONETOOL [11] and R (v3.6.3).

3. Gastric cancer identification

All GC cases in the SNUBH cohort were confirmed histologically following surgery or endoscopic therapy. Participants with endoscopy results that showed no evidence of GC, dysplasia, mucosa-associated lymphoid tissue lymphoma, or esophageal cancer were included in the control group. In both GENIE and KoGES cohort, GC cases were identified based on self-reported GC history during follow-up, and participants who did not provide information on GC history were excluded. Any controls with a previous cancer history were excluded from our study. There were 563 GC cases and 8,930 controls in the validation set and 680 GC cases and 67,459 controls in the test set.

4. Assessment of alcohol consumption

The participants responded to a study-specific questionnaire about their alcohol consumption habits. This included questions about drinking status and quantity and frequency of alcohol consumption among current drinkers. The SNUBH questionnaire asked about drinking frequency (nondrinker, < 1 time/mo, 2-3 times/mo, 1 time/wk, 2-4 times/wk or daily) and the quantity of alcohol consumed per session (< 2, 2, 4, 8, or 16 standard units), with one unit equivalent to 14g of ethanol [3]. Similarly, the GENIE questionnaire asked about drinking frequency (< 1 time/mo, 2-4 times/mo, 2 times/wk, 3 times/wk, 4 times/wk, 5 times/week, 6 times/wk, or daily) and the quantity of alcohol consumed (1-2, 3-4, 5-6, 7-9 or ≥ 10 glasses of

alcoholic beverage). In South Korea, one glass of an alcoholic beverage is generally considered to contain approximately 8 g of ethanol [12]. Alcohol consumption was calculated by multiplying the frequency (times/day) with the quantity (glasses/time) and ethanol content (8 g/glass). To account for heterogeneity in the survey questions within the validation set, we standardized alcohol consumption within each cohort for subsequent analyses. The KoGES questionnaire asked about the frequency (nondrinker, 1 time/mo, 2-3 times/mo, 1 time/wk, 2-3 times/wk, 4-6 times/wk, or ≥1 time/day) and the amount (free-text) of Korean traditional alcohols consumed, including Soju (ethanol content, 22%), Cheongju (15%), and Makgeolli (6%), beer (4.5%), wine (13%) and liquor (40%). Alcohol consumption (ethanol, g/day) was calculated by multiplying the frequency (times/day) by the amount (mL/time), ethanol content (%), and ethanol density (0.7893 g/mL).

Each of the three Korean cohorts (SNUBH, GENIE, and KoGES) had a different study design. SNUBH was a case-control study with a prospective cohort, whereas GENIE and KoGES were longitudinal studies. For the SNUBH cohort, alcohol consumption was calculated during the first hospital visit. For the GENIE and KoGES cohorts, alcohol consumptions were measured at multiple time points, and their mean values were calculated. It is commonly observed that patients diagnosed with GC stop consuming alcohol. Thus, alcohol consumption was measured before the onset of GC. For GENIE and KoGES cohorts, we calculated alcohol consumption by averaging their consumption prior to the GC diagnosis. For the validation set, alcohol consumption was scaled within each cohort and subsequently rescaled to achieve a total mean and SD. The participants were also classified into two categories: non-to-light drinker (< 10 g/day for females and < 20 g/day for males), or heavy drinker (≥ 10 g/day for females and ≥ 20 g/day for males).

After excluding participants with missing alcohol consumption information, 9,004 and 71,508 participants remained in the validation and test set, respectively. Patients with GC who did not report alcohol consumption before their cancer diagnosis were excluded from the study. As a result, 8,846 participants (531 cases and 8,315 controls) remained in the validation set, and 67,711 participants (313 cases and 67,458 controls) remained in the test set for GC analyses.

5. Polygenic risk score calculation and evaluation

The construction of the PRS required summary statistics from the GWAS and those for GC and alcohol consumption were downloaded from the BBJ [13,14] (http://jenger.riken.jp/en). Genetic correlations were estimated using LDSC method with the

default option [15,16]. To ascertain the appropriateness of employing summary statistics derived from Japanese GWAS for Korean, we estimated the SNP-based heritability using GWAS results from the BBJ and KoGES [17], in addition to the genetic correlations between the two summary statistics.

We developed PRS-GC and PRS-Alcohol using data from GWAS of GC, conducted by Ishigaki et al. in 2020 [13], which analyzed 6,563 cases (4,885 males; 1,678 females) and 195,745 controls (97,655 males; 98,090 females) across 8,678,256 variants and the GWAS of alcohol consumption (drinks per week), conducted by Matoba et al. in 2020 [14], which analyzed 58,610 Japanese individual (44,232 males; 14,378 females) across 5,961,480 variants. We checked the MAFs of the SNPs and focused on SNPs with a MAF > 0.005.

