bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome

Tom G. Richardson[1*], Sean Harrison[1], Gibran Hemani[1], George Davey Smith[1]

[1] *MRC Integrative Epidemiology Unit (IEU), Population Health Sciences, Bristol Medical School, University of Bristol, Oakfield House, Oakfield Grove, Bristol, BS8 2BN, United Kingdom*

*Corresponding author: Dr Tom G. Richardson, MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, Oakfield House, Oakfield Grove, Bristol BS8 2BN, UK. Tel: +44 (0)117 3313370; E-mail: Tom.G.Richardson@bristol.ac.uk

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

The age of large-scale genome-wide association studies (GWAS) has provided us with an unprecedented opportunity to evaluate the genetic liability of complex disease using polygenic risk scores (PRS). In this study, we have analysed 162 PRS ($P<5x10^{-05}$) derived from GWAS and 551 heritable traits from the UK Biobank study (N=334,398). Findings can be investigated using a web application (http://mrcieu.mrsoftware.org/PRS_atlas/), which we envisage will help uncover both known and novel mechanisms which contribute towards disease susceptibility.

To demonstrate this, we have investigated the results from a phenome-wide evaluation of schizophrenia genetic liability. Amongst findings were inverse associations with measures of cognitive function which extensive follow-up analyses using Mendelian randomization (MR) provided evidence of a causal relationship. We have also investigated the effect of multiple risk factors on disease using mediation and multivariable MR frameworks. Our atlas provides a resource for future endeavours seeking to unravel the causal determinants of complex disease.

Word count: 150

Key words: polygenic risk scores, phenome-wide association study, Mendelian randomization, polygenic prediction

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Developing our understanding of how modifiable social, behavioural and physiological factors influence risk of disease is of vital importance to improve effective medical treatment and preventative interventions[1]. Genetic factors may also contribute substantially to disease susceptibility, as demonstrated by recent large-scale genome-wide association studies (GWAS) which have uncovered thousands of trait-associated single nucleotide polymorphisms (SNPs) throughout the human genome. However, typically the magnitude of effect and variance explained by one of these common genetic variants is small[2]. Polygenic risk scores (PRS), commonly defined as the sum of trait-associated SNPs weighted by their effect sizes, harness findings from GWAS to provide an overall measure of an individual's genetic liability to develop disease[3]. Although early applications of PRS were found to be underwhelming in terms of disease prediction[4], breakthroughs in the scale of GWAS and accessibility to biobank scale datasets have considerably improved their performance[5, 6]. As such, they hold considerable potential to improve early disease prognosis and treatment plan formulation[7].

Along with the emerging utility of PRS to predict disease, they have also been previously used to evaluate putative causal relationships[8, 9]. For example, instead of using a coronary heart disease (CHD) PRS to predict incidence of this disease, studies have investigated whether scores for known risk factors, such cholesterol and lipid levels[10], are also strongly associated with CHD incidence. One such approach in this paradigm is Mendelian randomization (MR), a method by which genetic variants are leveraged as instrumental variables to investigate causal relationships between modifiable risk

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

factors and disease outcomes[11, 12]. MR is typically limited to using SNPs which survive conventional GWAS corrections (i.e. $P<5x10^{-08}$), which may lack statistical power if these variants do not explain a large proportion of trait variance[13]. In contrast, PRS derived using a more lenient threshold (e.g. $P<5x10^{-05}$) can help recover some of this missing heritability due to a larger number of SNPs being included. This may help improve detection rates for causal relationships, which can be particularly useful when evaluating associations between genetic liability for a given trait and hundreds of diverse health outcomes. Such endeavours are commonly referred to as phenome-wide association studies[14, 15, 16 17].

To investigate this we undertook a preliminary simulation study to compare the performance of using a PRS to detect causal relationships with a popular MR approach (the inverse variance weighted (IVW) method[18]) (Figure 1). Results indicated that, although using a PRS provides higher statistical power, it also suffers from substantive false positive rates due to horizontal pleiotropy, the phenomenon whereby a gene influences multiple traits via independent biological pathways[12]. SNPs which are known to be pleiotropic with large effects on different and diverse traits have been found to distort findings from PRS analyses[19]. As a consequence, findings from phenome-wide association studies using a PRS may be useful in terms of highlighting putative causal associations, although robust evaluations are necessary to investigate results. We therefore propose that investigating whether PRS associations are robust to various sensitivity analyses developed in the field of MR are necessary to discern whether they represent causal relationships. To facilitate such future analyses, an accessible resource to evaluate associations between disease genetic liability and complex traits from across the human phenome should prove to be of considerable value.

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

In this study, we have constructed 162 different PRS (based on $P<5x10^{-05}$) using findings from large-scale GWAS and evaluated their association with 551 traits in up to 334,398 individuals enrolled in the UK Biobank study[20, 21]. To disseminate these findings, we have developed a web application to examine and visualise this derived atlas of associations. We have also undertaken follow-up analyses to demonstrate the usefulness of this resource to help identify putative causal relationships. Firstly, we have interpreted findings from a hypothesis-free scan of associations between the schizophrenia PRS and each of the 551 traits. We demonstrate that amongst these findings are associations which may likely reflect underlying causal relationships. We have also showcased the utility of evaluating the association between all 162 PRS and a single outcome using our atlas. Using gout susceptibility as an example, we demonstrate how recently developed methodology (mediation MR and multivariable MR) can be applied to evaluate the effects of multiple risk factors on disease risk.

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Results

An atlas of polygenic risk score associations across the human phenome

Overall, we undertook 89,262 tests to investigate the association between 162 different PRS derived from GWAS (Supplementary Table 1) and 551 complex traits from the UK Biobank study (Supplementary Table 2). PRS were constructed using independent SNPs for each GWAS ($P<5x10^{-05}$) based on $r^2 < 0.001$ using genotype data from European individuals (CEU) from phase 3 (version 5) of the 1000 Genomes project[22]. As opposed to the conventional GWAS cut-off of $P<5x10^{-08}$, the threshold of $P<5x10^{-05}$ was selected to incorporate additional SNPs into scores which may explain additional heritability for GWAS traits. Furthermore, this allowed us to create PRS for traits which had no SNPs surviving conventional GWAS corrections, as well as increasing the number of SNPs used in scores for traits with only a small number of GWAS hits. Of the 162 GWAS we identified, 11 reported that they included UK Biobank participants in their analysis. As this may lead to overfitting, the PRS for these 11 traits were not weighted to reduce this source of bias. In case they are still useful for follow-up analyses despite overlapping with UK Biobank, these scores have been clearly flagged in Supplementary Table 1 by being allocated to the 'unweighted' subcategory.

