FACULTEIT GENEESKUNDE EN GEZONDHEIDSWETENSCHAPPEN

DE ROL VAN PREDICTIEMODELLEN IN GWAIS

Emile Van Hecke
01809021

Promotor: Prof. dr. dr. Kristel Van Steen

Masterproef voorgelegd in het kader tot het behalen van de graad Master of Medicine in de Geneeskunde

Academiejaar: 2022 – 2023

UNIVERSITEIT GENT

"De auteur en de promotor geven de toelating dit afstudeerwerk voor consultatie beschikbaar te stellen en delen ervan te kopiëren voor persoonlijk gebruik. Elk ander gebruik valt onder de beperkingen van het auteursrecht, in het bijzonder met betrekking tot de verplichting uitdrukkelijk de bron te vermelden bij het aanhalen van resultaten uit dit afstudeerwerk."

Datum

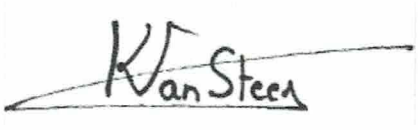
13/11/2022

(handtekening)

Naam (student)

Van Hecke Emile

Prof Dr Dr Kristel VAN STEEN



(promotor)

Inhoudsopgave

Abstract

Sinds de vervolledigiging van het Humane Genome Project in 2003-2005 zijn GWAS studies een populair instrument geworden binnen de genetische epidemiologie. In navolging van de ontwikkelingen in GWAS ontstond GWAIS: dit gebruikt GWAS data en voegt extra informatie in de vorm van interacties toe. Het kan gebruikt kan worden via de ontwikkeling van predictiemodellen, welke ziekte of andere zaken kunnen voorspellen. Deze tak van genetische studies is volop in ontwikkeling en in deze studie wordt gekeken op welke vlakken dit manifesteert. De studie behandelt de mogelijke toepassingsgebieden, de designs, het gebruik van interacties, de predictiemodellen en validatie van interactiestudies en specifiek deze met predictie als doel. Er is een evolutie te merken in de verscheidenheid aan behandelde ziektebeelden, terwijl de studiedesigns veelal case-control blijven. Interacties worden vooralsnog niet courant gebruikt als uitgangspunt van onderzoek. Verder wordt een shift van het gebruik van PRS naar ingewikkeldere modellen die ruwe data rechtstreeks kunnen incorporeren opgemerkt. Ook op vlak van validatie is er een evolutie in richtlijnen omtrent de aanpak, maar deze worden voorlopig zelden toegepast. Verder behandelt deze studie de problemen die optreden bij de toepassing van de ontwikkelde modellen in de kliniek en de vereiste aanpassingen om dit in de toekomst mogelijk te maken.

Inleiding

GWAS (genome-wide association study) data kunnen voor verscheidene doeleinden gebruikt worden. Een voorbeeld hiervan is predictie op individueel vlak, waar deze studie op focust. GWAS data houden onder meer DNA gebaseerde genetische informatie in waarbij gekeken kan worden naar genetische varianten (specifiek Single Nucleotide Polymorphysms (SNPs)) in verschillende individuen om een verband te leggen tussen deze SNPs en een bepaald ziektebeeld (fenotype). Indien dit gebruikt wordt voor predictie kan er een polygenic risk score (PRS) ontwikkeld worden. Deze wordt berekend als de gewogen som van de risico-allelen, waarbij de gewichten gebaseerd zijn op de effect sizes uit de GWAS. Deze gewichten worden ook wel summary data genoemd, aangezien ze het resultaat zijn van ruwe data waarop al een berekening is uitgevoerd (1).

In deze studie wordt als uitgangspunt van predictie GWAIS (genome wide association and interaction study) genomen. Deze gebruiken veelal dezelfde data als GWAS, maar bouwen een extra SNP-SNP interactie (ook wel epistasis genoemd) component in. Biologische interacties zijn individu gebonden en maken deel uit van een interactoom. Dit laatste omvat verschillende types interacties die eiwitten, DNA, RNA,… met elkaar aangaan. Aangezien het effect van SNPs op het fenotype verloopt via een interactoom, is het dus aannemelijk dat de toevoeging van interacties aan predictiemodellen een betere voorspellende waarde kan hebben. Hier gaat het specifiek over interacties die in de context van een uitkomstmaat worden bekeken. Het fenotype van een individu zijn zijn/haar meetbare kenmerken en is het resultaat van het genotype (dus het DNA) en de invloed van de omgeving hierop. Het is een ruim begrip en omvat naast bijvoorbeeld oogkleur en lengte ook de aan- of afwezigheid van ziekten.

De moeilijkheid voor GWAIS ten opzichte van GWAS is het toevoegen van deze interactiecomponent. Hiervoor worden verschillende methoden gebruikt die elk hun voor- en nadelen hebben, maar dit gebied is nog steeds in ontwikkeling. Het heeft ook tijd gekost om predictiemodellen te ontwikkelen die deze data op een nuttige manier kunnen gebruiken.
Het primaire objectief in GWAIS is een beter begrip van de moleculaire grondslagen van het bestudeerde fenotype. Een mogelijks ander objectief zou zijn het voorspellen van het risico dat een individu een bepaalde ziekte ontwikkelt. Dit kan uitgebreid worden naar predictie over de reactie op geneesmiddelen, hoe een bepaalde ziekte tot uiting komt, et cetera. Hierbij tracht men dus een risicomodel te ontwikkelen via GWAS data, gebruik makend van interacties, dat bruikbaar is in een medische context. Translatie naar de kliniek toe van de gevonden resultaten is het finale objectief. Er bestaan verschillende methoden om predictiemodellen op te bouwen die gebruik maken van GWAIS, waardoor er verschillende resultaten bekomen worden. Hierdoor is de translatie moeilijk, aangezien men deze telkens op een andere manier moet uitvoeren. Er zou een combinatie van verschillende modellen gebruikt kunnen worden in de klinische setting, waarbij de voordelen van elk model optimaal benut worden (2).

Vraagstelling:

Is er een evolutie in het gebruik van GWAIS met betrekking tot predictiemodellen over de laatste 10 jaar?

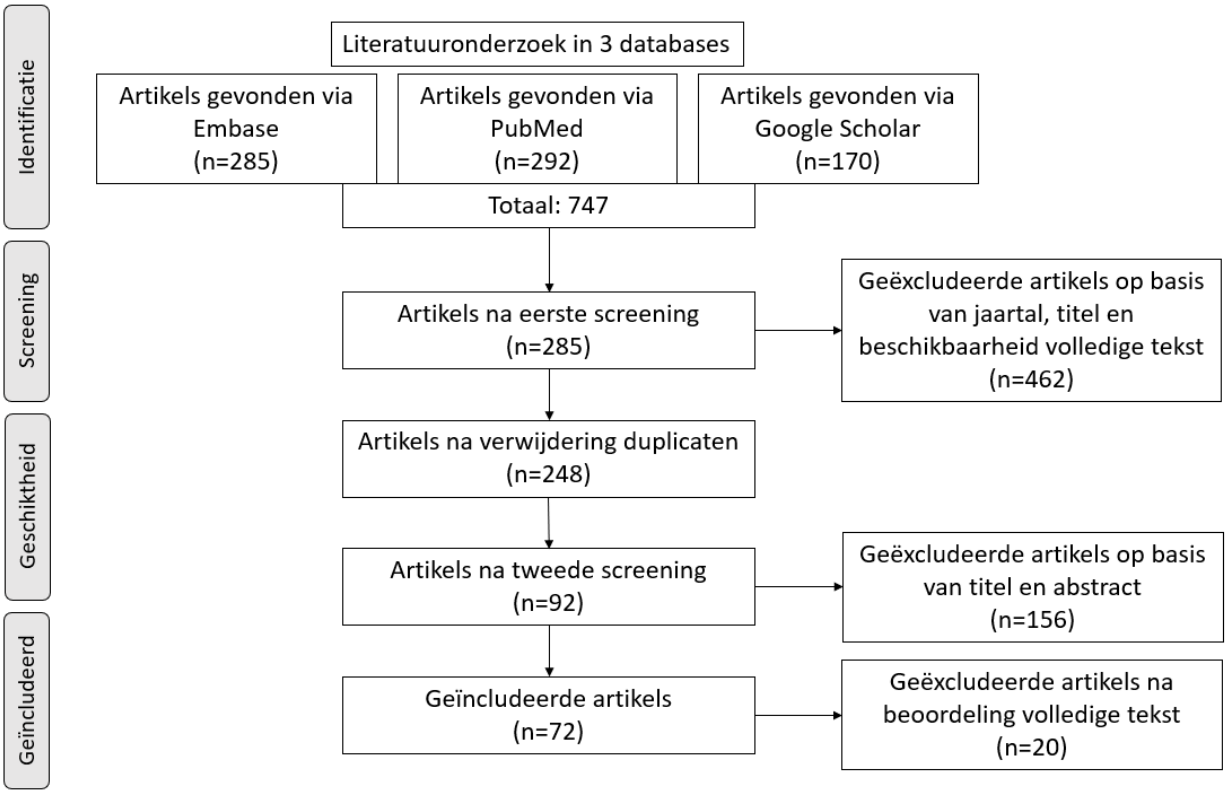
Methodologie:

Selectie artikels:

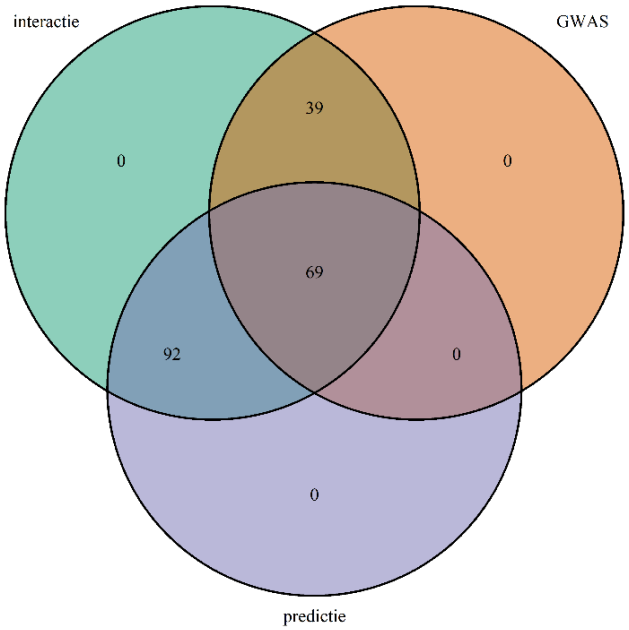
In de periode van 18/09/2021 tot en met 8/11/2021 werd er op Embase, Pubmed en Web of Science gezocht met volgende zoektermen: ‘genome wide association study’ en ‘SNP’ (in figuur 2 ondergebracht onder GWAS); ‘epistasis’, ‘interaction’, ‘gene regulatory networks’ en ‘gene-gene networks’ (figuur 2: ondergebracht onder interacties); ‘prediction’, ‘prediction model’, ‘risk prediction’, ‘polygenic risk score’, ‘risk assessment’, ‘prognosis’ en ‘Bayes Theorem’ (figuur 2: ondergebracht onder predictie); ‘precision medicine’, ‘algorithm’, ‘network’ en ‘immune mediated diseases’ (in figuur 2 niet beschouwd).

Er werden enkel artikels geïncludeerd die in de laatste 10 jaar gepubliceerd werden (dus vanaf 2012). Op basis van de titel werd er verder geselecteerd: artikels over planten/dieren en artikels waarbij de focus niet op epistasis lag werden niet beschouwd. Verder werden ook de meeste artikels over kanker geëxcludeerd, aangezien er anders een overaanbod van deze artikels zou zijn. 4 artikels over kanker werden wel geïncludeerd omdat deze een duidelijk beeld schetsen rond predictiemodellen gebruik makend van epistasis en dit in een verdere stap pas toepassen op kanker. Enkel de artikels waarbij de tekst gevonden werd via Endnote werden verder bekeken. Zo werden uiteindelijk 278 artikels geselecteerd: 112 via Embase, 107 via PubMed en 59 via Google Scholar. Vervolgens werden de duplicaten verwijderd, waardoor er nog 241 artikels overbleven. Ten slotte werd gekeken naar de titel en abstract, waarbij voornamelijk artikels over toepassingen van GWAIS en predictie werden geïncludeerd. Zo bleven nog 85 artikels over.

In de periode van 13/03/2022 tot en met 20/04/2022 werden nog 20 artikels verwijderd uit deze 85, omdat ze na verdere inspectie niet over SNP-SNP interacties gingen. Er werden ook nog 7 artikels toegevoegd uit PubMed na samenspraak met de promotor, waarbij gebruikt werd gemaakt van de zoektermen ‘neural networks’ en ‘epistasis’. In figuur 2 werd de zoekterm ‘neural network’ niet beschouwd. Dezelfde exclusiecriteria werden gebruikt om tot deze 7 artikels te komen en opnieuw werden enkel de artikels beschouwd waarvan de tekst te vinden was via Endnote. In totaal werden dus 72 artikels opgenomen in deze studie (supplement 1).
Figuur 2 toont welke zoektermen het meest gebruikt werden bij de artikels die geëxporteerd werden naar Endnote en waarvan de tekst beschikbaar was. De zoektermen die niet opgenomen zijn in deze figuur werden gebruikt in 85 artikels (figuur 1 en 2).



Figuur 1.
Flow diagram over de identificatie en screening van artikels, beoordeling van geschiktheid en aantal geïncludeerde en geëxcludeerde artikels. Geïnspireerd door Ertaylan et al. (3).



Figuur 2.
Venn-diagram van de belangrijkste zoektermen; zie tekst voor details. Geïnspireerd door Ertaylan et al. (3).

Indeling artikels:

De volgende stap was de indeling van de artikels. Dit gebeurde op jaartal van publicatie, besproken ziekte, type abnormaliteit, studiedesign, type interacties, interactie als uitgangspunt, type predictie, methode voor predictie, besproken confounding en besproken population stratification. Deze indeling werd gemaakt in een Excel bestand (supplement 2).

Voor de indeling in besproken ziekten werd eerst gedetailleerd weergegeven welke ziekte besproken werd in het artikel, met uitzondering van cardiovasculaire ziekten die meteen onder deze term werden ingedeeld. Indien er meerdere of geen ziekten besproken werden, werd dit respectievelijk als 'meerdere ziekten' en 'geen ziekten' ingedeeld. Deze ziekten werden vervolgens ondergebracht in de categorieën onder phenotypic abnormality volgens de website https://hpo.jax.org/app/. Deze indeling is in het Excel bestand te vinden onder 'type abnormaliteit'.

De indeling van het studiedesign gebeurde als volgt: cross-sectioneel, prospectieve cohorte, retrospectieve cohorte, case-control, meta-analyse en niet van toepassing. Het type interacties werd enkel ingedeeld in SNP-SNP en SNP-environment + SNP-SNP. Initieel werden ook andere types interacties (proteïn-proteïn, miRNA-miRNA en SNP-environment zonder bijkomend SNPNP) beschreven, maar deze artikels zijn finaal niet beschouwd in de studie (zie figuur 2 laatste stap). Er werd gekeken of de besproken studies interacties gebruikten als uitgangspunt. Indien dit het geval was werd dit genoteerd als 'ja'. Indien dit niet het geval was en interacties eerder een bijkomend deel waren in plaats van een belangrijk deel van de vraagstelling, werd dit genoteerd als 'neen'. Om deze indeling te maken werd voornamelijk gekeken naar de titel en het abstract, en indien het hieruit moeilijk af te leiden was ook naar het volledige artikel.

