A Phenome-Wide Mendelian Randomization Study of Pancreatic Cancer Using Summary Genetic Data

Ryan J. Langdon1,2, Rebecca C. Richmond1,2, Gibran Hemani1,2, Jie Zheng1,2, Kaitlin H. Wade1,2, Robert Carreras-Torres3, Mattias Johansson4, Paul Brennan4, Robyn E. Wootton1,2,5, Marcus R. Munafò1,2,5, George Davey Smith1,2, Caroline L. Relton1,2, Emma E. Vincent1,2, Richard M. Martin1,2, and Philip Haycock1,2

Abstract

Background: The 5-year mortality rate for pancreatic cancer is among the highest of all cancers. Greater understanding of underlying causes could inform population-wide intervention strategies for prevention. Summary genetic data from genome-wide association studies (GWAS) have become available for thousands of phenotypes. These data can be exploited in Mendelian randomization (MR) phenotype-wide association studies (PheWAS) to efficiently screen the phenotype for potential determinants of disease risk.

Methods: We conducted an MR-PheWAS of pancreatic cancer using 486 phenotypes, proxied by 9,124 genetic variants, and summary genetic data from a GWAS of pancreatic cancer (7,110 cancer cases, 7,264 controls). ORs and 95% confidence intervals per 1 SD increase in each phenotype were generated.

Results: We found evidence that previously reported risk factors of body mass index (BMI; 1.46; 1.20–1.78) and hip circumference (1.42; 1.21–1.67) were associated with pancreatic cancer. We also found evidence of novel associations with metabolites that have not previously been implicated in pancreatic cancer: ADP5SGEGDFXAEGGGVR, a fibrinogen-cleavage peptide (1.60; 1.31–1.95), and O-sulfo-L-tyrosine (0.58; 0.46–0.74). An inverse association was also observed with lung adenocarcinoma (0.63; 0.54–0.74).

Conclusions: Markers of adiposity (BMI and hip circumference) are potential intervention targets for pancreatic cancer prevention. Further clarification of the causal relevance of the fibrinogen-cleavage peptides and O-sulfo-L-tyrosine in pancreatic cancer etiology is required, as is the basis of our observed association with lung adenocarcinoma.

Impact: For pancreatic cancer, MR-PheWAS can augment existing risk factor knowledge and generate novel hypotheses to investigate.

Introduction

People diagnosed with pancreatic cancer have a very poor prognosis, with a less than 5% 5-year survival rate (http://globocan.iarc.fr); symptoms do not manifest until the cancer is at an advanced stage and the disease is rarely detected early. Greater understanding of the etiology of pancreatic cancer could reduce its burden by informing whole-population or risk-stratified prevention strategies.

Risk factors previously reported for pancreatic cancer include cigarette smoking (1), type II diabetes (2), adiposity (3), and chronic pancreatitis (4). However, these reports are based on observational epidemiologic studies, which are prone to unmeasured or residual confounding and reverse causation, precluding robust causal inference. Furthermore, conventional epidemiologic studies often test a narrow set of hypotheses using prior subject knowledge, typically based on other observational studies. While essential, these approaches can constrict a field of research, and preoccupation with previously hypothesized risk factors can prevent both the identification of novel risk factors and prioritization of causal associations (5).

Mendelian randomization (MR) is a well-established type of instrumental variable (IV) analysis that addresses some of the shortcomings of conventional observational studies by using genetic anchors to appraise the causal relevance of exposures in disease (6). It is an increasingly recognized and powerful tool for identifying causes of a broad spectrum of outcomes, including cancer (7, 8). Two-sample MR uses summary-level data from published genome-wide association studies (GWAS) to allow causal appraisal of hypothesized exposure–outcome associations using gene–exposure and gene–outcome associations collected in separate studies (9–11). This method can be extended to appraise...
causality in a hypothesis-free manner, appraising 1-to-many, many-to-1, or many-to-many exposure–outcome combinations, in an approach known as a MR phenotype-wide association study (MR-PheWAS; refs. 12, 13).

