Polygenic Risk Scores

Executive Summary

- Polygenic risk scores, which consist of combining information from thousands of individual genetic variants, can be a powerful approach to identify individuals with markedly higher (or lower) risk of particular diseases.
- Polygenic risk scores are not stand-alone diagnosis tools. They complement existing risk factors to improve screening programmes by better targeting at-risk individuals. They can also be used to better interpret screening tests.
- For many diseases improved prevention is possible for those individuals at high risk through either lifestyle changes or medical intervention.
- In cases where diagnosis is challenging, genetic information can provide additional information to support the medical team.
- Our ongoing research will continue to improve the power of these predictions, their applicability to a diverse range of ancestries, and the suite of clinical applications where they can be applied.

1. Polygenic Risk Scores Background

More than a decade of research has established that for all common diseases many thousands of genetic variants (single nucleotide polymorphisms, or SNPs) contribute to disease risk. Most of these individual risk-SNPs are common in the population, but each one typically has only a small effect on risk. The same is true for many other continuous traits in humans, including anthropometric, physiological, biochemical, and cognitive measurements. This contrasts with many serious rare diseases where a single, rare, genetic change often has a large effect. For some common diseases (e.g. breast cancer) both types of genetics are in play: there are known genes where single changes can have substantial effects on risk (e.g. *BRCA1* and *BRCA2* for breast cancer), and also thousands of SNPs with individually small effects on risk.

To identify rare mutations which cause rare genetic diseases it is usually essential to sequence, or read, the entire genome of the individual (or at least the part that encodes genes). In contrast, the many common variants which affect risk of common diseases can be measured in an individual with a different, and much cheaper, technology — so-called genotyping chips — which measure a predetermined set of about 1 million of the 3 billion positions in the human genome.

For a particular disease or trait, a so-called *Polygenic Risk Score* (PRS), combines information from large numbers of SNPs across the genome (hundreds to millions) to give a single numerical score which is an aggregate summary of an individual's propensity to develop that disease on the basis of the DNA variants they have inherited. The idea of combining information across many common SNPs to predict disease risk has been around for more than 10 years, but it has only recently been possible to validate such methods and assess their potential clinical utility, largely thanks to the UK Biobank. In principle, PRS can be constructed for many diseases – the SNPs involved in a PRS, and the

weightings for them, will typically differ from disease to disease. For an individual, one could thus calculate PRS for many diseases.

Across a large set of individuals, for a particular trait, there will be a distribution of values for the PRS. Individuals with higher values of the score will be at higher risk of developing the disease on the basis of the common genetic variants they have inherited.

Recent studies have shown, for example, that individuals in the top 5% of the coronary disease PRS distribution are at about 3 fold increased risk of developing the disease compared to the population average (Khera *et al.*). This effect is large enough to warrant clinical attention, and for many of these individuals there will be clinical interventions (and behavioural changes) which will reduce disease risk. Women in the UK in the top 1% of the breast cancer PRS distribution (excluding *BRCA* genes) have a 30% lifetime risk of developing breast cancer. In this case it may well make sense to target screening preferentially at such individuals. For yet other diseases, including some of the mental health disorders, knowledge of increased genetic risk could be helpful in differential diagnosis, shortening the time between first seeing a doctor and getting the correct treatment. PRS are also likely to be helpful in guiding treatment choices, improving efficacy and reducing the risk of side effects.

For any particular disease, clinically meaningful increase in disease risk predicted by PRS typically only applies to a few percent of individuals (although as for coronary disease, in a large population, this information can direct treatment to hundreds of thousands of individuals). Critically, while most individuals will have average risk for any particular disease, it is very likely that they will be at the extreme of genetic risk for at least one disease. Early identification of these risks, through the availability of genome-wide genetic information, could have a profound effect on individual and population health, and on health-related expenditure. The possibility of generating PRS for many common diseases at a population scale would identify individuals in the tail of the risk distribution for a subset of diseases, and present an exciting opportunity to optimize care, prevention and screening accordingly.