In this study we have only interpreted findings from associations with PRS derived using the $P<5x10^{-05}$ threshold. However, analyses have been repeated using scores derived using the conventional GWAS threshold of $P<5x10^{-08}$ for future studies that wish to evaluate these results. Complex traits from the UK Biobank study were selected based on $P<0.05$ from previously undertaken heritability analyses within this study[23]. This threshold was chosen as a heuristic to highlight associations worth pursuing in

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

further detail. A web app to query and visualise these results can be found at http://mrcieu.mrsoftware.org/PRS_atlas/.

Stratifying the UK Biobank sample into deciles based on their PRS supported previous findings in the literature demonstrating the ability of PRS to predict risk of disease. For example, comparing the highest and lowest deciles of the coronary heart disease (CHD) PRS found that individuals had increased odds of 3.64 to develop this disease (based on the ICD10 code 'I25'). Combining this PRS with scores for established causal risk factors for CHD suggested that they can help improve polygenic prediction (namely low density lipoprotein (LDL) cholesterol and myocardial infarction), although integrating any associated scores in a hypothesis-free manner may hinder prediction (Supplementary Figure 1). This could potentially be attributed to the increase in variance incorporated into prediction analyses from scores that do not directly influence CHD, or alternatively may indicate that they are spurious associations. Amongst other findings, we observed that participants had increased odds of 2.43 in terms of obtaining a University or College degree when comparing top and bottom deciles for the years of schooling PRS. Other noteworthy examples included a 3.48 fold increase in odds of taking atorvastatin as medication when comparing the extreme deciles for the LDL PRS. We also observed that participants in the highest decile for the ulcerative colitis PRS had increased odds of 5.36 in terms of developing this disease in comparison to those in the lowest decile (based on the ICD10 code 'K51').

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Uncovering known and novel findings by conducting a phenome-wide evaluation of associations

To demonstrate the value of this atlas of results, we have investigated some of the strongest associations detected between the schizophrenia PRS and all 551 complex traits analysed in the UK Biobank study (Figure 2, Supplementary Table 3). Associations within our atlas could potentially be identified due to underlying epidemiological relationships, although there are various other possible explanations such as a shared genetic aetiology between traits. To investigate this for our associations with the schizophrenia PRS, we have used various methods in two-sample MR as an example of how future studies could evaluate findings from our atlas. For these analyses we only used SNPs with $P<5x10^{-08}$ as instrumental variables to reduce the likelihood of weak instrument bias in our analysis[24]. Our systematic approach involved the following:

1. As an initial evaluation, we investigated evidence of association using the inverse variance weighted (IVW)[18] method and derived Cochran's Q statistic[25] as an indicator of potential heterogeneity. Weak evidence of association in this analysis suggests that a causal effect is unlikely.
2. If the IVW method provides strong evidence of association but in the presence of heterogeneity, we suggest undertaking two additional MR analyses using the weighted mode[26] and weighed median[27] methods. If there is a lack of strong evidence in both of these analyses then associations are unlikely to be causal.
3. As a sensitivity analysis, repeat steps 1 and 2 but only using SNPs as instruments which are not filtered out by applying the MR directionality test[28]. We also recommend evaluating the MR-Egger intercept term[29] to discern whether estimates may be biased by directional pleiotropic effects.

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The top association with the schizophrenia PRS suggests that individuals with high schizophrenia genetic liability have increased odds of seeing a psychiatrist at some point in their lives due to nerves, anxiety, tension of depression (OR=1.09 per standard deviation increase in PRS, 95% CI=1.08 to 1.10, $P=1.55x10^{-50}$). The schizophrenia PRS was also strongly associated with various neurological traits, such as neuroticism (Beta=0.066, SE=0.006, $P=8.17x10^{-27}$), being 'tense or highly strung' (OR=1.07, 95% CI=1.07 to 1.08, $P=2.25x10^{-47}$) and self-reported depression (OR=1.07, 95% CI=1.06 to 1.08, $P=4.91x10^{-18}$).

We identified strong evidence that schizophrenia genetic liability influences this set of neurological traits (Supplementary Table 4), except for self-reported depression where strong evidence was only detected using the inverse variance weighted (IVW) method (Beta=0.004, SE=0.001, P=0.009). There was also no strong evidence of directional horizontal pleiotropy for these results based on the MR Egger intercept term and associations were detected after repeating analyses after applying MR directionality filtering.

Along with using MR to investigate the effect of PRS traits on outcomes, we recommend investigating the converse direction of effect where possible (also known as 'bi-directional' MR[30]). Undertaking this analysis detected suggested that neuroticism influences schizophrenia risk. (Supplementary Table 5), although we detected evidence of directional horizontal pleiotropy based on the MR Egger intercept term (Beta=0.043, SE=0.018, P=0.018). After applying MR directionality filtering, we also identified evidence of association between being 'tense or highly strung' and schizophrenia risk.

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Therefore, the most parsimonious explanation for these findings could be that they have been observed due to a shared genetic aetiology between schizophrenia and other neurological traits. This is also likely to be a plausible explanation for other associations within our atlas. In particular, caution is advised when interpreting findings between autoimmune traits which are known to be influenced by highly correlated genes residing in the HLA region of the genome[31]. Although these findings could still be of interest in terms of genetic correlations between traits, they may not reflect underlying causal relationships[32].