Here, we used MR-PheWAS to screen the phenotype for potential causes of pancreatic cancer. Our aims were twofold: to identify potentially novel causes of pancreatic cancer that may not have been captured using previous epidemiologic approaches, and to prioritize hypotheses identified in current literature.

Materials and Methods

Data preparation

Genetic instruments for phenotypes. Two-sample MR was conducted using the TwoSampleMR R package (14). Genetic data on cognitive, anthropometric, metabolic, immune, and behavioral phenotypes were obtained from the MR-Base database of harmonized GWAS summary data (Supplementary Fig. S1). All phenotypes possessing robust genetic proxies (defined as \( P < 5 \times 10^{-8} \)) with which to conduct MR analyses were considered for further analysis (\( N = 523 \)). Duplicate (\( N = 17 \)) and non-European studies (\( N = 8 \)) were excluded from the analysis at this stage, leaving 498 potential phenotypes for analysis. Genetic instruments for each phenotype were single-nucleotide polymorphisms (SNP) independently associated with the phenotype of interest after linkage disequilibrium (LD) clumping (window = 10,000 kb; \( r^2 = 0.1 \)). For each identified SNP, the reported effect size was expressed as a one SD increase in the level of the phenotype per risk allele, along with the SE. In the case of a binary phenotype (e.g., presence or absence of coronary heart disease), the reported effect size was expressed as a log-OR. The single largest or most recent summary GWAS data were used per phenotype, systematically prioritized by the instrument extraction function (extract_instruments) of the TwoSampleMR R package and preventing bias from sample overlap from multiple GWAS for exposure phenotypes. For each genetic variant associated with the identified phenotypes, effect-estimates and SEs were extracted from the summary genetic data for pancreatic cancer.

To harmonize the data, effect alleles in the pancreatic cancer summary data were coded to reflect the phenotype-increasing allele; using allele frequencies to resolve strand ambiguities for palindromic SNPs (A/T or C/G). Those phenotypes that did not have genetic variants in the pancreatic cancer GWAS were excluded (\( N = 12 \)), resulting in a final list of 486 phenotypes on which to perform MR analyses (15–48). These phenotypes are tabulated in Supplementary Table S1, which details the phenotype type name, the corresponding author or contributing consortium, the sample size of the contributing GWAS, the number of SNPs in the GWAS, and the original units of each phenotype.

Pancreatic cancer data. GWAS data from people of European descent with pancreatic cancer and matched controls were obtained from the PanScan (12 studies) and PanC4 (10 studies) consortia through the National Centre for Biotechnology Information (NCBI) Database of Genotypes and Phenotypes (dbGaP; ref. 49; Study Accession: phs000206.v3.p2 and phs000648.v1.p1; project reference #9314). PanScan and PanC4 were initially published in three releases: PanScan I (1,788 cases and 1,769 controls), PanScan II (1,696 cases and 1,563 controls), and PanC4 (3,626 cases and 3,932 controls; refs. 50–52). The samples were originally genotyped using Illumina HumanHap550 (PanScan I), Human610-Quad (PanScan II), and HumanOmniExpressExome-8v1 (PanC4) arrays. A summary of the characteristics of the consortia and contributing studies is provided in Supplementary Table S2A and S2B.

Initial quality control steps and analyses were performed within each publication set at the International Agency for the Research of Cancer (IARC), Lyon. After removing duplicates, related samples, samples with sex discrepancy and population outliers, 7,110 cases and 7,264 controls remained across the three combined consortia. Genotype imputation was performed using the Michigan Imputation Server (53). Genotypes were prephased using SHAPEIT v2 (54) and imputed with Minimach v3 (55) using the Haplotype Reference Consortium panel (56). After imputation, SNPs with an imputation quality (\( R^2 \)) lower than 0.7 were removed from the datasets. Effect-estimates for pancreatic cancer risk were obtained after adjusting for age, sex, and principal components for population stratification using R software (R version 3.3.1). Results from each PanScan release were then combined using a fixed-effects inverse-variance approach implemented in METAL (57). Finally, outcome data were converted from a “chromosome: position” format to reference SNP cluster ID (rsID), using the “biomaRt” R package (58) with human genome build 19 (hg19) as reference, to generate SNP IDs that were in the format expected by the TwoSampleMR R package.

