TheScientist
EXPLORING LIFE, INSPIRING INNOVATION

READ THIS ARTICLE ONLINE

AI-Assisted Genome Studies Are Riddled with Errors

Researchers used artificial intelligence in large genomics studies to fill in gaps in patient information and improve predictions, but new research uncovers false positives and misleading correlations.

Sahana Sitaraman, PhD
Nov 13, 2024

The genome serves as the blueprint for the body, influencing every trait from the shape of the face to the arches of the feet, and even the development of certain diseases. While some disorders, like cystic fibrosis, are linked to single genes and can be reliably predicted based on a person's genetic data, many others—such as autism spectrum disorder, Alzheimer's disease, depression, and obesity— are not.

ABOVE:
Scientists often rely on incomplete survey data from biobanks to predict correlations between diseases and genetics. This approach can introduce bias and produce unreliable or false associations.
©ISTOCK, TANYAJOY

For the past 15 years, scientists have used genome-wide association studies (GWAS) to compare genomes of large groups of people to identify hundreds of thousands of genetic variants that are associated with a trait or disease.[1] This method has helped scientists unravel the underlying biology and risk factors of complex diseases and has also led to the discovery of novel drug targets. Despite these advancements, GWAS studies have their limitations, which scientists have tried to address with the help of artificial intelligence (AI). However, in two studies published in *Nature Genetics*, researchers at the University of Wisconsin-Madison identified pervasive biases these new approaches can introduce when working with large but incomplete datasets.[2,3]

GWAS rely on large biobanks with extensive patient data. However, these repositories could be lacking anything from blood reports, scans, and patient history to family data. Even with a thorough survey, challenges such as the lack of data on late onset diseases in a cohort of young participants can throw a wrench into researchers' plans.

To address gaps in the data, scientists developed two approaches: machine learning and GWAS-by-proxy (GWAX), which relies on family history data as predictors of late-onset diseases. Many researchers combine GWAS and GWAX to improve the statistical power of their predictions. However, the University of Wisconsin-Madison research team has found that these "solutions" can erroneously link gene variants with diseases.

"It has become very popular in recent years to leverage advances in machine learning, so we now have these advanced machine-learning AI models that researchers use to predict complex traits and disease risks with even limited data," said Qiongshi Lu, a biostatistician at the University of Wisconsin-Madison and coauthor of the studies, in a press release.

With AI-assisted GWAS, Lu and his colleagues noticed false associations between gene variants and type II diabetes. For example, four gene variants showed a high correlation with the disease in an AI-assisted GWAS, but not when using a conventional GWAS approach. However, previous research has shown that although these genes act on a cellular pathway that is indirectly connected to blood glucose levels it does not strongly influence them.

In cohorts where all samples have genetic data but only a fraction of samples have desired phenotypic data, AI-assisted GWAS algorithms try to fill in the gaps based on learned patterns. But without knowledge of physiological intricacies, this approach can lead researchers down the wrong path.

"The problem is if you trust the machine learning-predicted diabetes risk as the actual risk, you would think all those genetic variations are correlated with actual diabetes even though they aren't," Lu said.

Compensating for the holes in the data banks with proxies is also problematic. For example, when analyzing the correlation of multiple traits with the risk of developing Alzheimer's disease, Lu observed a divergence from GWAS results, which are based on actual data. A key discrepancy was the association between education attainment and the risk of Alzheimer's disease. Multiple groups have reported an inverse correlation between these variables, a result that is backed by GWAS. However, Lu observed a positive correlation when GWAX approaches were used. The proxy-information approach also failed to show a link between the disease and lower cognition later in life,

contrary to previous data and GWAS findings.

The team proposed new statistical methods that researchers can use to correct these biases and increase the reliability of their findings. They urge the research community to transparently report findings and to adopt a more rigorous and cautious outlook when drawing conclusions from these methods.

“Our group’s recent studies provide humbling examples and highlight the importance of statistical rigor in biobank-scale research studies,” Lu said.

REFERENCES

1. Uffelmann E, et al. Genome-wide association studies. *Nat Rev Methods Primers*. 2021;1(1):59.
2. Wu Y, et al. Pervasive biases in proxy genome-wide association studies based on parental history of Alzheimer’s disease. *Nat Genet*. 2024:1-8.
3. Miao J, et al. Valid inference for machine learning-assisted genome-wide association studies. *Nat Genet*. 2024;56(11):2361-2369.