2. Genomics plc methodology for deriving Polygenic Risk Scores.

Genomics plc has amassed and curated data from hundreds of publicly available sources[1], and using this can estimate the effect of millions of genetic variants on >10,000 measurements of human traits and disease outcomes. This resource builds on the output of the medical and academic community worldwide. Using these data, we have built a Bayesian non-parametric machine learning approach to identify clusters of traits and diseases affected by the same causal genetic variant in any particular region in the genome. Importantly, our algorithm can differentiate between clusters with association signals interleaved along the chromosome, but distinct causal variants. This approach allows both the detection of variant-trait correlations that would be missed in a trait-by-trait analysis (because they fall below the traditional significance threshold), as well as improved identification of causal variants.

[1] See our website for more information: https://www.genomicsplc.com/our-approach-to-using-data/

For the purpose of computing PRS, we leverage these clusters to better sift signal from noise in the large number of possible genetic predictors. In addition to the unsupervised clustering described above (where the algorithm has no *a priori* information about biological relationships between traits), we can also build on the knowledge that certain traits have explicit causal relationships from established epidemiology (e.g. high blood pressure and high lipid levels increase risk of coronary artery disease). Furthermore, better estimates of which variants are causal helps make our PRS applicable to a wider range of ancestries than competing approaches. While this does not completely overcome the challenge imposed by the heavy bias of existing studies towards European ancestry, it is a first step towards the key goal of making PRS as widely useful as possible.

3. Assessing the potential impact of risk stratification by PRS: four examples.

One helpful way of assessing the impact of a PRS for a specific disease is to see how effective that PRS is in stratifying individuals, whose genetic data was not involved in the original calculation, into groups with different risk profiles for the disease. In our PRS research, we have relied on the data from the UK Biobank cohort for such validation assessments. UK Biobank provides a very valuable resource for this process: it involves a large number of individuals on whom there is both genome wide genetic data (from a genotyping chip) and extensive health information.

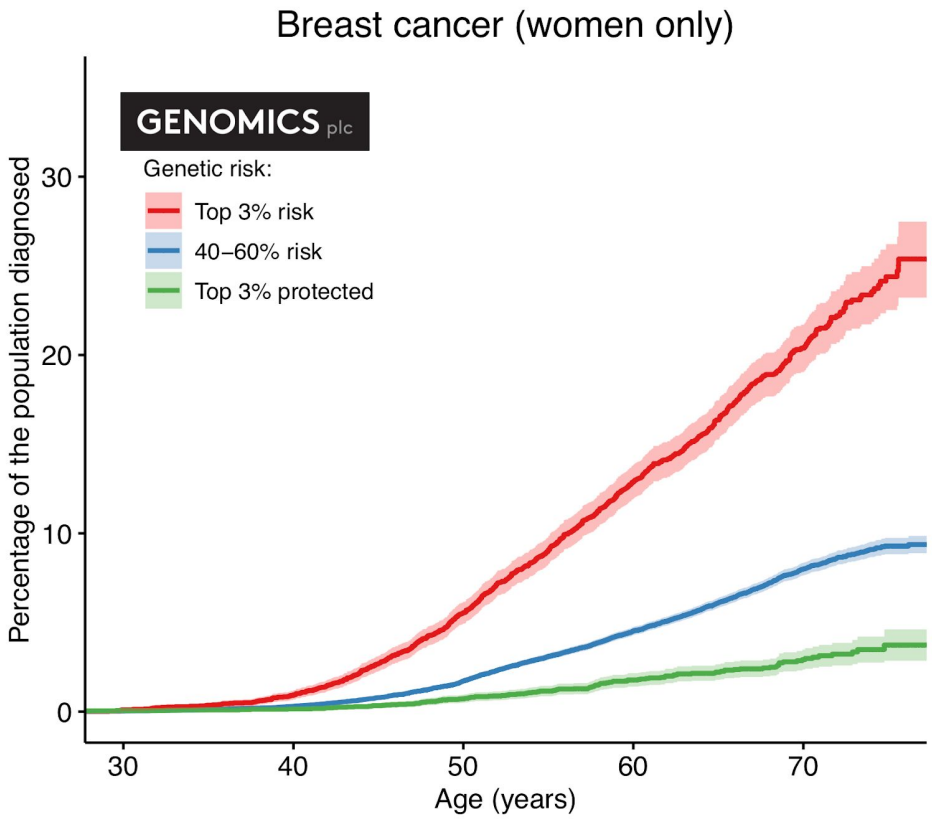
For a specific disease we use genetic studies outside UK Biobank to determine which SNPs to include in the PRS and the appropriate weightings for each of those SNPs. We can then transfer the algorithm for calculating a score into UK Biobank and calculate the score for each of a large number of individuals in UK Biobank. These can then be stratified into different groups depending on their PRS, for example high-risk individuals (those with the highest PRS for that disease), median-risk individuals (those with PRS in the middle of the distribution) and low-risk individuals (those with the lowest PRS for that disease). The UK Biobank data is then used to estimate, for each group, the cumulative incidence of disease in each of these groups, and these cumulative incidence curves can be compared visually. It is important to note that the UK Biobank's recruitment process does not perfectly reflect the average British individual. This cohort is generally healthier, and the rate of disease will tend to be lower than in the broader UK population.