Power calculations

Low power can be a limitation of MR because genetic polymorphisms typically explain a small amount of phenotypic variance. We calculated power for this analysis based on a sample size of 14,374 (7,110 cases, 7,264 controls) across a range of predefined phenotypic variances and effect sizes. The median variance explained by SNP IVs for our 486 phenotypes was 3.3%. At this variance, our power calculations indicated we had 80% power to detect a minimum OR of 1.52 (beta of 0.42), at an alpha of 1.6 \( \times 10^{-4} \) (0.05/312 independent tests).

Mendelian randomization analyses

We used maximum likelihood (59) and multiplicative random effects inverse-variance weighted (MRE IVW; refs. 60, 61) MR analyses when the number of SNP’s instrumenting a phenotype was greater than 1. Both have been proposed for MR analyses when using summary genetic data with phenotype instruments containing multiple SNPs (62). An MRE model allows for heterogeneity between the causal estimates targeted by the genetic variants by allowing overdispersion in the regression model. Underdispersion is not permitted (in case of underdispersion, the residual SE is set to 1, as in a fixed-effect analysis). For phenotypes instrumented by a single SNP, we derived Wald ratio effect-estimates (62, 63). Results were expressed ORs with a corresponding 95% confidence interval (CI) per 1 SD increase in continuous traits (e.g., height), and as ORs with 95% CIs per increase in log odds for binary traits (e.g., type II diabetes; ref. 64).

To correct for multiple testing, the correlation structure among the analyzed phenotypes was estimated using PhenopD (65), which implements principal component analysis to identify independent variables using GWAS summary–level statistics. Firstly, a correlation matrix of phenotypes was built using metaCCA (66), estimating Pearson pair-wise correlations.
between the GWAS summary data for each phenotype. Once the correlation matrix was built, the effective number of independent phenotypes was estimated using matrix spectral decomposition (67, 68). PhenoSpD overestimates the number of independent variables as it treats phenotypes from separate studies as entirely independent when it is likely they are not. Therefore, our Bonferroni correction for multiple testing is likely particularly conservative.

### Sensitivity analyses

MR-Egger regression (69) was used as a sensitivity analysis to detect bias due to horizontal pleiotropy in the causal estimates. Horizontal pleiotropy is where a genetic variant affects the outcome via a different biological pathway from the phenotype under investigation and is a violation of a key assumption of MR (see Supplementary Fig. S2). MR-Egger regression performs a weighted linear regression of the effect-estimates. Horizontal pleiotropy is where a genetic variant affects the outcome via a different biological pathway from the phenotype under investigation and is a violation of a key assumption of MR (see Supplementary Fig. S2). MR-Egger regression performs a weighted linear regression of the effect-estimates. Horizontal pleiotropy is where a genetic variant affects the outcome via a different biological pathway from the phenotype under investigation and is a violation of a key assumption of MR (see Supplementary Fig. S2). MR-Egger regression performs a weighted linear regression of the effect-estimates. Horizontal pleiotropy (69). Deviation from the origin in an MR-Egger analysis where phenotypes were proxied by 3 or fewer SNPs. In addition, we assessed evidence of heterogeneity between SNPs (another potential indication of horizontal pleiotropy and other violations of MR assumptions) for the causal effect-estimates of the phenotype on pancreatic cancer using forest plots and Cochran Q test. Finally, we investigated whether the MR-Egger regression intercept tests from MR-Egger regression did not, however, meet the conservative Bonferroni correction for multiple testing (67, 68). PhenoSpD overestimates the number of independent variables as it treats phenotypes from separate studies as entirely independent when it is likely they are not. Therefore, our Bonferroni correction for multiple testing is likely particularly conservative.