Four different methods were used to calculate PRSs: CT [18], lassosum [19], LDpred [20], and PRS-CS [21], following the protocols suggested by Choi et al. [22]. Hyperparameters were set for each model. For CT, the p-value threshold of 1×10^{-5}, and the r^2 threshold was 0.5. For lassosum, a reference LD block for the Asian population, lambdas in sequence of 20 numbers in logarithmic scale between 0.001 and 0.1, and a shrinkage parameter of 0.2, 0.5, 0.9 and 1 (default). For LDpred, ρ values of 0.03, 0.01, 0.3, 0.1 and 1 were used. For PRS-CS, 1000 Genome Project phase 3 East Asian reference panels were used, as well as a global shrinkage parameter of 0.01, and scale and shape parameters of gamma-gamma prior distributions of 1 and 0.5, respectively (default).

Two types of PRSs were derived from the participants in the validation set. The first was computed for both sexes, while the second was derived separately for each sex. These were standardized and combined. Using the validation set, we fitted the logistic regression of PRS-GC on GC and the linear regression of PRS-Alcohol on alcohol consumption (g/day) after adjusting for age, sex, and the top four PCs. The best PRSs were selected based on the 5-fold cross-validated area under the curve (AUC) for PRS-GC, and r^2 and Pearson's correlation coefficients for PRS-Alcohol. The SNPs used to calculate the best PRSs were mapped to the genome using the Variant Effect Prediction tool [23] and genes in which the SNPs were located within 5kb. Pearson's correlation coefficients between the best PRS-GC and PRS-Alcohol were estimated. The best PRS on the test set, as a weighted sum of risk alleles, was then calculated using the weights obtained from the best PRSs, which were estimated in the validation set.

6. Statistical analysis

Clinical characteristics of participants were compared according to GC cancer status. Continuous variables were summarized as mean and standard deviation, and categorical variables as counts and percentages. Comparison between cancer status was tested using Wilcoxon signed rank test for continuous variables and chi-squared test for categorical variables.

We performed multinomial logistic regression to investigate the association between specific types of GC and PRSs, utilizing a validation set. The covariates incorporated in our study were age, sex, BMI, smoking status, salt preference, family history of GC, alcohol consumption and H. pylori infection. Based on these analyses, we created a prediction model and DeLong's test was used to compare the AUCs of the models, with and without PRS-GC and PRS-Alcohol.

In the test set, we examined the association of PRS-GC and PRS-Alcohol (both in continuous value and quartiles) with each phenotype. We adjusted for age, sex and the top four PCs, while employing logistic or linear regression. Subsequently, we compared models containing different sets of covariates. Model 1 contained PRS-GC and PRS-Alcohol. Model 2 contained conventional risk factors (age, sex, body mass index [BMI], smoking, salted food intake, family history of GC and alcohol consumption). Model 3 contained the parameters in Model 2, PRS-GC, and PRS-Alcohol. Model 4 contained the parameters in Model 3, the interaction terms of alcohol consumption with PRS-GC and PRS-Alcohol. For model 4, we visualized the interaction effects of PRS using the R package 'interactions' [24]. We plotted alcohol consumption and the log odds of GC comparing three levels of PRSs: mean +/− 1 × standard deviation from the mean. Continuous alcohol consumption and drinking status (non-to-low drinker versus moderate-to-heavy drinker) were examined.

To investigate the relationship between PRS-GC, PRS-Alcohol, and alcohol consumption, we conducted a mediation analysis using the Medflex R package (v.0.6-10) [25]. Data expansion was conducted using the imputation approach and bootstrapping with 1,000 iterations was conducted. We assumed that alcohol consumption acted as a mediator in the effect of PRS-Alcohol on GC with an interaction effect involving PRS-Alcohol. Age, sex, BMI, smoking status, salted food intake, and family history of GC were considered as covariates. Moreover, effect modification by the PRS-GC was assumed, and the interaction terms of binary PRS-Alcohol (cutoff at the 50th percentiles), alcohol consumption, and binary PRS-GC (cutoff

at the 70th percentiles) were included as covariates. The total effect was decomposed into three components, as suggested by Van der Weele: (1) a pure direct effect (2) a pure indirect effect, and (3) a mediated interactive effect [26]. The summation of the pure direct/indirect and medicated interactive effect indicates the overall direct/indirect effect.

