# Nederlandse Samenvatting

Dit proefschrift behandelt nieuwe statistische methoden, die ontwikkeld zijn om de statistische power in genetische associatiestudies te verbeteren. De focus ligt op epidemiologische studies met een *response-selective sampling design*, zoals *case-control* studies met niet-verwante individuen en *case-control* studies met families. In deze samenvatting beschrijven we in detail nieuwe statistische methoden die (a) profiteren van de beschikbare informatie in de verdeling van de covariabelen in *case-control* studies door het *ascertainment* proces te modelleren; (b) informatie van familie-gebaseerde en *case-control* studies met niet-verwante individuen combineren; (c) gebruik maken van uitgebreidere modellen voor het beschrijven van de relatie tussen genetische varianten en fenotypen in standaard genetische associatiestudies; en (d) verschillende soorten data, zoals genomische, epigenomische, transcriptomische informatie integreren. Deze viertal punten kunnen samen de power verbeteren om de genetische basis van complexe menselijke eigenschappen te achterhalen.

Hoofdstuk 1 geeft een algemene inleiding op de bestaande methoden voor de statistische analyse van genetische associatiestudies met *response-selective sampling designs*. We introduceren de relevante terminologie en de belangrijkste concepten binnen genetische associatiestudies. Daarna vergelijken we de twee meest populaire *response-selective sampling designs* in genetische studies, namelijk *case-control* studies met niet-verwante individuen en die met families. De voordelen van methoden die rekening houden met *ascertainment* zijn de potentiële toename in statistische power voor het detecteren van associaties en het uitvoeren van een secundaire fenotype analyse.

De rest van de introductie is opgesplitst in twee delen. Het eerste deel laat drie verschillende likelihoods zien voor het modelleren van *ascertainment* in *case-control* studies met familiedata. De eerste is de *prospective likelihood*, waarin de verdeling van de uitkomst conditioneel op de covariabelen en de *ascertainment* wordt gemodelleerd. De tweede is de *ascertainment* gecorrigeerde *joint likelihood* die de gezamenlijke (joint) verdeling van de uitkomst en de covariabelen modelleert conditioneel op de *ascertainment*. De laatste is de *retrospective likelihood*. Deze modelleert de verdeling van de covariabelen gegeven de uitkomst. De *retrospective likelihood* is ook geschikt voor het analyseren van *case-control* data met niet-verwante individuen. We vergelijken de drie likelihoods met betrekking tot efficiëntie van de parameterschatters en de computationele kosten.

Het tweede gedeelte van de introductie beschrijft verschillende modellen voor de relatie tussen genetische varianten en de uitkomst. De huidige standaard voor genoombrede analyse is om voor elke *SNP* apart de relatie met de uitkomst te evalueren. Dit terwijl complexe ziekten meestal niet één enkele genetische oorzaak hebben, maar het gevolg zijn van een combinatie van meerdere genetische en omgevingsfactoren (bijv. polygenetisch). In dit gedeelte van de introductie presenteren we alternatieve methoden, die beter de onderliggende biologische mechanismen modelleren door het effect van meerdere genetische varianten of het effect van genetische en intermediare cellulaire fenotypen mee te nemen.

Hoofdstuk 2 beschrijft een nieuwe methode die de power van genetische associatie studies verbetert door data uit *multi-case* familie- en tweelingen studies met elkaar te combineren. Hierbij wordt ook het proces van *ascertainment* gemodelleerd. Om de efficiëntie van de parameterschatters te verhogen, gebruiken we de *ascertainment* gecorrigeerde *joint*

*likelihood*. Door gebruik te maken van een familie-specifiek random effect houden we rekening met de correlatiestructuur binnen families die veroorzaakt wordt door ongemeten genetische of omgevingsfactoren.