### Results

Using PhenoSpD, we estimated that the 486 phenotypes we investigated corresponded to 312 independent tests (65). To aid interpretation of our MR analyses, we set a P value threshold of $1.6 	imes 10^{-4}$ (0.05/312) to suggest evidence of association and to prioritize phenotypes for follow-up analyses. Five phenotypes were associated with pancreatic cancer at this threshold (Fig. 1; Table 1). The results of the MR analyses for all phenotypes are shown in Supplementary Table S3. Of the 5 associations, 2 were inversely related to pancreatic cancer: lung adenocarcinoma [OR for pancreatic cancer (95% CI): 0.63 (0.54–0.74) per doubling in the odds of lung adenocarcinoma; P: 1.68e$^{-4}$] and the metabolite O-sulfo-L-tyrosine [0.58 (0.46–0.74) per SD increase; P: 2.45e$^{-4}$]. The other 3 phenotypes were positively related to pancreatic cancer [OR (95% CI/SD increase): ADpSGEGDFXAEGGGVR$^*$ (a fibrinogen cleavage peptide) 1.60 (1.31–1.95); P: 1.50e$^{-4}$; hip circumference [1.42 (1.21–1.67); P: 3.92e$^{-4}$]; and body mass index [BMI; 1.46 (1.20–1.78); P: 4.02e$^{-4}$]. Maximum likelihood effect-estimates were consistent with IVW estimates for these associations (Table 1).

There was clear evidence of heterogeneity in associations with pancreatic cancer among the individual SNPs used as IVs for body mass index (Q: 186.61; P: 0.01), hip circumference (Q: 25.3; P: 4.86e$^{-3}$), but not ADpSGEDFXAEGGGVG$^*$ (Q: 2.67; P: 0.10), lung adenocarcinoma (Q: 0.43; P: 0.51), O-sulfo-L-tyrosine (Q: 0.13; P: 0.72), or BMI (Q: 0.00; P: 0.95; Fig. 3).

There was clear evidence of heterogeneity in associations with pancreatic cancer among the individual SNPs used as IVs for body mass index (Q: 186.61; P: 0.01), hip circumference (Q: 25.3; P: 4.86e$^{-3}$), but not ADpSGEDFXAEGGGVG$^*$ (Q: 2.67; P: 0.10), lung adenocarcinoma (Q: 0.43; P: 0.51), O-sulfo-L-tyrosine (Q: 0.13; P: 0.72), or BMI (Q: 0.00; P: 0.95; Fig. 3).

There was clear evidence of heterogeneity in associations with pancreatic cancer among the individual SNPs used as IVs for body mass index (Q: 186.61; P: 0.01), hip circumference (Q: 25.3; P: 4.86e$^{-3}$), but not ADpSGEDFXAEGGGVG$^*$ (Q: 2.67; P: 0.10), lung adenocarcinoma (Q: 0.43; P: 0.51), O-sulfo-L-tyrosine (Q: 0.13; P: 0.72), or BMI (Q: 0.00; P: 0.95; Fig. 3).

There was clear evidence of heterogeneity in associations with pancreatic cancer among the individual SNPs used as IVs for body mass index (Q: 186.61; P: 0.01), hip circumference (Q: 25.3; P: 4.86e$^{-3}$), but not ADpSGEDFXAEGGGVG$^*$ (Q: 2.67; P: 0.10), lung adenocarcinoma (Q: 0.43; P: 0.51), O-sulfo-L-tyrosine (Q: 0.13; P: 0.72), or BMI (Q: 0.00; P: 0.95; Fig. 3).

There was clear evidence of heterogeneity in associations with pancreatic cancer among the individual SNPs used as IVs for body mass index (Q: 186.61; P: 0.01), hip circumference (Q: 25.3; P: 4.86e$^{-3}$), but not ADpSGEDFXAEGGGVG$^*$ (Q: 2.67; P: 0.10), lung adenocarcinoma (Q: 0.43; P: 0.51), O-sulfo-L-tyrosine (Q: 0.13; P: 0.72), or BMI (Q: 0.00; P: 0.95; Fig. 3).

There was clear evidence of heterogeneity in associations with pancreatic cancer among the individual SNPs used as IVs for body mass index (Q: 186.61; P: 0.01), hip circumference (Q: 25.3; P: 4.86e$^{-3}$), but not ADpSGEDFXAEGGGVG$^*$ (Q: 2.67; P: 0.10), lung adenocarcinoma (Q: 0.43; P: 0.51), O-sulfo-L-tyrosine (Q: 0.13; P: 0.72), or BMI (Q: 0.00; P: 0.95; Fig. 3).