Het type predictie werd onderverdeeld in 'development only', 'validation only', 'development and validation in the same publication' en 'niet van toepassing'. Dit gebeurde volgens de methode aangeraden door Wolff et al (4): 'development only' indien het artikel focust op het creëren van een model door een of meerdere nieuwe predictors toe te voegen, 'validation only' indien er enkel validatie van een al bestaand model gebeurde en 'development and validation in the same publication' indien een model werd ontwikkeld dat bovendien validatie onderging of indien een bestaand model gevalideerd en aangepast/uitgebreid werd. De methode voor predictie werd steeds gezocht in de sectie methods van de beschreven artikels. Indien er geen risicoscore werd opgebouwd, werd dit aangeduid als 'geen risk score opgebouwd'. Om te detecteren of artikels iets

over predictie vermeldden, werd voornamelijk op volgende zoektermen gelet: prediction, predictor, risk score, risk (factor), PRS en varianten van deze termen.

Voor de indeling van de besproken confounding en population stratification werd er voornamelijk op de volgende zoektermen gelet: confounding, confounders, confounded, covariates, bias, population stratification, population (sub)structure, admixture, shared genetic ancestry en andere varianten of synoniemen van deze termen. Indien deze termen voorkwamen, werden de betreffende secties integraal uitgelicht voor verdere verwerking.

Software:

De figuren werden aangemaakt via RStudio versie 4.1.2, in het bijzonder via de R-paketten RColorBrewer versie 1.1-3, ggplot versie 4.1.3, dplyr versie 1.0.10 en VennDiagram versie 1.7.3. De data gebruikt in R werd meestal uit het Excelbestand gehaald en soms toegevoegd vanuit eigen notities (supplementen 2 en 3).

Resultaten en discussie:

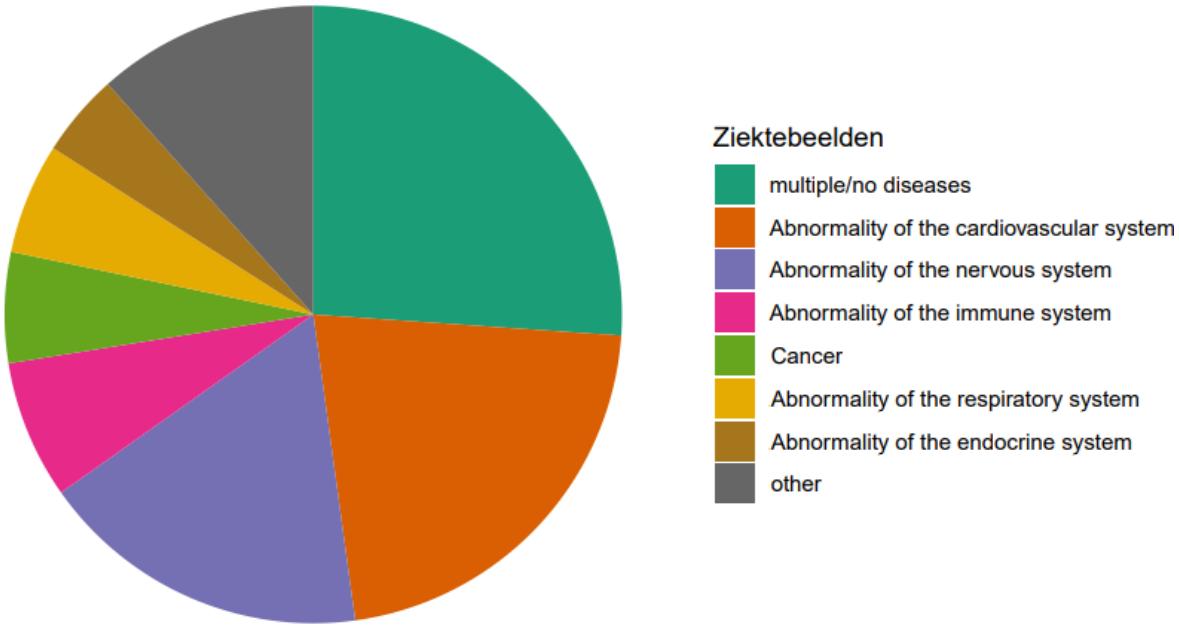
Er werd gekeken naar de manier waarop verschillende onderdelen (ziektebeelden, design, interacties, predictiemodellen en validatie) evolueren in de tijd.

Ziektebeelden:

In figuur 3 is te zien dat de onderzochte ziektebeelden in interactiestudies waarbij gefocust wordt op predictie voornamelijk cardiovasculaire ziekten en ziekten van het zenuwstelsel zijn. De studies over cardiovasculaire ziekten gingen voornamelijk over het risico op cardiovasculaire events en hartinfarcten. De studies over ziekten van het zenuwstelsel bevatten een bredere groep ziekten, met voornamelijk de ziekte van Alzheimer, schizofrenie en depressie.

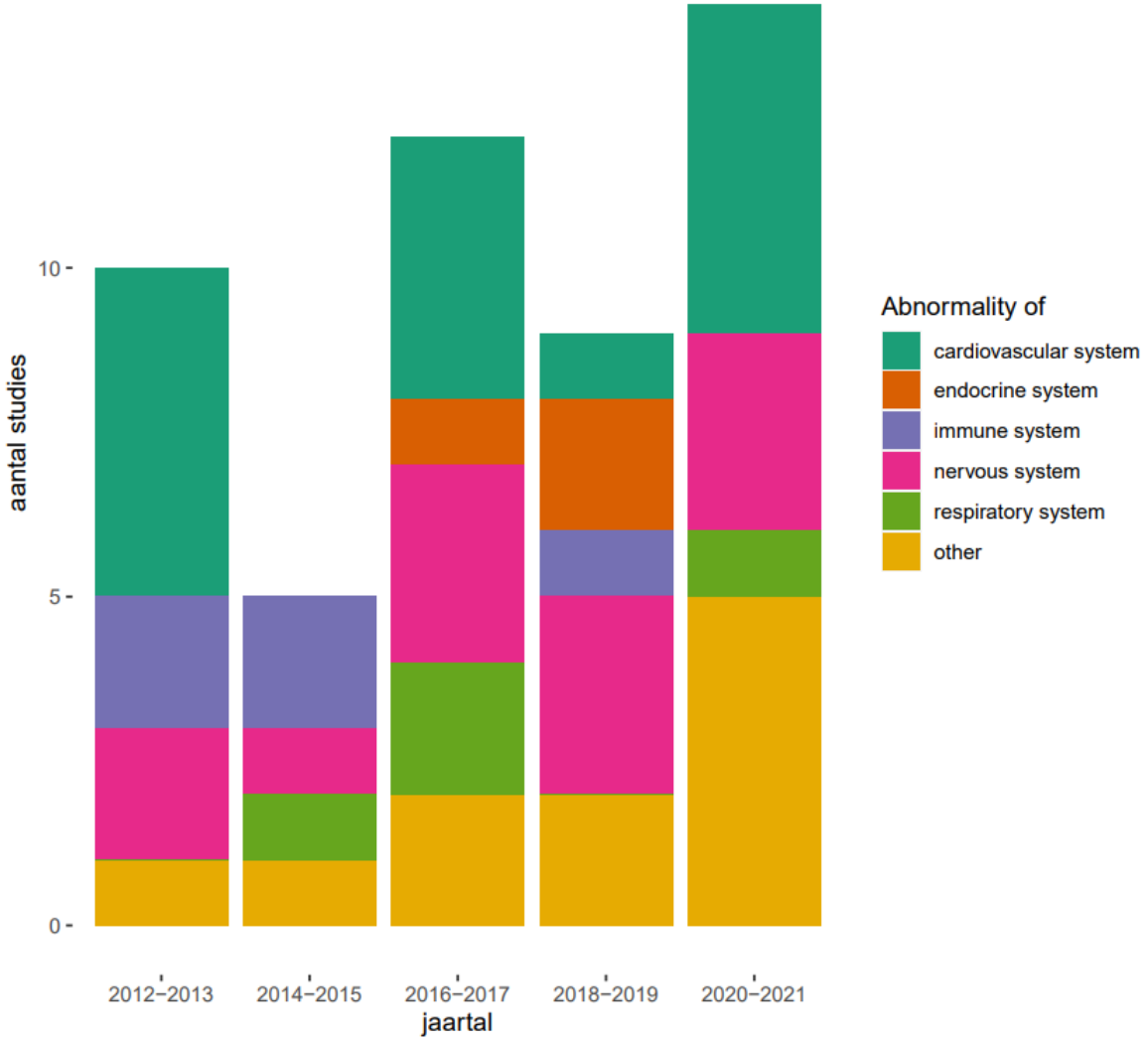
Onder ‘multiple/no diseases’ vallen de studies waarbij men onderzoek doet naar algemene zaken in verband met interactie en predictie, en waarbij men deze soms toepast op meerdere ziektebeelden. Over kanker kan er uiteraard niets gezegd worden, aangezien artikels hierover op voorhand grotendeels uitgefilterd werden. Er is een grote verscheidenheid aan besproken ziekten: 33 verschillende pathologieën werden besproken (multiple/no diseases niet meegerekend). Het valt te begrijpen dat cardiovasculaire ziekten een populair onderwerp zijn, aangezien deze een