We have so far worked on a set of 16 diseases for which predictions are both effective and actionable in some form for the individuals tested: age related macular degeneration, asthma, atrial fibrillation, breast cancer, chronic obstructive pulmonary disease, Crohn's disease, coeliac disease, coronary artery disease, glaucoma, hypertension, multiple sclerosis, obesity, prostate cancer, systemic lupus erythematosus, type 2 diabetes, and ulcerative colitis.

The figures below illustrate this risk stratification for four of these diseases: breast cancer, coronary artery disease, prostate cancer and type 2 diabetes. These analyses were based on published GWAS studies for which the summary statistics data were made available: Schumacher *et al.* (prostate cancer); Scott *et al.* (type 2 diabetes); Anand *et al.* (coronary artery disease); and Michailidou *et al.* (breast cancer). In each case we have evaluated performance in unrelated individuals in UK Biobank, matched on ancestry and sex (where appropriate) to the source studies. In the figures, we include

three indicative groups from UK Biobank, those individuals in the top 3%, median (ranging from 40-60%), and bottom 3% of PRS.

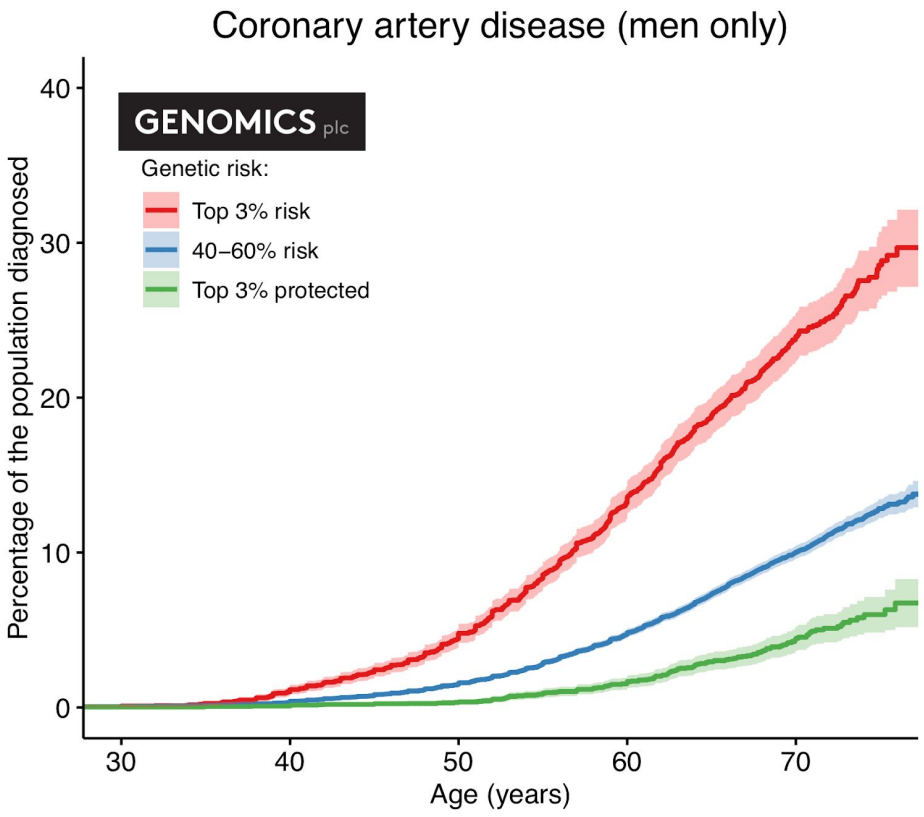
3.1. Breast cancer



In **breast cancer**, the disease incidences in the three groups are very different. A 45-year-old woman in the high-risk group has the same chance of having had breast cancer as a 55-year-old woman in the average group and a 75-year-old woman in the low-risk group. The chances of breast cancer by age 75 for a woman in the top group are around 25%, about three times higher than for one in the average group, and eight times higher than the low-risk group.