There was clear evidence of heterogeneity in associations with pancreatic cancer among the individual SNPs used as IVs for body mass index (Q: 186.61; P: 0.01), hip circumference (Q: 25.3; P: 4.86e$^{-3}$), but not ADpSGEDFXAEGGGVG$^*$ (Q: 2.67; P: 0.10), lung adenocarcinoma (Q: 0.43; P: 0.51), O-sulfo-L-tyrosine (Q: 0.13; P: 0.72), or BMI (Q: 0.00; P: 0.95; Fig. 3).

There was clear evidence of heterogeneity in associations with pancreatic cancer among the individual SNPs used as IVs for body mass index (Q: 186.61; P: 0.01), hip circumference (Q: 25.3; P: 4.86e$^{-3}$), but not ADpSGEDFXAEGGGVG$^*$ (Q: 2.67; P: 0.10), lung adenocarcinoma (Q: 0.43; P: 0.51), O-sulfo-L-tyrosine (Q: 0.13; P: 0.72), or BMI (Q: 0.00; P: 0.95; Fig. 3).
indicate strong evidence for bias from unbalanced pleiotropy for body mass index (OR: 1.00; 95% CI: 0.98–1.02; P: 0.84) and hip circumference (OR: 1.00; 95% CI: 0.98–1.03; P: 0.75). In addition, effect-estimates from MR-Egger regression for hip circumference (OR: 1.18; 95% CI: 0.54–2.50; P: 0.68) and body mass index (OR: 1.35; 95% CI: 0.71–2.51; P: 0.36) were broadly compatible with results based on the maximum likelihood and IVW methods, albeit with wide CIs (see Table 1). While an inverse association was seen for lung adenocarcinoma and pancreatic cancer, the intercept from MR-Egger regression was negative (OR: 0.83; 95% CI: 0.51–1.35; P: 0.52) and the slope was in the opposite direction to the effect observed in the main analysis (OR: 1.57; 95% CI: 0.20–11.17; P: 0.71).

ADpSGEGDFXAEGGGVR and O-sulfo-L-tyrosine were both instrumented by 2 SNPs; thus, MR-Egger could not be used to assess horizontal pleiotropy for these phenotypes. Associations for both metabolites appeared to be largely driven by rs651007 (a SNP found in the ABO blood group region). The evidence for a causal effect of ADpSGEGDFXAEGGGVR on pancreatic cancer was weaker for the second SNP (rs6151429) used to instrument ADpSGEGDFXAEGGGVR (OR: 1.04; 95% CI: 0.75–1.44; P:0.81). Similarly, the evidence for a causal effect of O-sulfo-L-tyrosine was weaker for the other SNP (rs6151429) used to instrument O-sulfo-L-tyrosine (OR: 0.84; 95% CI: 0.62–1.14; P:0.26).

For the BMI and hip circumference analyses, 17 SNPs were common IVs for both phenotypes (Supplementary Table S4). We repeated MR analysis of these phenotypes after removing common SNPs between the hip circumference and BMI IVs (71). We obtained OR estimates similar to our original estimates (BMI OR: 1.49, 95% CI: 1.17–1.88; hip circumference OR: 1.41, 95% CI: 1.18–1.69), suggesting these associations were independent of each other.