Amongst other findings, there were associations which suggested individuals with high schizophrenia genetic liability had a lower fluid intelligence score (Beta=-0.083, SE=0.006, $P=1.49x10^{-39}$). We also observed evidence that these individuals performed worse than others in an assessment of cognitive function concerning memorising pairs of cards (Beta=0.020, SE=0.002, $P=6.66x10^{-34}$ for 'number of incorrect matches'). Follow-up MR analyses provided evidence from multiple methods that schizophrenia genetic liability influences both of these outcomes (Supplementary Table 6). These results were robust to sensitivity analyses using MR directionality filtering and MR Egger intercepts did not indicate that findings were prone to directional horizontal pleiotropy. In contrast, there was a weak evidence of a causal effect in the opposite direction for these associations, in particularly after applying MR directionality filtering (Supplementary Table 7). We also conducted a leave-one out analysis which suggested that no individual SNPs were responsible for driving observed effects (Supplementary Figures 2 & 3). Taken together, these analyses support evidence that schizophrenia genetic liability may lead to reduced cognitive function.

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Elsewhere, there were associations indicating that participants with a high schizophrenia PRS were more likely to be unsuccessful when attempting to quit smoking (Beta=0.028, SE=0.003, $P=3.87x10^{-22}$) and, accordingly reduced odds of being a past smoker (OR=0.97, 95% CI=0.97 to 0.98, $P=9.71x10^{-17}$). We observed strong evidence of association that schizophrenia genetic liability influences these outcomes (Supplementary Table 8), whereas the converse direction of effect provided weak evidence of an effect (Supplementary Table 9). However, the 'number of unsuccessful smoking attempts' outcome could only be instrumented using a single variant which limits our ability to investigate this effect. Moreover, a recent study has uncovered a large number of SNPs robustly associated with smoking cessation and provided evidence of a bi-directional relationship between smoking and schizophrenia using MR[33]. Leave-one out analyses suggested that no individual SNP was responsible for driving observed associations (Supplementary Figures 4 & 5).

We also observed a strong inverse association between the schizophrenia PRS and various anthropometric traits. However, evaluating the relationship between schizophrenia and body mass index (BMI) provided weak evidence of a causal effect in both directions (Supplementary Tables 10 & 11). This result reinforces our recommendation that all findings within our atlas require in-depth evaluation to discern whether they represent potential causal associations.

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Elucidating risk factors which may play a mediating role along the causal pathway to disease

Another strength of our atlas is that findings can be evaluated by selecting an outcome of interest and evaluating which of the 162 PRS are most strong associated with it. Doing so may motivate future endeavours to investigate the effect of multiple risk factors on disease risk. As a demonstration of this, we have evaluated the associations between all PRS and self-reported gout in the UK Biobank study (Supplementary Table 12). In this analysis, there was strong evidence of association using the PRS for gout itself (OR=1.16, 95% CI=1.13 to 1.19), although we also observed a much larger magnitude of effect using the urate PRS (OR=1.75, 95% CI=1.72 to 1.78). This result is likely representative of other findings within our atlas, where the PRS for the disease of interest may not always necessarily be the best polygenic predictor of it.

A receiver operating characteristic plot (Supplementary Figure 6) illustrates this point, where the area under curve for the gout PRS was 0.54 in comparison to the urate PRS which had a value of 0.65. This may be attributed to gout being a binary outcome heavily influenced by the number of cases analysed in its corresponding GWAS (N=2,115). In comparison, urate is a continuous trait measured in all participants for its respective GWAS (N=110,347). After urate, the next strongest positive associations with self-reported gout were triglycerides (TG) and body mass index (BMI) (OR=1.14, 95% CI=1.11 to 1.16 and OR=1.09, 95% CI=1.06 to 1.12 respectively). However, it is unclear whether these risk factors influence gout risk independently of one and other or if they reside on the same causal pathway to disease.

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We investigated this by firstly using an MR mediation framework[34] which involved evaluating bi-directional relationships for each risk factor in turn. As before, only SNPs with $P<5x10^{-08}$ for each PRS were used as instrumental variables in MR analyses. There was strong evidence that BMI had a causal effect on each other trait in turn (TG, urate and gout), where effect estimates appeared to be consistent between different MR methods (Supplementary Table 13). Repeating this analysis for TG as our exposure provided evidence of a causal effect on urate and gout risk, but not BMI (Supplementary Table 14). We then modelled urate as our exposure variable, which suggested that increased urate positively influences gout risk, although there was weak evidence of an effect on either BMI or TG (Supplementary Table 15). In all analyses there was no strong evidence of horizontal pleiotropy based on the MR-Egger intercept terms and findings were robust to sensitivity analyses using MR directionality filtering (Supplementary Tables 13-15). We also undertook leave-one out analyses which found that no single SNP was driving observed effects (Supplementary Figures 7-10). In conclusion, as illustrated in Figure 2a, findings from the mediation MR analysis suggests that BMI influences TG levels (Figure 3a (1)), which has an effect of urate (Figure 3a (2)), and this subsequently influences gout risk (Figure 3a (3)). Using the effect estimates from our IVW analysis, we estimated that 77% of the overall effect of BMI on gout risk (Figure 3a (4)) is mediated through this causal pathway.

We also used a related approach to investigate the effect of these multiple risk factors on gout susceptibility, known as multivariable MR[35]. In this analysis genetic instruments for all exposures (i.e. BMI, TG and urate) are modelled simultaneously to investigate whether these risk factors influence our outcome (i.e. gout) independently of one and other. We observed the effects of BMI and TG on gout risk attenuate when analysed in

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the same model as urate (Supplementary Table 16). Furthermore, in subsequent analyses we applied multivariable MR to investigate each pairwise combination of these risk factors on gout risk. There was evidence of an attenuation of the effect of BMI on gout risk when accounting for either the TG or urate effect (Supplementary Table 17 & 18). We also observed the effect of TG on gout risk attenuate when accounting for urate levels (Supplementary Table 19). These findings therefore support the same direction of effect observed using the mediation framework (Figure 3b).

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

In this study we have developed an atlas of associations between PRS and complex traits across the human phenome. Along with contributing to mounting evidence that PRS can be valuable in predicting later life disease outcomes, we have provided examples of how this resource can be harnessed to help identify potential risk factors in disease which warrant further investigation. We envisage that the inferences we have made in this study are just the beginning of potential findings which can be uncovered using such catalogues of associations. Multiple lines of evidence from robust follow-up studies of putative causal risk factors will help improve our understanding of disease susceptibility[36].