zeer grote groep patiënten bevat en deze een groot aandeel hebben in de mortaliteit in westerse landen (5).



Figuur 3.
Piechart van de besproken ziektebeelden.

Figuur 4 toont de evolutie van de besproken ziektebeelden. Over de grote categorieën valt niet veel te zeggen, maar er is wel te zien dat interactiestudies in steeds meer verschillende domeinen (te zien als het deel other dat toeneemt) hun intrede doen. Other bevat onder andere obesitas, prematuriteit en zelfs de respons op pokkenvaccinatie. Dit toont aan dat steeds meer domeinen in aanraking komen met epistasis en predictiemodellen die hierop gebaseerd zijn. Een van de manieren waarop deze predictiemodellen een nuttige tool kunnen zijn, is via de respons op behandeling. Dit wordt al vaak toegepast en er is binnen het vakgebied veel interesse voor. 13 van de 72 besproken studies gingen hier dan ook over.

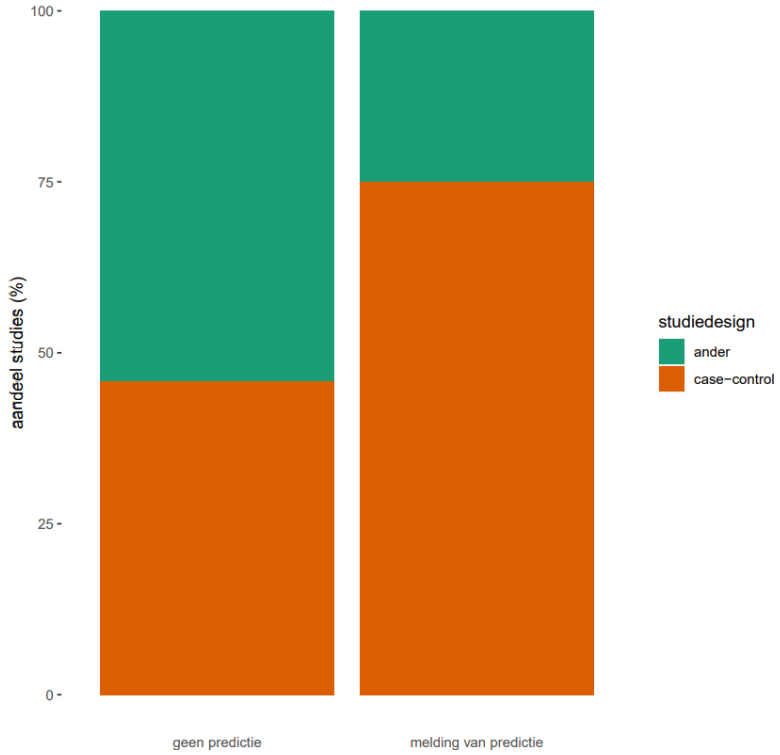


Figuur 4.
Barplot: evolutie in besproken ziektebeelden.

Design:

Het grootste deel van de besproken artikels gebruikt een case-control studiedesign: 47 artikels met tegenover 25 zonder case-control studiedesign. Dit komt voornamelijk omdat de meeste SNP-SNP interactiestudies voortbouwen op GWAS designs. Deze GWAS komen voornamelijk tot stand via case-control. Het is namelijk vrij eenvoudig om de verschillen in SNPs te ontdekken tussen een groep met een bepaald kenmerk en een groep zonder dat kenmerk (1). Indien er interactiedata aan toegevoegd worden, blijft dit gelden. Dit argument kan uiteraard ook toegepast worden op prospectieve cohortestudies (studiedesign gebruikt in 8 artikels), maar deze zijn vaak moeilijker op te zetten.

In figuur 5 is te zien dat wanneer het om predictie gaat, het case-control design nog pertinenter is Hierbij dient wel vermeld te worden dat ‘melding van predictie’ ruim werd genomen. Het wijst er niet op dat de studie tot doel had een predictiemodel te maken, maar dat er op zijn minst werd gekeken naar de mogelijkheid dat het onderzoek en/of resultaat iets kon betekenen voor de predictie van de onderzochte ziekte, het antwoord op medicatie, et cetera. Niet te verwonderen aangezien de meeste GWAIS geijkt zijn op GWAS en als iets secundairs beschouwd worden: eerst wordt een GWAS opgezet, en vervolgens worden hier interacties aan toegevoegd.



Figuur 5.
Barplot: aandeel studies dat een case-control design heeft, ingedeeld in studies die predictie vermelden en studies die dat niet doen.

De meeste analytische modellen binnen GWAS zijn associatiemodellen (dus regressiemodellen): verklarende factoren worden gekoppeld aan een fenotype (6). In GWAIS is de opzet om via gen-gen interacties een beter inzicht te krijgen in de moleculaire mechanismen onderliggend aan de ziekten waardoor dezelfde modellen gebruikt kunnen worden. Veelal gebruiken deze associatiemodellen niet gerelateerde individuen vanuit pragmatisch en economisch oogpunt. Daarom blijven case-controle studies populair ongeacht of het objectief verklaren/begrijpen (associatiemodellen) of predictie (predictiemodellen) is.

Interacties:

In deze studie werd gekeken of interacties als primair doel werden beschouwd in de artikels, wat bij 20 artikels niet het geval was. Deze hebben veelal als primaire doel onderzoek te voeren naar de oorzaken of predictie van bepaalde ziekten, waarbij epistasis dan werd gebruikt als een extra component in het verklaringsmodel of een extra factor waar rekening mee gehouden werd in het predictiemodel. In deze studie werd voornamelijk gekeken naar de onderdelen waar het effectief over epistasis ging, en hoe dit dan aangepakt en gebruikt werd. Het is opvallend dat er nog een vrij groot aandeel studies interacties niet als primaire doel beschouwen, hoewel hier de nadruk op werd gelegd bij het zoeken van artikels (figuur 1). Dit toont aan dat interactiestudies veel minder frequent zijn dan GWAS: ze zijn zelden een doel op zich. Dit kan onder andere liggen aan het ontbreken van duidelijke richtlijnen voor deze studies (waar deze wel bestaan voor klassieke GWAS), de complexiteit van deze studies en de computationele kracht die de vele mogelijkheden vereisen. Dit valt ook te merken aan de vele studies die onderzoek voeren naar de interactie tussen omgeving en SNPs om tot een verklaringsmodel te komen. Deze werden veelvuldig tegengekomen bij het zoeken naar artikels voor deze studie. Dit type interacties heeft het voordeel dat er veel minder testen moeten uitgevoerd worden, aangezien er vaak slechts enkele omgevingsfactoren in rekening worden gebracht in tegenstelling tot de vele SNPs die paarsgewijs dienen onderzocht te worden. De berekeningen bij gen-omgevingsfactoren zijn een stuk eenvoudiger uit te voeren. Bovendien is het niet zo dat de toevoeging van interacties aan een predictiemodel steeds een betere predictie oplevert. Dit hangt natuurlijk ook af van welk predictiemodel wordt gebruikt en hoe de interactiedata verwerkt worden.