**Table 1.** MR-PheWAS results passing study multiple testing threshold

<table>
<thead>
<tr>
<th>Exposure</th>
<th># SNPs</th>
<th>ML OR</th>
<th>ML CI</th>
<th>P</th>
<th>P₀ₙ₅</th>
<th>R²</th>
<th>F</th>
<th>Power</th>
<th>IVW OR</th>
<th>IVW CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung adenocarcinoma</td>
<td>4</td>
<td>0.63</td>
<td>0.54–0.74</td>
<td>1.68e⁻⁰⁶</td>
<td>5.47e⁻⁰⁶</td>
<td>N/A</td>
<td>N/A</td>
<td>0.72</td>
<td>0.48–1.08</td>
<td></td>
</tr>
<tr>
<td>ADpSGEGDFXAEGGGVR</td>
<td>2</td>
<td>1.60</td>
<td>1.31–1.95</td>
<td>3.08e⁻⁰⁶</td>
<td>1.50e⁻⁰⁳</td>
<td>3.59%</td>
<td>71.6</td>
<td>94.0</td>
<td>1.59</td>
<td>0.85–2.97</td>
</tr>
<tr>
<td>O-sulfo-L-tyrosine</td>
<td>2</td>
<td>0.58</td>
<td>0.46–0.74</td>
<td>8.07e⁻⁰⁶</td>
<td>2.45e⁻⁰⁴</td>
<td>1.02%</td>
<td>37.9</td>
<td>33.0</td>
<td>0.58</td>
<td>0.24–1.59</td>
</tr>
<tr>
<td>Hip circumference</td>
<td>113</td>
<td>1.42</td>
<td>1.21–1.67</td>
<td>2.41e⁻⁰⁵</td>
<td>4.02e⁻⁰⁴</td>
<td>4.46%</td>
<td>76.1</td>
<td>48.0</td>
<td>1.34</td>
<td>1.05–1.70</td>
</tr>
<tr>
<td>Body mass index</td>
<td>109</td>
<td>1.46</td>
<td>1.20–1.78</td>
<td>1.25e⁻⁰⁴</td>
<td>1.00e⁻²³</td>
<td>2.98%</td>
<td>91.3</td>
<td>51.0</td>
<td>1.44</td>
<td>1.32–1.86</td>
</tr>
</tbody>
</table>

NOTE: Phenotypes passing multiple testing correction for the MR-PheWAS analysis. Maximum likelihood ORs, CIs, and P values are shown for each phenotype in addition to the number of SNPs used in the IV, a Q-test P value for SNP heterogeneity, the variance explained, power statistics, and the inverse-variance weighted OR and CI for each phenotype.

Abbreviations: ML, maximum likelihood; N/A, not available; P₀ₙ₅, P value of heterogeneity from Q test.
Of the most established observational phenotypes with pancreatic cancer (smoking, diabetes, chronic pancreatitis, and adiposity; refs. 72, 73), only pancreatitis could not be instrumented and only adiposity passed our $P$ value threshold for further evaluation. The ORs (95% CI) for pancreatic cancer per SD increase in cigarettes smoked per day was 1.27 (0.67–2.42; $P$: 0.46) and was 1.02 (0.95–1.10; $P$: 0.56) per doubling in the odds of type II diabetes (Supplementary Table S3).

**Discussion**

We undertook an MR-PheWAS of the association of 486 phenotypes with pancreatic cancer, including cognitive, anthropometric, metabolic, immune, and behavioral phenotypes. We provide evidence that 5 of the 486 phenotypes we tested were associated with pancreatic cancer: BMI; hip circumference; ADpSGEGDFXAEGGGVR* (a fibrinogen cleavage peptide); O-sulfo-L-tyrosine; and lung adenocarcinoma.

The association of higher BMI with risk of pancreatic cancer is similar to findings from conventional observational studies, including the IARC Handbook Working Group (73), who reference Genkinger and colleagues (3) as the largest meta-analysis of body fatness on pancreatic cancer (OR for highest BMI category vs. normal: 1.5, 95% CI: 1.2–1.8). Our results also agree with the BMI finding in a MR study using PanScan data by Carreras-Torres and colleagues (74). In addition, they did not change substantially in our sensitivity analyses, thus are compatible with a causal effect.

Hip circumference, while potentially reflecting the observational association of general adiposity with pancreatic cancer, has not been previously implicated as a specific risk factor. Despite evidence of heterogeneity in effect-estimates when we stratified our analyses by PanScan study and sex, the direction of effect of sex- and study-specific estimates for hip circumference were the same. Thus, only the magnitude of the positive effect is uncertain for hip circumference. The SNPs for hip circumference show little evidence of sex-specific effects in the original GWAS (75), but consistent with findings in observational studies (76), the observed heterogeneity in this study suggests the effect of hip circumference on pancreatic cancer is stronger in females than males. Alternatively, the observed heterogeneity could reflect differences in strength of association between the IV SNPs and hip circumference between males and females; a violation of two-sample MR assumptions, casting doubt on the reliability of this result.