Large-scale biobank datasets provide an unparalleled opportunity to undertake hypothesis-free causal inference. Such efforts can help identify evidence supporting established causal relationships, as well as potentially implicating novel ones[12, 37]. We have illustrated this type of approach in our study by evaluating the results of a phenome-wide association study of schizophrenia genetic liability. We identified various associations with different neurological traits such as depression and neuroticism, which may likely be explained due to having a shared genetic component with schizophrenia risk.

There were also strong associations with measures of cognitive function and smoking behaviour which MR follow-up analyses suggested may be due to putative causal relationships with schizophrenia genetic liability. There is long standing evidence from the literature that cognitive impairment is a recognised characteristic of

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

schizophrenia[38]. Although PRS may prove useful in determining lifelong risk of developing schizophrenia, based on currently available data they may be less effective in terms of predicting age of schizophrenia onset as well as the severity of its progression. Characterization of cognitive decline in individuals with a high schizophrenia PRS may therefore help to better understand its neurological basis, and therefore improve our capability to treat it[39].

There is also a wealth of evidence in the literature from observational studies that individuals diagnosed with schizophrenia smoke more frequently compared to the general population[40]. Our results indicate that UK Biobank participants with a high schizophrenia genetic liability are more likely to be unsuccessful in their attempts to stop smoking. This may therefore suggest that the high frequency of schizophrenia patients who smoke could be attributed to their inability to quit smoking. However, we were unable to support recent evidence which suggests that smoking is a risk factor for schizophrenia which could be attributed to weak instruments in our analysis[33]. The positive association with smoking behaviour may also provide a possible explanation for the inverse association we observed between schizophrenia genetic liability and anthropometric traits. Evaluating the relationship between schizophrenia genetic liability and body mass index supports this as we identified weak evidence of a direct causal relationship in this analysis. Not only does this result emphasise the importance of evaluating associations detected in our atlas, but also suggests that findings could be valuable in terms of uncovering traits that mediate the effect of genetic risk on disease.

In this study we have also provided an example of how investigating various PRS associations with the same outcome may help motivate studies evaluating the effect of

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

multiple risk factors on disease risk. Our analysis detected evidence of an association between body mass index and gout risk, putatively mediated by triglycerides and urate levels. The findings from this analysis therefore appear to recapitulate known biology regarding the established causal pathway to gout[41,42]. Speculatively, a diet including high calorie and alcohol consumption, which are known risk factors for increased body mass index and triglyceride levels, may result in elevated circulating uric acid level and in turn increase gout risk. A recent study has suggested that genetic factors may have a greater impact on serum urate levels than environmental factors such as diet[43]. Our findings suggest that genetic drivers of appetite which may influence higher BMI levels are likely to predominantly influence gout risk via increased urate levels. We hope this illustration will motivate creative hypotheses for future endeavours to investigate the effect of multiple risk factors on disease risk.

The application of PRS is a topic which has sparked considerable recent debate, particularly concerning whether scores are relevant for clinical decision making[44]. Although resources such as the UK Biobank provide an unparalleled opportunity to investigate the determinants of complex disease as we have done in this study, findings regarding genetic liability may not be generalizable to individuals who are not of European descent. As such, there is likely to be an emphasis in the forthcoming years on efforts to establish disease-specific datasets for a diverse range of ancestries. We also note that, although we have adjusted all analyses in our study using the top 10 principal components from the UK Biobank, there may still be an influence of geographic clustering which remains unaccounted for[45]. Furthermore, although we have flagged the PRS traits in our study derived using GWAS who samples overlap with the UK Biobank, we are unable to assess this for scores whose GWAS predate this cohort. Future efforts

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

to link anonymous identifiers between the UK Biobank and UK cohorts would be of helpful in terms of ascertaining this information to prevent overfitting. Lastly, certain complex traits in our study may benefit from being combined to improve statistical power. For instance, a more powerful approach to identify associations between genetic liability and statin medication could involve deriving a combined measure of all the different types of statins reported. Investigating these results in a hypothesis-free manner as we have described in this study may also prove useful for drug repurposing efforts.

Polygenic risk scores hold huge promise in the era of large-scale genetic epidemiology to identify individuals who are at high risk of disease. Associations detected between these scores and outcomes undertaken by large-scale analyses should prove powerful for future studies that wish to unravel causal relationships between complex traits. Doing so will help improve disease prevention by developing a stronger understanding of complex epidemiological pathways.

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Methods

Constructing polygenic risk scores from large-scale genome-wide association studies

We have used the MR-Base platform[46] to identify SNPs from large-scale GWAS to include in our PRS. Our inclusion criteria for selected GWAS was having a sample size of more than 1,000 participants, over 100,000 SNPs measured on genotyping arrays and based on European/mixed populations. If multiple studies were found for the same trait, we selected the most recent study or the one with the largest sample size.

PRS were constructed using SNPs for each GWAS trait based on $P < 5 \times 10^{-05}$. A threshold of $r^2 < 0.001$ was selected to identify independent SNPs using genotype data from European individuals (CEU) from phase 3 (version 5) of the 1000 genomes project[22]. When a GWAS SNP was not available from the UK Biobank study genotype data, we used a proxy SNP instead based on $r^2 \geq 0.8$ using the same reference panel. Scores were then calculated as the sum of the effect alleles for all SNPs weighted by their reported regression coefficients. However, a small subset of PRS were left unweighted to reduce the likelihood of overfitting. This was due to their GWAS including participants from the initial release of the UK Biobank study. As such, additional caution should be exercised when interpreting findings from these unweighted PRS. Prior to analysis, each PRS was normalised to have a mean of zero and a standard deviation (SD) of one. Our PRS construction pipeline was also applied using a more stringent threshold of $P<5x10^{-08}$. Although we have not interpreted any of the results using these more stringent scores in this report, they are available within our atlas for future use.