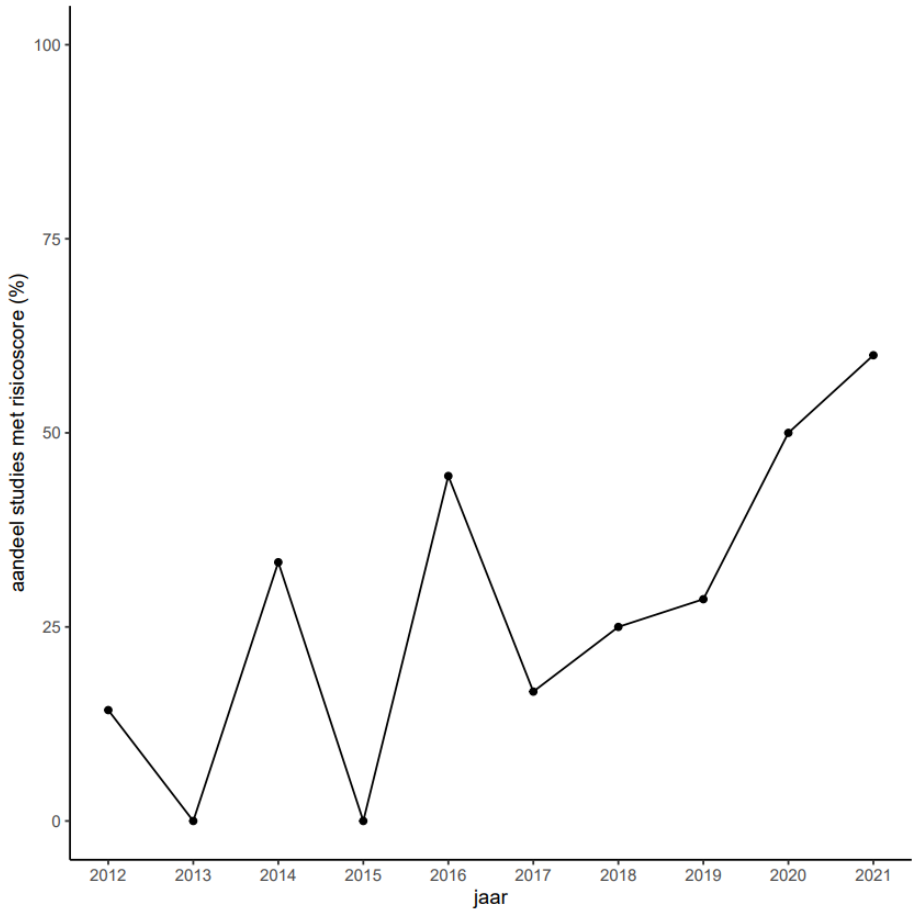
Predictiemodellen:

Er zijn meerdere manieren om tot predictie te komen. Zo kan er een risicoscore opgebouwd worden, maar kan er ook gefocust worden op de predictieve waarde van 1 of enkele SNPs, omgevingsfactoren, et cetera. Van de 48 onderzochte artikels die een melding maakten van predictie, werd er bij 21 een risicoscore opgesteld. Een van de manieren om dit op te stellen is via de opbouw van een PRS, die, zoals vermeld in de inleiding, gebruik maakt van summary data. Er zijn echter ook andere manieren om tot predictie te komen die ruwe data (dus data waarop nog geen berekening op uitgevoerd is) gebruiken. 15 van de 21 artikels met een risicoscore ontwikkelden een PRS. Het is nog helemaal niet duidelijk of PRS de beste manier is om tot predictie te komen, en hier is ook nog veel discussie over (7, 8). Er zijn ondertussen ook nieuwe manieren ontwikkeld om risicoscores te berekenen die op een efficiënte en effectieve manier interactiedata

gebruiken, zoals bijvoorbeeld de multilocus risk score (MRS) ontwikkeld door Le et al., dat een soort uitbreiding is van de klassieke PRS (9).

Om tot een risicoscore te komen kunnen meerdere methoden gebruikt worden. Voornamelijk regressiemodellen (lineaire en logistische), deep learning en neural networks werden hiervoor gebruikt. Pas sinds 2016 worden de ingewikkeldere methoden (deep learning en neurale netwerken) toegepast, maar in aantal blijven ze vooralsnog beperkt (4 in totaal). Deze hebben het voordeel dat ze ontwikkeld werden om tot predictie te komen. Ze kunnen dan ook veel variabelen met relatief weinig individuen aan, wat een groot voordeel is bij het gebruik van GWAIS. Het nadeel aan de ingewikkeldere methoden is dat men minder zicht heeft op wat erachter schuilt, en ze lopen dus het risico bias te missen (zie paragraaf validatie). Bovendien zijn ze ook minder gebruiksvriendelijk. Deze methoden zijn pas recenter ontwikkeld en vooralsnog minder bekend. Regressiemodellen werken eerder beschrijvend: ze kunnen gebruikt worden om nieuwe mechanismen te ontdekken maar werden niet ontwikkeld om tot predictiemodellen te komen. Ze zijn echter relatief gemakkelijk te gebruiken en er bestaan ook vele handleidingen hoe deze gebruikt kunnen worden, voornamelijk om tot een PRS te komen (6).

Figuur 6 toont aan dat de laatste jaren het aandeel artikels dat een risicoscore opbouwt, gebruik makend van GWAIS, toeneemt. Dit kan onder andere een gevolg zijn van de evolutie in onderzoek naar de aanpak ervan, alsook de evolutie in predictiemodellen en in de hoeveelheid beschikbare gegevens. Deze trend is voornamelijk te zien in predictiemodellen die geen gebruik maken van PRS: de eerste studies waarin predictiemodellen werden ontwikkeld die rechtstreeks gebruik maken van ruwe data zijn pas in 2016 gepubliceerd, en sindsdien zijn 6 artikels gepubliceerd die dit ontwikkelden. Ook dit is waarschijnlijk het gevolg van de technische vooruitgang dankzij de ontwikkeling van neurale netwerken en deep learning: deze kunnen de ruwe data meteen incorporeren. Of deze modellen tot betere predictie komen is nog niet duidelijk (7, 8). Door de heterogeniteit op vlak van ziektebeelden is het mogelijk dat verschillende modellen in verschillende disciplines toegepast zullen worden. Idealiter zou een combinatie van verschillende modellen worden toegepast, waarbij de voordelen van elk model worden benut. Deze implementatie is echter nog onvoldoende onderzocht en is waarschijnlijk een volgende belangrijke stap in het gebruik van GWAIS (2).



Figuur 6.
Aandeel studies dat een risicoscore opbouwt per jaar.

Validatie:

Er waren 47 artikels die rekening hielden met confounding. Een voorbeeld van een confounder die we kennen is population stratification of admixture. Dit houdt er rekening mee dat verschillen in allelfrequenties kunnen ontstaan door systematische verschillen in voorouders in plaats van associatie van genen met ziekte. Het kan voor bias zorgen indien er onderzoek gebeurt in verschillende populaties. Daarom is het een belangrijke factor om rekening mee te houden bij GWAS. Er zijn dan ook manieren ontwikkeld om hier mee om te gaan in GWAS, maar in GWAIS zijn hier nog niet voldoende oplossingen voor en wordt het probleem ook onderschat, al zijn er recent wel methoden ontwikkeld om hier mee om te gaan (10-12).