Currently in the NHS, breast cancer screening is based solely on age, with all women being offered screening when they turn 50. Adding PRS information to the current screening system could direct screens to those most at risk while reducing unnecessary screens for low risk individuals. In addition to targeting screening more effectively, polygenic risk scores affect the interpretation of screening results. A positive mammogram result for a woman in the top group is much less likely to be a false positive than one for a woman in the bottom group.

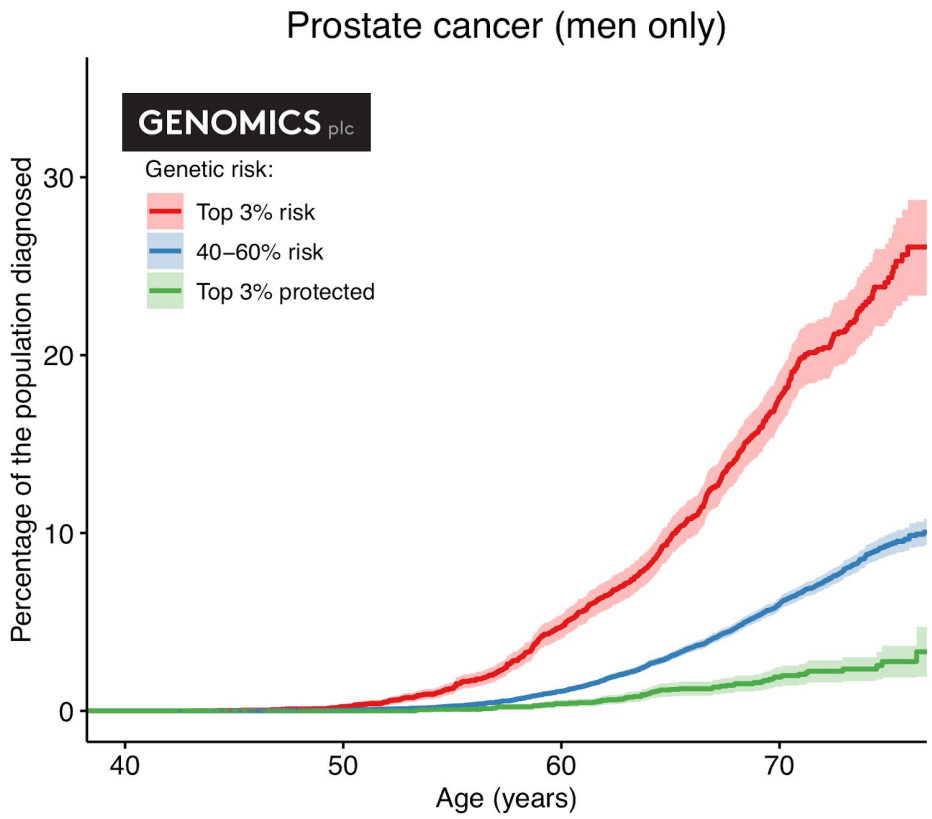
3.2. Coronary artery disease



In **coronary artery disease**, a similar pattern is observed with very different risk profiles across the genetically defined groups. The red curve relates to men with the top 3% of PRS for heart disease, the green curve to those with the lowest 3% and the blue curve to the middle 20%. A 50-year-old man in the high-risk (red) group has the same chance of having developed heart disease as a typical 60-year-old and the same chance as a 70-year-old in the low-risk group. Almost a third of the men in the red group will have developed heart disease by the time they are 75.