To our knowledge, the two metabolites ADpSGEGDFXAEGGGVR* and O-sulfo-L-tyrosine have not previously been associated with pancreatic cancer. There was clear heterogeneity among the SNPs used as instruments for these metabolites, with the associations being largely attributable to a single SNP (rs651007). The other SNPs (rs601338 and rs6151429, instrumenting ADpSGEGDFXAEGGGVR* and O-sulfo-L-tyrosine, respectively) showed weaker evidence of an association with pancreatic cancer. This suggests that the observed association of these metabolites with pancreatic cancer could reflect horizontal pleiotropy, and that the effect of rs651007 on pancreatic cancer may be mediated by some other pathway. A lookup of rs651007 in the National Human Genome Research Institute-European Bioinformatics Institute (NHGRI-EBI) GWAS Catalog revealed it to be mapped to the ABO gene; a locus that has...
been shown to be significantly associated with risk of pancreatic cancer genetically (16) and observationally (77). The ABO locus is associated with the serum inflammatory markers of TNFα (78) and soluble intercellular adhesion molecule 1 (sICAM-1; ref. 79). Inflammation has been reported to play an important role in the initiation of pancreatic tumors (80); the ABO locus may therefore influence pancreatic cancer risk by affecting systemic inflammation, thus promoting pancreatic carcinogenesis. Alternatively, these metabolites may cause pancreatic cancer, but rs601338 and rs6151429 could be subject to negative pleiotropy or not truly be associated with the metabolites, biasing our results toward the null. The limited availability of SNPs that could be used as instruments for ADPSGEGFXAEGGGVGR and O-sulfo-L-tyrosine constrained our ability to conduct sensitivity analyses to investigate these further.

Our results suggest evidence of an association between pancreatic cancer and genetic liability to lung adenocarcinoma. A potential explanation for this finding is sample overlap between our exposure and outcome, as we cannot reject the possibility that the PanScan control population contained individuals who were lung adenocarcinoma cases. However, Wolpin and colleagues report using cancer-free controls in their PanScan GWAS manuscript (81); sample overlap would therefore need to be undiagnosed lung cancer cases at the time of study. Given that the 5-year prevalence of lung cancer in the general population of Europe is 4.1% (82), we find it unlikely that there would be enough sample overlap to substantially bias our effect-estimate in this instance. The association between pancreatic cancer and lung adenocarcinoma could reflect a shared genetic architecture for pancreatic cancer that is translated in opposing directions to affect risk in these two diseases. In either case, our finding should not be interpreted as a direct causal effect of lung cancer on pancreatic cancer (or vice versa). The association between these SNPs and pancreatic cancer requires validation in larger GWAS and independent replication.

Smoking and type II diabetes, although previously reported risk factors (1, 2, 83, 84), did not show strong evidence of an association with pancreatic cancer in our analysis. While the lack of association shown for smoking in our analysis could indicate that previous observational associations are biased due to confounding or reverse causation, it is also possible that our results reflect low power. The SNPs comprising the instrument for smoking (cigarettes/day) are within the CHRNA3 gene region, which is reported to proxy for smoking heaviness among smokers rather than being representative of cigarettes per day in a general population (85, 86). As such, the outcome GWAS data would have to be restricted to current smokers to produce a meaningful effect-estimate. We could not stratify in this way due to the sole use of summary GWAS statistics; therefore, the effect-estimate generated by our analysis is not conclusive.

Numerous meta-analyses and pooled analyses have been performed looking at the association of diabetes and pancreatic cancer, all showing that long-term diabetes is associated with a $\geq 50\%$ increased risk of pancreatic cancer (2, 87–92). Our analysis found little evidence to suggest genetic liability to type II diabetes has a causal effect on pancreatic cancer; a finding also reported by Carreras-Torres and colleagues (74).