Idealiter zou confounding een item zijn in alle artikels. Van de artikels die een risicoscore opmaken waren er 4 die dit niet benoemden, hoewel hun resultaten natuurlijk te betwijfelen vallen indien men hier geen rekening mee hield. Zeker wanneer gewerkt wordt met regressiemodellen (2 van de 4 artikels met een risicoscore en zonder melding van confounding) valt dit vrij gemakkelijk te

incorporeren in het model. Indien gebruik wordt gemaakt van deep learning modellen of neurale netwerken, is er veel minder zicht op confounders en zijn deze ook moeilijker te incorporeren. 2 van de 4 artikels die neurale netwerken of deep learning modellen gebruiken hebben niets vermeld over confounding.

Wanneer confounding dan wel vermeld werd, was dit vaak beperkt. Nochtans bestaan hier tools voor (bijvoorbeeld QUIPS, Chochrane ROB tool, ROBINS-I en PROBAST) waarmee je op een systematische manier bias kan nagaan. PROBAST is ontwikkeld voor meta-analyses, waarmee deze de ROB (risk of bias) van de besproken artikels in kaart kunnen brengen. Het kan echter ook gebruikt worden bij de ontwikkeling van een risicoscore of de bespreking van risicofactoren om een idee te hebben met welke bias rekening moet gehouden worden. In deze studie werd PROBAST niet gebruikt aangezien dit te ver zou leiden en dit moeilijk uit te voeren is zonder de vereiste ervaring (4). Choi et al. beschrijven ook methoden om kwaliteitscontrole van de data uit te voeren (6). Deze zijn soms terug te vinden in de artikels die een risicoscore opmaken, maar het is zelden duidelijk of alle stappen zijn doorlopen om bias zo veel mogelijk te beperken. Belangrijk hierbij is dat deze studies pas in 2019 en 2020 gepubliceerd werden, waardoor het voor artikels hiervoor onmogelijk was om deze te gebruiken en het misschien nog niet bekend genoeg was om te gebruiken in de recentere artikels.

Over het algemeen werd bias werden niet voldoende geadresseerd in deze artikels en indien ze wel geadresseerd werden, gebeurde dit nog niet op een systematische wijze. Hierdoor is er een groot risico dat er met sommige bias geen rekening gehouden werd. Het is dan ook weinig waarschijnlijk dat de ontwikkelde modellen en gevonden gegevens voldoende toepasbaar zijn in de klinische setting en al zeker niet over de hele wereld.

Conclusie:

Predictiemodellen gebaseerd op GWAIS hebben in theorie zeer veel toepassingen en kunnen op verschillende gebieden een rol spelen, ook in de respons op medicatie, die vaak erg verschilt tussen individuen, kan dit een verklaringsmodel creëren. Dit vakgebied is volop in ontwikkeling, met onder andere de betrekking van meerdere ziektebeelden, de ontwikkelingen en het gebruik van nieuwe predictiemodellen en evoluties in de validatie van deze modellen. Er is echter nog veel

ruimte voor verbetering: de optimalisatie van predictiemodellen, de ontwikkeling van nieuwe modellen en richtlijnen rond pragmatische implementatie, validatie, de combinatie van verschillende methodologische gezichtspunten en consensus besluitvorming in het kader van precisiegeneeskunde.

Referenties:

1. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nature Reviews Methods Primers. 2021;1(1):59.
2. Hao L, Kraft P, Berriz GF, Hynes ED, Koch C, Kumar KV, et al. Development of a clinical polygenic risk score assay and reporting workflow. Nature medicine. 2022;28(5):1006-13.
3. Ertaylan G, Le Cornet C, Van Roekel EH, Jung AY, Bours MJ, Damms-Machado A, et al. A Comparative Study on the WCRF International/University of Bristol Methodology for Systematic Reviews of Mechanisms Underpinning Exposure–Cancer AssociationsA Comparative Study on Systematic Review of Mechanisms. Cancer Epidemiology, Biomarkers & Prevention. 2017;26(11):1583-94.
4. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med. 2019;170(1):51-8.
5. Wilkins E, Wilson L, Wickramasinghe K, Bhatnagar P, Leal J, Luengo-Fernandez R, et al. European cardiovascular disease statistics 2017. 2017.
6. Choi SW, Mak TS-H, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nature protocols. 2020;15(9):2759-72.
7. Cope JL, Baukmann HA, Klinger JE, Ravarani CN, Böttinger EP, Konigorski S, et al. Interaction-Based Feature Selection Algorithm Outperforms Polygenic Risk Score in Predicting Parkinson's Disease Status. Frontiers in Genetics. 2021;12.
8. Gola D, Erdmann J, Müller-Myhsok B, Schunkert H, König IR. Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status. Genetic epidemiology. 2020;44(2):125-38.
9. Le TT, Gong H, Orzechowski P, Manduchi E, Moore JH, editors. Expanding Polygenic Risk Scores to Include Automatic Genotype Encodings and Gene-gene Interactions. BIOINFORMATICS; 2020.
10. Abegaz F, Van Lishout F, John JMM, Chiachoompu K, Bhardwaj A, Gusareva ES, et al. Epistasis Detection using Model Based Multifactor Dimensionality Reduction in Structured Populations. bioRxiv. 2019:541946.
11. Abegaz F, Van Lishout F, Mahachie John JM, Chiachoompu K, Bhardwaj A, Duroux D, et al. Performance of model-based multifactor dimensionality reduction methods for epistasis detection by controlling population structure. BioData Mining. 2021;14(1).
12. Abegaz F, Van Lishout F, Mahachie John JM, Chiachoompu K, Bhardwaj A, Gusareva ES, et al. Epistasis detection in genome-wide screening for complex human diseases in structured populations. Systems Medicine. 2019;2(1):19-27.

Supplement 1.
Oplijsting van alle weerhouden artikels corresponderend met figuur 1.

1. Adams SM, Harralson AF, Feroze H, Nguyen T, Eum S, Cornelio C. Genome wide epistasis study of on-statin cardiovascular events with iterative feature reduction and selection. Journal of Personalized Medicine. 2020;10(4):1-14.
2. Ahluwalia TS, Eliasen AU, Sevelsted A, Pedersen CET, Stokholm J, Chawes B, et al. FUT2–ABO epistasis increases the risk of early childhood asthma and Streptococcus pneumoniae respiratory illnesses. Nature Communications. 2020;11(1).
3. Arning A, Hiersche M, Witten A, Kurlemann G, Kurnik K, Manner D, et al. A genome-wide association study identifies a gene network of ADAMTS genes in the predisposition to pediatric stroke. Blood. 2012;120(26):5231-6.
4. Aslibekyan S, Goodarzi MO, Frazier-Wood AC, Yan XF, Irvin MR, Kim E, et al. Variants Identified in a GWAS Meta-Analysis for Blood Lipids Are Associated with the Lipid Response to Fenofibrate. Plos One. 2012;7(10).
5. Beam AL, Motsinger-Reif A, Doyle J. Bayesian neural networks for detecting epistasis in genetic association studies. BMC bioinformatics. 2014;15(1):368.
6. Chang SH, Fang KC, Zhang KL, Wang J. Network-Based Analysis of Schizophrenia Genome-Wide Association Data to Detect the Joint Functional Association Signals. Plos One. 2015;10(7).
7. Chattopadhyay A, Lu TP. Gene-gene interaction: The curse of dimensionality. Annals of Translational Medicine. 2019;7(24).
8. Chicco D, Faultless T. Brief Survey on Machine Learning in Epistasis. 2021. p. 169-79.
9. Chiesa A, Lia L, Lia C, Lee SJ, Han C, Patkar AA, et al. Investigation of possible epistatic interactions between GRIA2 and GRIA4 variants on clinical outcomes in patients with major depressive disorder. J Int Med Res. 2013;41(3):809-15.
10. Cong W, Meng X, Li J, Zhang Q, Chen F, Liu W, et al. Genome-wide network-based pathway analysis of CSF t-tau/Aβ1-42 ratio in the ADNI cohort. BMC genomics. 2017;18(1):421.
11. de Oliveira FF, Chen ES, Smith MC, Bertolucci PHF. Selected LDLR and APOE Polymorphisms Affect Cognitive and Functional Response to Lipophilic Statins in Alzheimer's Disease. J Mol Neurosci. 2020;70(10):1574-88.
12. El-Lebedy D. Interaction between endothelial nitric oxide synthase rs1799983, cholesteryl ester-transfer protein rs708272 and angiopoietin-like protein 8 rs2278426 gene variants highly elevates the risk of type 2 diabetes mellitus and cardiovascular disease. Cardiovasc Diabetol. 2018;17(1):97.
13. Fergus P, Montanez CC, Abdulaimma B, Lisboa P, Chalmers C, Pineles B. Utilizing Deep Learning and Genome Wide Association Studies for Epistatic-Driven Preterm Birth Classification in African-American Women. IEEE/ACM transactions on computational biology and bioinformatics. 2020;17(2):668-78.
14. Fernández-Santiago R, Martín-Flores N, Antonelli F, Cerquera C, Moreno V, Bandres-Ciga S, et al. SNCA and mTOR Pathway Single Nucleotide Polymorphisms Interact to Modulate the Age at Onset of Parkinson's Disease. Mov Disord. 2019;34(9):1333-44.
15. Franco NR, Massi MC, Ieva F, Manzoni A, Paganoni AM, Zunino P, et al. Development of a method for generating SNP interaction-aware polygenic risk scores for radiotherapy toxicity. Radiotherapy and Oncology. 2021;159:241-8.
16. Freytag S, Manitz J, Schlather M, Kneib T, Amos CI, Risch A, et al. A network-based kernel machine test for the identification of risk pathways in genome-wide association studies. Human heredity. 2013;76(2):64-75.