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Complex trait and genotype data from the UK Biobank study

We selected traits from the UK Biobank study[21] which had $P < 0.05$ in the heritability analyses conduct by the Neale lab[23]. Genotype data were available for approximately 490,000 individuals enrolled in the study. Phasing and imputation of these data are explained elsewhere[20]. Individuals with withdrawn consent, evidence of genetic relatedness or who were not of 'white European ancestry' based on a *K*-means clustering ($K = 4$) were excluded from analysis. After exclusions there were up to 334,398 individuals with both genotype and complex trait data who were eligible for analysis.

Statistical analysis

We evaluated the association between each combination of PRS and complex trait in the UK Biobank study using linear regression (for continuous traits), logistic regression (for case/control traits), ordinal logistic regression (for ordered categorical traits) and multinomial logistic regression (for unordered categorical traits). All analyses were adjusted for age, sex, the first 10 genetic principle components (to adjust for population stratification) and genotyping chip used to measure genetic data in participants. Only female participants were included in the 'Age at menarche' and 'Age at menopause' PRS analyses.

We also calculated R^2 coefficients for continuous traits and McFadden pseudo R^2 coefficients for other models by repeating analyses unadjusted for covariates. McFadden's R^2 is defined as:

$$R^2_{McF} = 1 - \ln(L_m / \ln(L_0)$$

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

where ln is the natural logarithm, L_0 is the value of the likelihood function of the model with no predictors and L_m is the likelihood of the model being estimated. We note that pseudo R^2 coefficients should not be interpreted in a similar manner to those derived using linear regression[47].

Mendelian randomization analysis

We used various two-sample MR methods to evaluate associations detected in the PRS analysis. This involved using the observed effects of the genetic variants used in the PRS on both the GWAS trait that the score was based on (treated as the exposure in our MR analysis) as well as the UK Biobank trait (treated as the outcome in our MR analysis). For all MR analyses we only selected SNPs with $P<5x10^{-08}$ based on GWAS findings as instrumental variables to reduce the likelihood of weak instrument bias[24]. In terms of MR methods, we applied the inverse variance weighted (IVW)[18], weighed median[27] and weighted mode[26] approaches. We also conducted several different sensitivity analyses to evaluate findings. We derived Cochran's Q statistic[25] when undertaking the IVW approach as an indicator of heterogeneity, as well as repeating all analyses after filtering out SNPs which the MR directionality test[28] suggested did not influence the outcome of interest through the analysed exposure. The intercept of the MR-Egger approach[29] was used to investigate directional horizontal pleiotropy and leave-one-out analyses (i.e. reapplying the IVW method after removing each SNP in turn with replacement) were conducted to discern whether any individuals SNPs were driving observed associations. These types of analyses are particularly important when assessing findings from our atlas, as one possible explanation is that they could be attributed to a single pleiotropic SNP which has a large effect size (e.g. the *APOE* locus which is associated with Alzheimer's disease and lipid levels).

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

To investigate the direction of effect for associations identified in the PRS analysis we undertook bi-directional MR[30]. This involves firstly modelling our PRS trait as our exposure and complex trait as our outcome, and subsequently the complex trait as our exposure and PRS trait as our outcome in a separate analysis. Lastly, we have undertaken two recent developments in the MR paradigm; mediation MR[34] and multivariable MR[35]. These methods can be used to investigate the effect of multiple risk factors on a single outcome, as well as uncover potential mediators in disease. In this study we have evaluated findings from the PRS analysis based on the $P < 5 \times 10^{-05}$ threshold. We note however that it is only advisable to apply techniques in MR using this threshold as long as in-depth sensitivity analyses (e.g. leave-one out, MR-Egger intercept) are also undertaken to evaluate alternative explanations for associations, as opposed to genetic liability.

All analyses were undertaken using R (version 3.5.1). The R packages 'shiny' v1.1 was used to develop the web application and 'highcharter' v0.5 was used to generate interactive plots. Figures in this manuscript were generated using 'ggplot2' v2.2.1.

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data availability

All summary statistics for the analyses undertaken in this study can be downloaded using our web application (http://mrcieu.mrsoftware.org/PRS_atlas/). Our dataset was derived from the UK Biobank study as part of projects 8786 and 15825. The same dataset can be created with an application to use data from the UK Biobank study (http://biobank.ctsu.ox.ac.uk/crystal/).

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Abraham G, *et al.* Genomic prediction of coronary heart disease. *Eur Heart J* **37**, 3267-3278 (2016).

2. Visscher PM, *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5-22 (2017).

3. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet*, (2018).

4. Ripatti S, *et al.* A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *Lancet* **376**, 1393-1400 (2010).

5. Khera AV, *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*, (2018).

6. Lee JJ, *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet* **50**, 1112-1121 (2018).

7. Lewis CM, Vassos E. Prospects for using risk scores in polygenic medicine. *Genome Med* **9**, 96 (2017).

8. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ* **362**, k601 (2018).

9. Palmer TM, *et al.* Using multiple genetic variants as instrumental variables for modifiable risk factors. *Stat Methods Med Res* **21**, 223-242 (2012).

10. Holmes MV, *et al.* Mendelian randomization of blood lipids for coronary heart disease. *Eur Heart J* **36**, 539-550 (2015).

11. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* **32**, 1-22 (2003).

12. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* **23**, R89-98 (2014).

13. Maher B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18-21 (2008).

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

14. Denny JC, *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* **31**, 1102-1110 (2013).

15. Fritsche LG, *et al.* Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. *Am J Hum Genet* **102**, 1048-1061 (2018).

16. Krapohl E, *et al.* Phenome-wide analysis of genome-wide polygenic scores. *Mol Psychiatry* **21**, 1188-1193 (2016).

17. Millard LA, Davies NM, Timpson NJ, Tilling K, Flach PA, Davey Smith G. MR-PheWAS: hypothesis prioritization among potential causal effects of body mass index on many outcomes, using Mendelian randomization. *Sci Rep* **5**, 16645 (2015).

18. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol* **37**, 658-665 (2013).

19. Felsky D, *et al.* Polygenic analysis of inflammatory disease variants and effects on microglia in the aging brain. *Mol Neurodegener* **13**, 38 (2018).

20. Bycroft C, *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).

21. 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).

22. 1000 Genomes Project, *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).