**Strengths**

We appraised the association of a multitude of phenotypes with a rare cancer type in a hypothesis-free manner. Our approach features a two-sample MR design, utilizing summary-level data; a particularly valuable method when the outcome of interest is rare, or when the capacity to investigate phenotypes in single studies is limited. For example, given limited power and sample size due to the cost of metabolomic platforms, many metabolites would unlikely have been investigated in relation to pancreatic cancer risk in observational studies. However, because genetic instruments for a multitude of metabolites have been obtained in previous studies with large sample sizes (93, 94), the two-sample MR framework allows the appraisal of the causal effect of the metabolome on health and disease.

**Limitations**

One limitation of the approach applied here is that not all possible phenotypes have genetic instruments or have not yet been curated in MR-Base. Therefore, some potentially associated phenotypes (e.g., occupational phenotypes and chronic pancreatitis) with pancreatic cancer could not be appraised. Because of the multiple testing burden of this analysis, there was potential for false negative findings. To remain conservative in such a broad approach, we chose to only present phenotypes that surpassed a strict Bonferroni correction in our main analysis. However, phenotypes showing weaker evidence for association (uncorrected $P < 0.05$) may contain some true associations and have therefore been included in our Supplementary Materials ($P < 0.05$; see Supplementary Table S3). On the other hand, the MR approach may identify false positive findings, particularly if there is a horizontal pleiotropic effect of a genetic instrument on the outcome, which was evident for some of the phenotypes identified here.

Given a binary outcome of pancreatic cancer, our MR models (maximum likelihood and IVW) are two-stage estimators where the second stage uses a log-linear regression model to derive an OR parameter. Estimates from such an approach will be overly precise, as uncertainty in the first-stage regression is not accounted for (95). However, this overprecision may be slight if the SE in the first-stage coefficients is low, and can be resolved by using a maximum likelihood method (95). We provide maximum likelihood estimates in addition to IVW estimates in our MR-PheWAS analysis; these estimates are similar across our main findings, indicating that the two-stage estimator with a logistic second-stage model is still a valid test of the null hypothesis here.

By systematically evaluating the association of all available phenotypes with GWAS data in the MR-Base repository of summary genetic data, we may not have had sufficient power to detect a true causal association for every analysis conducted; particularly those proxied by low numbers of SNPs, which may infer a low phenotypic variance explained. Low numbers of SNPs to proxy a phenotype are particularly prevalent when assessing the causal association of metabolites (93, 94) with pancreatic cancer; these phenotypes account for 255 of the 486 phenotypes tested, with a median 2 SNPs per metabolite. However, precise measurement of metabolites via nuclear magnetic resonance (NMR) and liquid chromatography-mass spectrometry (LC/MS) result in relatively large metabolite GWAS per-allele effect sizes and phenotypic variance explained (93, 94). The median variance explained by our metabolite phenotypes was 1.8%; at this variance explained, we had 80% power, with a sample size of 14,374 (7,110 cases, 7,264 controls) to detect an OR of 1.76 (beta of 0.5) at an alpha of 0.05; see Supplementary Table S3). On the other hand, our metabolite phenotypes was 1.8%; at this variance explained, we had 80% power, with a sample size of 14,374 (7,110 cases, 7,264 controls) to detect an OR of 1.76 (beta of 0.5) at an alpha of 1.6 $\times 10^{-4}$ (0.05/312 independent tests).
Conclusions
Within the context of a highly aggressive cancer for which the underlying causes are poorly understood, we undertook an MR-PheWAS study, which was able to suggest a causal association of a previously identified phenotype for pancreatic cancer in observational epidemiologic literature (BMI), suggest association between an anthropometric phenotype (hip circumference) with pancreatic cancer, and provide insights into some potentially novel mechanisms (metabolic factors and shared genetic architecture with lung cancer) for this disease.

Disclosure of Potential Conflicts of Interest
M.R. Munafo reports receiving commercial research grants from Pfizer and Cambridge Cognition, and is a consultant/advisory board member for Cambridge Cognition. No potential conflicts of interest were disclosed by the other authors.

Authors’ Contributions
Development of methodology: R.J. Langdon, K.H. Wade, P. Brennan
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): R.J. Langdon, R. Carreras-Torres, P. Brennan
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): R.J. Langdon, G. Hemani, J. Zheng, M. Johansson, P. Brennan, R. E. Wootton, M. R. Munafo, R. M. Martin, P. Haycock

References
Langdon et al.