Survival analysis of the association between GC incidence and drinking status was conducted using the Cox proportional hazards model, adjusting for the same covariates included in the logistic regression model. The proportionality assumption was verified, and no violations were observed. We dichotomized the samples into high and low PRS-GC (cutoff at the 70th percentiles) and high and low PRS-Alcohol (cutoff at the 50th percentiles) groups. The samples were then categorized into four groups based on the combination of low-high PRS-GC and PRS-Alcohol. Furthermore, subgroup analyses were conducted, and p for interactions were estimated using the log-likelihood ratio test.

Finally, we conducted subgroup analyses with the participants stratified by sex. All statistical analyses were performed using R (v3.6.3) and Rex (v3.6.1) [27] and statistical significance was set at a p-value of 0.05.

References

1. Lee C, Choe EK, Choi JM, Hwang Y, Lee Y, Park B, et al. Health and Prevention Enhancement (H-PEACE): a retrospective, population-based cohort study conducted at the Seoul National University Hospital Gangnam Center, Korea. BMJ Open. 2018;8:e019327.

2. Shin CM, Park K, Kim N, Won S, Ohn JH, Lee S, et al. rs2671655 single nucleotide polymorphism modulates the risk for gastric cancer in Helicobacter pylori-infected individuals: a genome-wide association study in the Korean population. Gastric Cancer. 2022;25:573-85.

3. Choi YJ, Kim N, Jang W, Seo B, Oh S, Shin CM, et al. Familial clustering of gastric cancer: a retrospective study based on the number of first-degree relatives. Medicine (Baltimore). 2016;95:e3606.

4. Kim Y, Han BG; KoGES Group. Cohort profile: the Korean Genome and Epidemiology Study (KoGES) consortium. Int J Epidemiol. 2017;46:e20.

5. Moon S, Kim YJ, Han S, Hwang MY, Shin DM, Park MY, et al. The Korea biobank array: design and identification of coding variants associated with blood biochemical traits. Sci Rep. 2019;9:1382.

6. Seo S, Park K, Lee JJ, Choi KY, Lee KH, Won S. SNP genotype calling and quality control for multi-batch-based studies. Genes Genomics. 2019;41:927-39.

7. Yoo SK, Kim CU, Kim HL, Kim S, Shin JY, Kim N, et al. NARD: whole-genome reference panel of 1779 Northeast Asians improves imputation accuracy of rare and low-frequency variants. Genome Med. 2019;11:64.
8. Loh PR, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016;48:1443-8.
9. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48:1284-7.
10. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559-75.
11. Song YE, Lee S, Park K, Elston RC, Yang HJ, Won S. ONETOOL for the analysis of family-based big data. Bioinformatics. 2018;34:2851-3.
12. Kim YG, Han KD, Choi JI, Boo KY, Kim DY, Lee KN, et al. Frequent drinking is a more important risk factor for new-onset atrial fibrillation than binge drinking: a nationwide population-based study. Europace. 2019;22:216-24.
13. Ishigaki K, Akiyama M, Kanai M, Takahashi A, Kawakami E, Sugishita H, et al. Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. Nat Genet. 2020;52:669-.
14. Matoba N, Akiyama M, Ishigaki K, Kanai M, Takahashi A, Momozawa Y, et al. GWAS of 165,084 Japanese individuals identified nine loci associated with dietary habits. Nat Hum Behav. 2020;4:308-16.
15. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015;47:1236-41.
16. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J; Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47:291-5.
17. Nam K, Kim J, Lee S. Genome-wide study on 72,298 individuals in Korean biobank data for 76 traits. Cell Genom. 2022;2:100189.
18. Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. Bioinformatics. 2015;31:1466-8.

19. Mak TS, Porsch RM, Choi SW, Zhou X, Sham PC. Polygenic scores via penalized regression on summary statistics. Genet Epidemiol. 2017;41:469-80.
20. Vilhjalmsson BJ, Yang J, Finucane HK, Gusev A, Lindstrom S, Ripke S, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am J Hum Genet. 2015;97:576-92.
21. Ge T, Chen CY, Ni Y, Feng YA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun. 2019;10:1776.
22. Choi SW, Mak TS, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc. 2020;15:2759-72.
23. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl variant effect predictor. Genome Biol. 2016;17:122.
24. Long JA. Interactions: comprehensive, user-friendly toolkit for probing interactions. R package version 1.1.0. R Foundation for Statistical Computing; 2019.
25. Steen J, Loeys T, Moerkerke B, Vansteelandt S. medflex: an R package for flexible mediation analysis using natural effect models. J Stat Softw. 2017;76:1-46.
26. VanderWeele TJ. A three-way decomposition of a total effect into direct, indirect, and interactive effects. Epidemiology. 2013;24:224-32.
27. Lee B, An J, Lee S, Won S. Rex: R-linked EXcel add-in for statistical analysis of medical and bioinformatics data. Genes Genomics. 2023;45:295-305.