23. Neale lab. Rapid GWAS of thousands of phenotypes for 337,000 samples in the UK Biobank. (ed^(eds) (2017).

24. Davies NM, von Hinke Kessler Scholder S, Farbmacher H, Burgess S, Windmeijer F, Davey Smith G. The many weak instruments problem and Mendelian randomization. *Stat Med* **34**, 454-468 (2015).

25. Cochran WG. The comparison of percentages in matched samples. *Biometrika* **37**, 256-266 (1950).

26. Hartwig FP, Davey Smith G, Bowden J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int J Epidemiol* **46**, 1985-1998 (2017).

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

27. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet Epidemiol* **40**, 304-314 (2016).

28. Hemani G, Tilling K, Davey Smith G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet* **13**, e1007081 (2017).

29. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* **44**, 512-525 (2015).

30. Timpson NJ*, et al.* C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal Mendelian randomization. *Int J Obes (Lond)* **35**, 300-308 (2011).

31. Gough SC, Simmonds MJ. The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action. *Curr Genomics* **8**, 453-465 (2007).

32. O'Connor LJ, Price AL. Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nat Genet*, (2018).

33. Wootton RE*, et al.* Causal effects of lifetime smoking on risk for depression and schizophrenia: Evidence from a Mendelian randomisation study. *https://wwwbiorxivorg/content/early/2018/08/01/381301*, (2018).

34. Relton CL, Davey Smith G. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *Int J Epidemiol* **41**, 161-176 (2012).

35. Sanderson E, Davey Smith G, Windmeijer F, Bowden J. An examination of multivariable Mendelian randomization in the single sample and two-sample summary data settings. *https://wwwbiorxivorg/content/early/2018/04/27/306209*, (2018).

36. Munafo MR, Davey Smith G. Robust research needs many lines of evidence. *Nature* **553**, 399-401 (2018).

37. Cai T*, et al.* Association of Interleukin 6 Receptor Variant With Cardiovascular Disease Effects of Interleukin 6 Receptor Blocking Therapy: A Phenome-Wide Association Study. *JAMA Cardiol*, (2018).

38. Mohamed S, Paulsen JS, O'Leary D, Arndt S, Andreasen N. Generalized cognitive deficits in schizophrenia: a study of first-episode patients. *Arch Gen Psychiatry* **56**, 749-754 (1999).

39. Green MF. What are the functional consequences of neurocognitive deficits in schizophrenia? *Am J Psychiatry* **153**, 321-330 (1996).

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

40. Sacco KA*, et al.* Effects of cigarette smoking on spatial working memory and attentional deficits in schizophrenia: involvement of nicotinic receptor mechanisms. *Arch Gen Psychiatry* **62**, 649-659 (2005).

41. Matsubara K, Matsuzawa Y, Jiao S, Takama T, Kubo M, Tarui S. Relationship between hypertriglyceridemia and uric acid production in primary gout. *Metabolism* **38**, 698-701 (1989).

42. Li X*, et al.* Serum uric acid levels and multiple health outcomes: umbrella review of evidence from observational studies, randomised controlled trials, and Mendelian randomisation studies. *BMJ* **357**, j2376 (2017).

43. Major TJ, Topless RK, Dalbeth N, Merriman TR. Evaluation of the diet wide contribution to serum urate levels: meta-analysis of population based cohorts. *BMJ* **363**, k3951 (2018).

44. Warren M. The approach to predictive medicine that is taking genomics research by storm. *Nature* **562**, 181-183 (2018).

45. Abdellaoui A*, et al.* Genetic consequences of social stratification in Great Britain. *https://wwwbiorxivorg/content/biorxiv/early/2018/10/30/457515*, (2018).

46. Hemani G*, et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**, (2018).

47. Hu B, Shao J, Palta M. Pseudo-R2 in logistic regression model. *Statistica Sinica* **16**, 847-860 (2006).

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgements

We are extremely grateful to all the authors of the genome-wide association studies who have made their summary statistics publicly available for the benefit of this study. We would also like to thank the efforts of the Neale Lab who conducted extensive heritability analyses in the UK Biobank which guided our selection of traits to analyse.

This work was supported by the Integrative Epidemiology Unit which receives funding from the UK Medical Research Council and the University of Bristol (MC_UU_00011/1). G.H is supported by the Wellcome Trust [208806/Z/17/Z]. T.G.R is a UKRI Innovation Research Fellow (MR/S003886/1).