17. Fung JN, Holdsworth-Carson SJ, Sapkota Y, Zhao ZZ, Jones L, Girling JE, et al. Functional evaluation of genetic variants associated with endometriosis near GREB1. Human Reproduction. 2015;30(5):1263-75.
18. Guan L, Wang Q, Wang L, Wu B, Chen Y, Liu F, et al. Common variants on 17q25 and gene-gene interactions conferring risk of schizophrenia in Han Chinese population and regulating gene expressions in human brain. Molecular Psychiatry. 2016;21(9):1244-50.
19. Guo M, Guo G, Ji X. Genetic polymorphisms associated with heart failure: A literature review. J Int Med Res. 2016;44(1):15-29.
20. Han S, Yang BZ, Kranzler HR, Liu X, Zhao H, Farrer LA, et al. Integrating GWASs and human protein interaction networks identifies a gene subnetwork underlying alcohol dependence. American journal of human genetics. 2013;93(6):1027-34.
21. Hillenmeyer S, Davis LK, Gamazon ER, Cook EH, Cox NJ, Altman RB. STAMS: STRING-assisted module search for genome wide association studies and application to autism. Bioinformatics (Oxford, England). 2016;32(24):3815-22.
22. Hong EP, Heo SG, Park JW. The liability threshold model for predicting the risk of cardiovascular disease in patients with type 2 diabetes: A multi-cohort study of korean adults. Metabolites. 2021;11(1):1-15.
23. Jabandziev P, Smerek M, Michalek J, Fedora M, Kosinova L, Hubacek JA, et al. Multiple gene-to-gene interactions in children with sepsis: a combination of five gene variants predicts outcome of life-threatening sepsis. Crit Care. 2014;18(1):R1.
24. Jiang X, Cai BH, Xue DY, Lu XH, Cooper GF, Neapolitan RE. A comparative analysis of methods for predicting clinical outcomes using high-dimensional genomic datasets. Journal of the American Medical Informatics Association. 2014;21(E2):E312-E9.
25. Jiang X, Neapolitan RE. Evaluation of a two-stage framework for prediction using big genomic data. Briefings in Bioinformatics. 2015;16(6):912-21.
26. Kafaie S, Chen Y, Hu T. A network approach to prioritizing susceptibility genes for genome-wide association studies. Genetic Epidemiology. 2019;43(5):477-91.
27. Keaton JM, Hellwege JN, Ng MC, Palmer ND, Pankow JS, Fornage M, et al. Genome-Wide Interaction with Insulin Secretion Loci Reveals Novel Loci for Type 2 Diabetes in African Americans. PloS one. 2016;11(7):e0159977.
28. Koo CL, Liew MJ, Mohamad MS, Salleh AH. A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology. BioMed research international. 2013;2013:432375.
29. Le TT, Urbanowicz RJ, Moore JH, McKinney BA. STatistical Inference Relief (STIR) feature selection. Bioinformatics (Oxford, England). 2019;35(8):1358-65.
30. Lee KY, Leung KS, Tang NLS, Wong MH. Discovering Genetic Factors for psoriasis through exhaustively searching for significant second order SNP-SNP interactions. Scientific reports. 2018;8(1):15186.
31. Li T, Zhang X, Sang L, Li XT, Sun HY, Yang J, et al. The interaction effects between TLR4 and MMP9 gene polymorphisms contribute to aortic aneurysm risk in a Chinese Han population. BMC Cardiovasc Disord. 2019;19(1):72.
32. Li X, Liu L, Zhou J, Wang C. Heterogeneity Analysis and Diagnosis of Complex Diseases Based on Deep Learning Method. Scientific reports. 2018;8(1):6155.
33. Li Y, Cho H, Wang F, Canela-Xandri O, Luo C, Rawlik K, et al. Statistical and Functional Studies Identify Epistasis of Cardiovascular Risk Genomic Variants From Genome-Wide Association Studies. J Am Heart Assoc. 2020;9(7):e014146.
34. Liao SY, Lin XH, Christiani DC. Genome-wide Association and Network Analysis of Lung Function in the Framingham Heart Study. Genetic Epidemiology. 2014;38(6):572-8.