Met behulp van simulaties en echte data-analyse laten we zien dat belangrijke parameters efficiënter worden geschat door gebruik te maken van de *ascertainment* gecorrigeerde *joint likelihood*, waarin de familie en tweelingen data worden gecombineerd. Deze methode is efficienter dan de *ascertainment* gecorrigeerde *joint likelihood* met alleen familie data en de *prospective likelihood* waarin het *ascertainment* proces niet meegenomen wordt. De gecombineerde aanpak heeft niet alleen meer statistische power voor het vinden van effecten van individuele genotypen, maar ook het effect van het genotype van de moeder op de uitkomst, bijv. niet-overerfbare maternale antigen effecten. Deze verbetering in efficiëntie van de *joint likelihood* ten opzichte van de *prospective likelihood* is groter wanneer er minder informatie is. Bijvoorbeeld voor datasets met kleine families (3 kinderen per gezin) met tenminste twee aangedane kinderen. We zien vooral een verbetering van de efficiëntie bij een kleine steekproef van 100 of minder families voor de *joint likelihood* waarin families en tweelingen gecombineerd worden ten opzichte van de *joint likelihood* met alleen de familiedata.

Hoofdstuk 3 beschouwt een alternatieve strategie voor het testen van *haplotypes* in vergelijking met de huidige gouden standaard, namelijk marginale testen. Deze *single SNPs* testen zijn het afgelopen decennium het meest gebruikt. Alhoewel deze *single SNPs analyses* voor vele ziekten geleid hebben tot het identificeren van honderden geassocieerde genetische varianten, zou meer statistische power verkregen kunnen worden wanneer er gebruik gemaakt wordt van op haplotype gebaseerde statistische methoden. Deze methodoen analyseren namelijk meerdere genetische markers tegelijkertijd door gebruik te maken van *linkage disequilibrium* (LD) informatie. Hierdoor kan de power verbeteren voor het vinden van genetische varianten voor een bepaalde eigenschap (ziekte). Een nadeel van deze haplotype-gebaseerde statistische methoden is dat het aantal parameters exponentieel toeneemt met het aantal *SNPs*. Dit gaat samen met een overeenkomstige toename van het aantal vrijheidsgraden wat tot een afname in power om associaties te detecteren kan leiden.

Wij introduceren een hiërarchisch *linkage disequilibrium model* dat flexibele teststrategieën geeft voor het vinden van genetische varianten van eigenschappen over een serie van statistische hypothesen: van standaard *single SNP* analyses tot en met testen van associatie met volledige haplotypeverdelingen. Voor veel van deze hiërarchisch *linkage disequilibrium modellen* blijft het aantal vrijheidsgraden relatief laag, en daarmee is de power voor het detecteren van associaties dan ook beter. Het model is gebaseerd op een *reparametrisering* van de multinomiale haplotype verdeling waarin iedere parameter overeenkomt met een *joint cumulant* van elke mogelijke deelverzamelingen van *loci*. Een uitgebreide simulatiestudie en echte data-analyses laten zien dat testen binnen het hiërarchisch *linkage disequilibrium model* vaak een hogere statistische power hebben dan de *global haplotype test* en de *single SNP* associetietesten.

Genetische associatie studies hebben tot doel de associatie tussen genetische varianten en complexe genetische eigenschappen te detecteren. Voor de analyse van deze eigenschappen, kunnen twee verschillende methoden gebruikt worden: *linkage mapping* en *association mapping*. In hoofdstuk 4 bestuderen we de eigenschappen van deze twee methoden. *Linkage mapping* methoden zijn krachtiger voor het identificeren van zeldzame varianten die een effect hebben op vatbaarheid voor ziekte terwijl *association mapping* meer geschikt is voor het detecteren van algemeen voorkomende varianten met matige effect groottes. Echter, genetische varianten komen of vaak voor en hebben een klein effect op de uitkomts, of zijn te zeldzaam om hen effect op een betrouwbare wijze te schatten. Wanneer de effecten van de zeldzame variant groot waren geweest, en het fenotype niet heterogeen, dan kunnen deze varianten gedetecteerd worden met bijvoorbeeld *linkage* studies gebaseerd op familiedata.

Oftewel, er zou een methode moeten zijn waarbij meerdere zeldzame varianten met matige tot kleine effecten of de uitkomst moeten zijn. Een dergelijk uitgangspunt zou ideaal zijn voor het combineren van *linkage*- en *association mapping*.