Competing interests

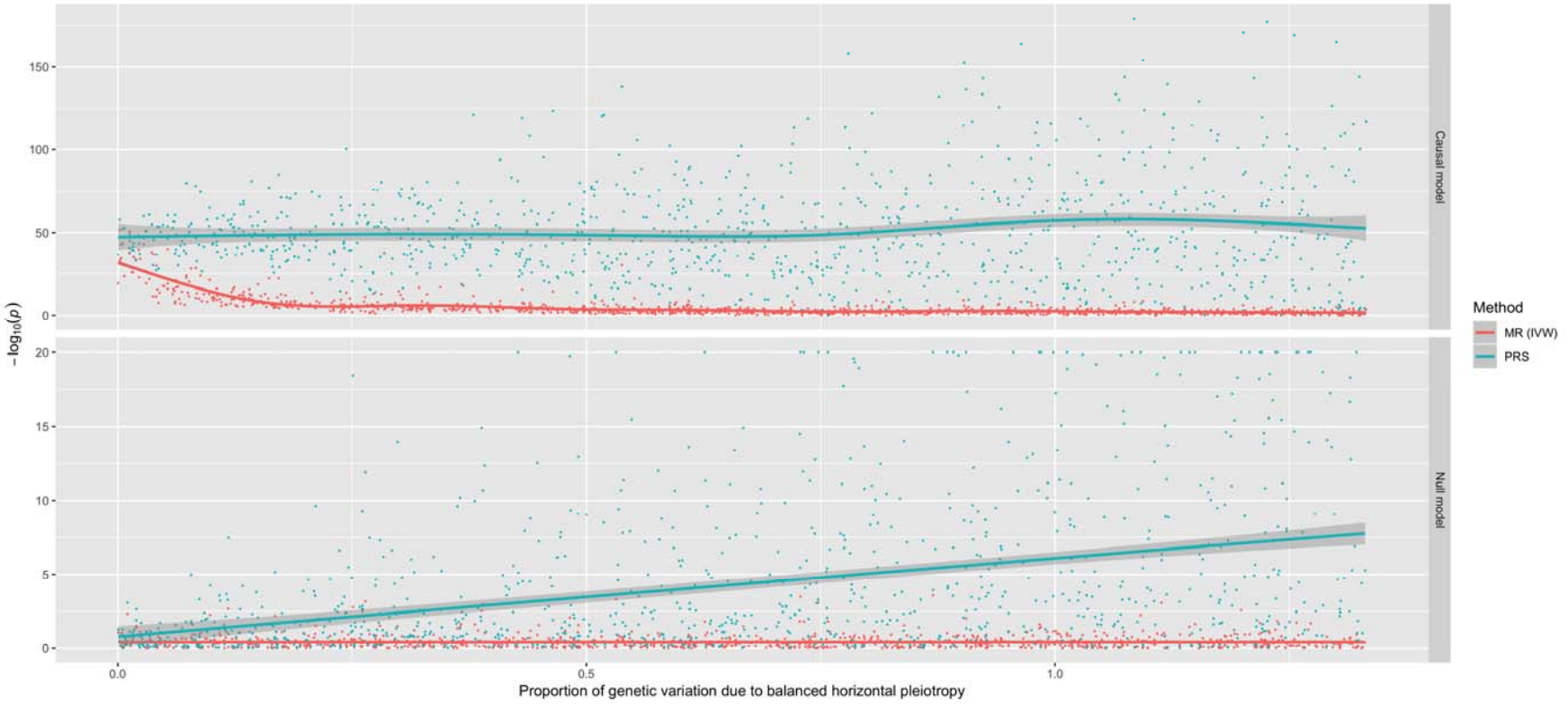
The authors declare no conflicts of interest.

Materials and Correspondence

This publication is the work of the authors and T.G.R. will serve as guarantor for the contents of this paper.

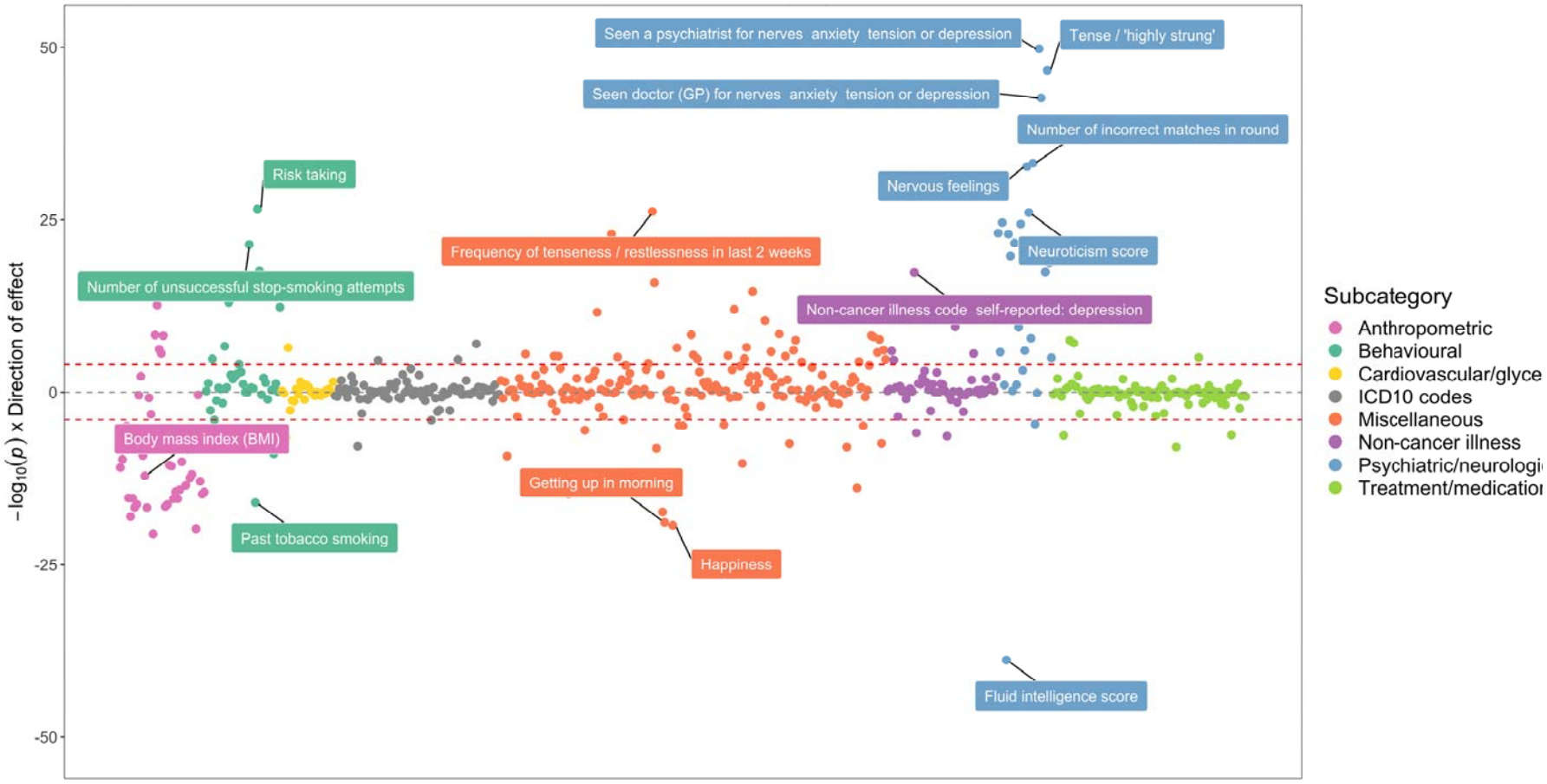
Figures

Figure 1: A simulation study to compare the performance of Mendelian randomization with polygenic risk score analysis



bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

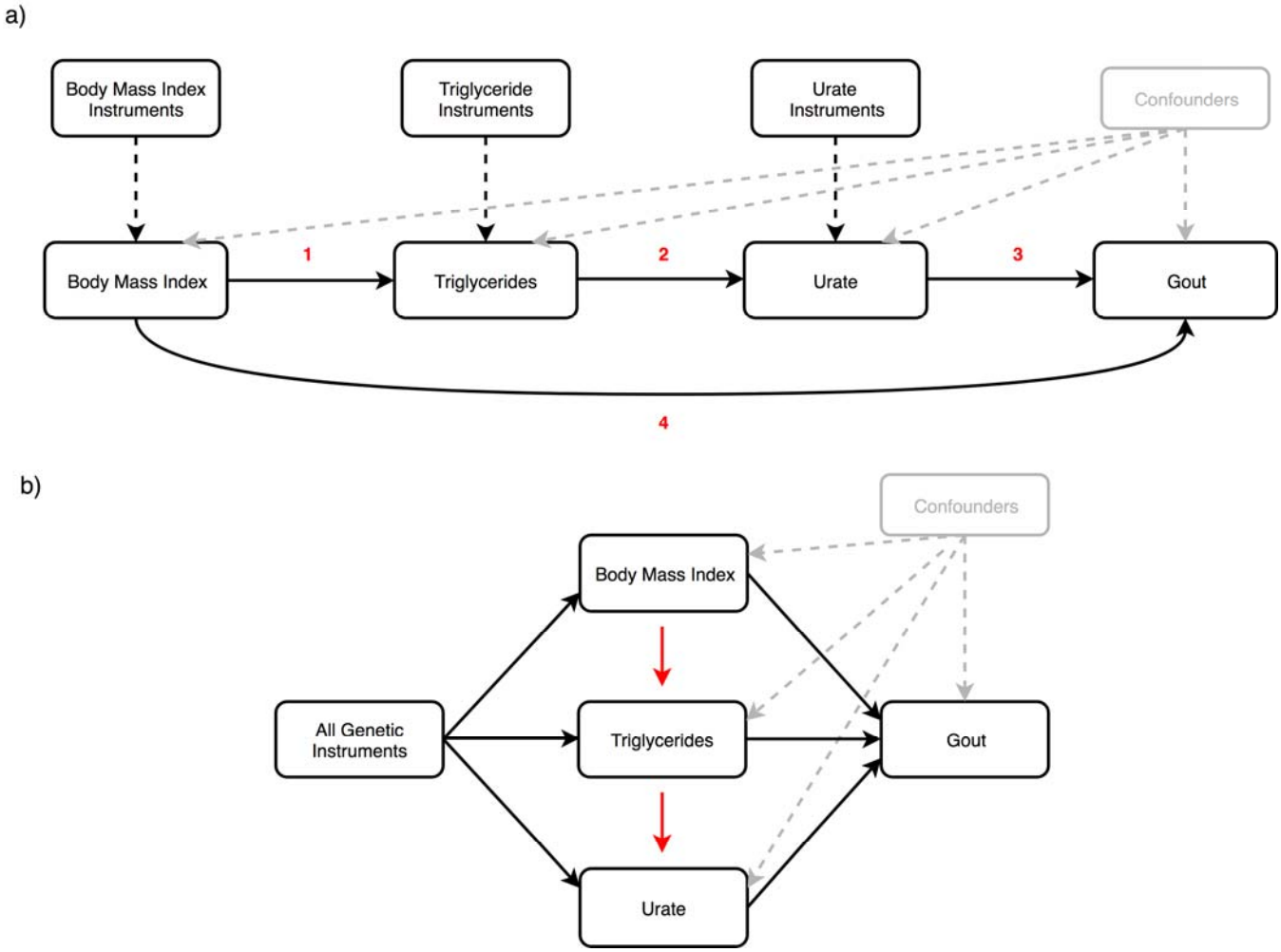
Figure 2: A bi-directional phenome-wide association plot for schizophrenia genetic liability



bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 3: Applying a) mediation and b) multivariable Mendelian randomization to investigate the causal effect of body mass index, triglycerides and urate on gout risk



bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure legends

Figure 1:

A comparison of the performance between the inverse variance weighted (IVW) Mendelian randomization (MR) model against polygenic risk score (PRS) analysis. Simulations were conducted under different levels of horizontal pleiotropy for two different models; the causal model (where the simulated exposure has a causal effect on the outcome) and the null model (where there is no causal effect between exposure and outcome).

Figure 2:

Each point on this plot represents the association between the schizophrenia polygenic risk score (based on $P<5x10^{-05}$) and a complex trait in the UK Biobank study. Along the y-axis are –log10 p-values for these associations multiplied by the direction of effect for their corresponding effect size. As such, traits positively associated with schizophrenia genetic liability reside above the horizontal grey line representing the null (i.e. –log10 (P) =0), whereas negative associations are below. Points are grouped and coloured based on their corresponding complex traits' subcategory. Horizontal red lines indicate the Bonferroni corrected threshold for the 551 tests undertaken (i.e. 0.05/551 = $9.07x10^{-05}$).

Figure 3:

a) Mediation Mendelian randomization (MR) framework to investigate whether urate mediates the effect of body mass index (BMI) and triglycerides (TG) on gout risk. The various analyses undertaken suggest that 1) elevated BMI

bioRxiv preprint doi: https://doi.org/10.1101/467910; this version posted November 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

increases TG levels 2) which subsequently has an effect on urate 3) and this in turn influences gout risk. This mediation pathway may help explain the manner by which BMI, potentially driven by lifestyle factors such as diet, is a risk factor for gout.

b) Multivariable MR framework attempting to reproduce findings from the mediation analysis. Genetic instruments for BMI, TG and urate were analysed simultaneously to evaluate the joint effect of these risk factors on gout risk. The effect of BMI and TG on gout risk attenuated compared to univariable analyses, suggesting that they influence gout risk through increased urate levels. Investigating each combination of pairwise risk factors using this framework suggested that BMI influences TG rather than the opposite direction of effect, which also supports findings from the mediation analysis.