35. Linseman T, Soubeyrand S, Martinuk A, Nikpay M, Lau P, McPherson R. Functional Validation of a Common Nonsynonymous Coding Variant in ZC3HC1 Associated With Protection From Coronary Artery Disease. Circ Cardiovasc Genet. 2017;10(1).
36. Liou YJ, Bai YM, Lin E, Chen JY, Chen TT, Hong CJ, et al. Gene-gene interactions of the INSIG1 and INSIG2 in metabolic syndrome in schizophrenic patients treated with atypical antipsychotics. Pharmacogenomics J. 2012;12(1):54-61.
37. Liu Y, Brossard M, Roqueiro D, Margaritte-Jeannin P, Sarnowski C, Bouzigon E, et al. SigMod: an exact and efficient method to identify a strongly interconnected disease-associated module in a gene network. Bioinformatics (Oxford, England). 2017;33(10):1536-44.
38. Lynch AI, Tang W, Shi G, Devereux RB, Eckfeldt JH, Arnett DK. Epistatic effects of ACE I/D and AGT gene variants on left ventricular mass in hypertensive patients: the HyperGEN study. J Hum Hypertens. 2012;26(2):133-40.
39. Lyu R, Sun J, Xu D, Jiang Q, Wei C, Zhang Y. GESLM algorithm for detecting causal SNPs in GWAS with multiple phenotypes. Briefings in bioinformatics. 2021.
40. Ma LJ, Chandel N, Ermel R, Sukhavasi K, Hao K, Ruusalepp A, et al. Multiple independent mechanisms link gene polymorphisms in the region of ZEB2 with risk of coronary artery disease. Atherosclerosis. 2020;311:20-9.
41. Ma YR, Zhao SX, Li L, Sun F, Ye XP, Yuan FF, et al. A weighted genetic risk score using known susceptibility variants to predict Graves disease risk. Journal of Clinical Endocrinology and Metabolism. 2019;104(6):2121-30.
42. Man M, Close SL, Shaw AD, Bernard GR, Douglas IS, Kaner RJ, et al. Beyond single-marker analyses: mining whole genome scans for insights into treatment responses in severe sepsis. Pharmacogenomics J. 2013;13(3):218-26.
43. Manso H, Krug T, Sobral J, Albergaria I, Gaspar G, Ferro JM, et al. Evidence for epistatic gene interactions between growth factor genes in stroke outcome. Eur J Neurol. 2012;19(8):1151-3.
44. McKinney BA, Lareau C, Oberg AL, Kennedy RB, Ovsyannikova IG, Poland GA. The integration of epistasis network and functional interactions in a GWAS implicates RXR pathway genes in the immune response to smallpox vaccine. PLoS ONE. 2016;11(8).
45. Naushad SM, Janaki Ramaiah M, Pavithrakumari M, Jayapriya J, Hussain T, Alrokayan SA, et al. Artificial neural network-based exploration of gene-nutrient interactions in folate and xenobiotic metabolic pathways that modulate susceptibility to breast cancer. Gene. 2016;580(2):159-68.
46. Nazarian A, Sichtig H, Riva A. A Knowledge-Based Method for Association Studies on Complex Diseases. Plos One. 2012;7(9).
47. Padmanabhan S, Joe B. TOWARDS PRECISION MEDICINE FOR HYPERTENSION: A REVIEW OF GENOMIC, EPIGENOMIC, AND MICROBIOMIC EFFECTS ON BLOOD PRESSURE IN EXPERIMENTAL RAT MODELS AND HUMANS. Physiological Reviews. 2017;97(4):1469-528.
48. Pei G, Sun H, Dai Y, Liu X, Zhao Z, Jia P. Investigation of multi-trait associations using pathway-based analysis of GWAS summary statistics. BMC genomics. 2019;20(Suppl 1):79.
49. Previde P, Thomas B, Wong M, Mallory EK, Petkovic D, Altman RB, et al. GeneDive: A gene interaction search and visualization tool to facilitate precision medicine. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing. 2018;23:590-601.
50. Qian Y, Besenbacher S, Mailund T, Schierup MH. Identifying disease associated genes by network propagation. BMC systems biology. 2014;8 Suppl 1:S6.
51. Rankinen T, Sarzynski MA, Ghosh S, Bouchard C. Are there genetic paths common to obesity, cardiovascular disease outcomes, and cardiovascular risk factors? Circ Res. 2015;116(5):909-22.

52. Ritchie MD. The success of pharmacogenomics in moving genetic association studies from bench to bedside: study design and implementation of precision medicine in the post-GWAS era. Hum Genet. 2012;131(10):1615-26.
53. Sabik OL, Calabrese GM, Taleghani E, Ackert-Bicknell CL, Farber CR. Identification of a Core Module for Bone Mineral Density through the Integration of a Co-expression Network and GWAS Data. Cell Rep. 2020;32(11):108145.
54. Shang L, Smith JA, Zhou X. Leveraging gene co-expression patterns to infer trait-relevant tissues in genome-wide association studies. PLoS genetics. 2020;16(4):e1008734.
55. Shen YL, Long SY, Kong WM, Wu LM, Fei LJ, Yao Q, et al. Single-nucleotide polymorphisms in genes predisposing to leprosy in leprosy household contacts in Zhejiang Province, China. Pharmacogenomics and Personalized Medicine. 2020;13:767-73.
56. Suppiah V, Armstrong NJ, O'Connor KS, Berg T, Weltman M, Abate ML, et al. CCR5-Δ32 genotype does not improve predictive value of IL28B polymorphisms for treatment response in chronic HCV infection. Genes and immunity. 2013;14(5):286-90.
57. Tefferi A, Guglielmelli P, Nicolosi M, Mannelli F, Mudireddy M, Bartalucci N, et al. GIPSS: genetically inspired prognostic scoring system for primary myelofibrosis. Leukemia. 2018;32(7):1631-42.
58. Tyler A, Matthew Mahoney J, Carter GW. Genetic interactions affect lung function in patients with systemic sclerosis. G3: Genes, Genomes, Genetics. 2020;10(1):151-63.
59. Upstill-Goddard R, Eccles D, Fliege J, Collins A. Machine learning approaches for the discovery of gene-gene interactions in disease data. Briefings in bioinformatics. 2013;14(2):251-60.
60. Uzun A, Dewan AT, Istrail S, Padbury JF. Pathway-based genetic analysis of preterm birth. Genomics. 2013;101(3):163-70.
61. Verma SS, Lucas A, Zhang X, Veturi Y, Dudek S, Li B, et al. Collective feature selection to identify crucial epistatic variants. BioData Mining. 2018;11(1).
62. Wang H, Bennett DA, De Jager PL, Zhang QY, Zhang HY. Genome-wide epistasis analysis for Alzheimer's disease and implications for genetic risk prediction. Alzheimer's Research and Therapy. 2021;13(1).
63. Wang S, Jeong HH, Kim D, Wee K, Park HS, Kim SH, et al. Integrative information theoretic network analysis for genome-wide association study of aspirin exacerbated respiratory disease in Korean population. BMC Medical Genomics. 2017;10.
64. Wang X, Peng Q, Fan Y. Detecting Susceptibility to Breast Cancer with SNP-SNP Interaction Using BPSOHS and Emotional Neural Networks. BioMed research international. 2016;2016:5164347.
65. Weigelt B, Reis JS. Epistatic interactions and drug response. Journal of Pathology. 2014;232(2):255-63.
66. Wollstein A, Walsh S, Liu F, Chakravarthy U, Rahu M, Seland JH, et al. Novel quantitative pigmentation phenotyping enhances genetic association, epistasis, and prediction of human eye colour. Scientific reports. 2017;7:43359.
67. Wu C, Pan W. Integration of Enhancer-Promoter Interactions with GWAS Summary Results Identifies Novel Schizophrenia-Associated Genes and Pathways. Genetics. 2018;209(3):699-709.
68. Xu T, Monir MM, Lou XY, Xu H, Zhu J. Conditional and unconditional genome-wide association study reveal complicate genetic architecture of human body weight and impacts of smoking. Scientific reports. 2020;10(1):12136.
69. Zhao J, Cheng F, Jia P, Cox N, Denny JC, Zhao Z. An integrative functional genomics framework for effective identification of novel regulatory variants in genome-phenome studies. Genome Med. 2018;10(1):7.

70. Zhao X, Luan YZ, Zuo X, Chen YD, Qin J, Jin L, et al. Identification of Risk Pathways and Functional Modules for Coronary Artery Disease Based on Genome-wide SNP Data. Genomics, Proteomics and Bioinformatics. 2016;14(6):349-5671. Zhou J, Passero K, Palmiero NE, Müller-Myhsok B, Kleber ME, Maerz W, et al. Investigation of gene-gene interactions in cardiac traits and serum fatty acid levels in the LURIC Health Study. PloS one. 2020;15(9):e0238304.
72. Zhu H, Xia W, Mo XB, Lin X, Qiu YH, Yi NJ, et al. Gene-Based Genome-Wide Association Analysis in European and Asian Populations Identified Novel Genes for Rheumatoid Arthritis. PloS one. 2016;11(11):e0167212.

Supplement 2.

Originele Excel-bestand waarbij de artikels werden ingedeeld op basis van auteur, jaartal van publicatie, besproken ziekte, type abnormaliteit, studiedesign, type interacties, interactie als uitgangspunt, type predictie, methode voor predictie, besproken confounding en besproken population stratification. Onder andere gebruikt als data voor de opbouw van figuren.

https://www.dropbox.com/scl/fi/h5uvx2gnjj78budlxkcu6/Ruwe_data.xlsx?dl=0&rlkey=t9vc7cenezgj6ldyofepr8iyd

Supplement 3.

Code gemaakt in RStudio versie 4.1.2, in het bijzonder via de R-paketten RColorBrewer versie 1.1-3, ggplot versie 4.1.3, dplyr versie 1.0.10 en VennDiagram versie 1.7.3., om tot figuren 2-6 te komen.

https://www.dropbox.com/s/0t5yuxqfh4u2rlo/figuren_def.R?dl=0