We hebben een twee staps methode ontwikkeld om te onderzoeken of linkage gebaseerde methoden, zoals *identity by descent (IBD) mapping*, een bijdrage kunnen leveren aan het detecteren van zeldzame varianten via associatie in (*next-generation sequencing data*). In de eerste stap passen we IBD mapping toe om regio's te vinden waar *cases* meer segmenten IBD met elkaar delen dan *controls* rondom een vermeende causale variant. In de tweede stap doen we een associatie analyse met behulp van een *two stage mixed-effect model*. Met dit model kunnen we een overzicht creëren van de *SNP* data binnen de gevonden regio's en vervolgens nemen we deze *SNPs* mee als covariabelen in het model voor de uitkomst. Om de power te verbeteren, nemen we ook een variabele op die het aantal zeldzame varianten per regio telt. Met deze methode hebben we een significante associatie gevonden in de *next generation sequencing* longitudinale familiedata van de *Genetic Association Workshop 18*.

Hoofdstuk 5 bestudeert de analyse van de uitkomst variabele met meerdere soorten *omics* data, zoals genomische, epigenomische, en transcriptomische data (*integrative omics*). De *integrative omics* methode heeft zich ontpopt tot een krachtige en biologisch relevante richting van onderzoek voor associatiestudies. In *integrative omics* methoden maakt men vaak gebruik van het *case-control* design. Ook is het van belang om het effect van andere risicofactoren en covariabelen op de *omics* data te modeleren. Een open vraag is hoe je het beste meerdere *omics* datasets en deze risicofactoren en covariabelen het best kan integreren. Recente studies van *integrative omics* data maken gebruik van een *prospective model* waarin de *case-control* status conditioneel op de *omics* en de risciofactoren gemodelleerd wordt. In vergelijking met de univariate modellen heeft het analyseren van de meerdere risciofactoren in een prospectief model meer statistische power wanneer de individuen niet geselecteerd zijn. Echter, in *case-control studies* is dit niet het geval, daarom is de *power* vaak minder in vergelijking met de univariate aanpak.

Wij presenteren een nieuwe statistische methode voor *case-control* associatiestudies die het verlies in power kunnen opvangen en ook de meerdere *omics* en niet-*omics* factoren kunnen modelleren. Deze methode is gebaseerd op een *retrospective likelihood* functie waarin, conditioneel op de *case-control* status, de gezamenlijke verdeling van *omics* en risicofactoren gemodelleerd wordt. Om de verdeling van de risciofactoren efficiënt te kunnen modelleren, benutten we kennis over de correlatiestructuur tussen de risciofactoren in de populatie, en maken we gebruik van parametrische aannamen over de verdeling van de risciofactoren.

Uit simulatiestudies blijkt dat deze nieuwe statistische methode voldoet met betrekking tot de Type I fout en meer efficiëntie heeft, ten opzichte van de prospectieve benadering. De winst in efficiëntie hangt af van het aantal parameters in het model en de effectgrootten van de risciofactoren. Deze winst is groter wanneer de risciofactoren continu zijn en wanneer ze elk groot effect hebben.

Hoofdstuk 6 beschouwt het probleem van het correct beschrijven van een fenotype. Voor bepaalde genetische aandoeningen is het fenotype slecht gedefinieerd. Heterogeniteit in fenotypes leidt tot een lage power voor het detecteren van genetische associaties. Daarnaast zijn gevonden significante associaties vaak moeilijk te interpreteren. Het doel van de fenotypische classificatiemethode is het verfijnen van de classificatie van het fenotype met behulp van een, vaak hoogdimensionele, verzameling aan *features*. We bestuderen genetische syndromen als fenotypes en de pixels van tweedimensionale afbeeldingen van gezichten als *features*. We presenteren een methode voor geautomatiseerde classificatie en visualisatie van dit soort data.

Wanneer de data eerst getransformeerd wordt kan een betere classificatie verkregen worden. We onderzoeken het effect van verschillende transformaties op de nauwkeurigheid

van de classificatie in een hoog-dimensionele ruimte van *features*. Deze transformaties hebben betrekking tot een klein aantal *features*. Wanneer geregulariseerde regressietechnieken toegepast wordt op deze features, kan de classificatie naukeuriger zijn dan een principale componenten analyse.

Een tweede doel is het visualiseren van de classificatiefactoren die we gevonden hebben. We hebben *importance* plots ontwikkeld die de invloed van coördinaten in het originele tweedimensionale afbeelding weergeven. *Features* die gebruikt worden in de classificatie worden toegewezen aan coördinaten in het oorspronkelijke beeld en samengebracht tot een maat van *importance* voor elke pixel. Deze plots dienen als hulp bij het beoordelen van de plausibiliteit en de interpretatie van de classificaties en het bepalen van de relevante *features*.