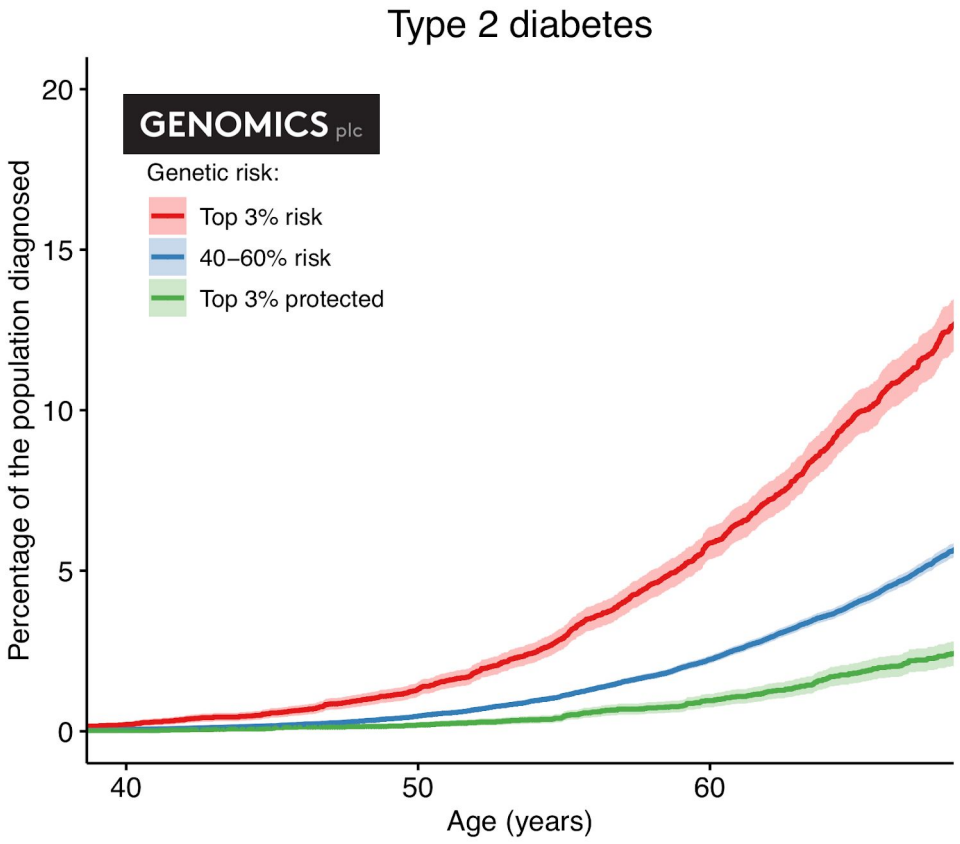
While further work is necessary, high-risk individuals could potentially be prescribed cholesterol-lowering statin medication. Deploying these medicines earlier to such individuals could potentially reduce coronary disease and its consequences such as myocardial infarction. At least some of the individuals in the high-risk group may also be more likely to make lifestyle adjustments to reduce their risk, although further empirical studies to assess this will be helpful.

3.3. Prostate cancer



In **prostate cancer**, there are also marked differences in incidence between groups defined by PRS values. One quarter of the men in the high-risk group will have developed prostate cancer by age 75, three times the proportion in the average group and almost 10 times the proportion in the low risk group. There is no early screening programme in the UK because of the lack of evidence that benefits outweigh the risks. A key concern is the high rate of false positives and the potential for harm resulting from unnecessary biopsies. While further work is required, PRS could potentially be used to target and screen at risk populations, as well as adjusting thresholds for prostate specific antigen (PSA) screening tests so that these are different for different PRS groups, which would reduce harmful false positives cases.

3.4. Type 2 diabetes



In **type 2 diabetes**, a similar stratification of the population is observed. Those with PRS scores in the top 3% of the population are almost three times more likely to develop type 2 diabetes than the median group. Lifestyle changes, including exercise, improved diet, and weight management, while relevant for the population as a whole, are likely to be particularly important for the high risk segment of the population. It may also be appropriate to monitor fasting glucose more regularly in high risk individuals.

References

Scott *et al.*, An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans, *Diabetes* 2017.
Michailidou *et al.*, Association analysis identifies 65 new breast cancer risk loci, *Nature Genetics* 2017.
CARDIoGRAMplusC4D Consortium, A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease, *Nature Genetics* 2015.
Schumacher *et al.*, Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci, *Nature Genetics* 2018.
Khera et *al*, Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations, *Nature Genetics* 2018.