# To fail or not to fail : clinical trials in depression

Sante, G.W.E.

Chapter

# 12

# Nederlandse samenvatting

# (Synopsis in Dutch)

# INLEIDING

## Achtergrond

Depressie is een ziekte die veel mensen direct (121 miljoen patiënten wereldwijd) of indirect treft. Op de lijst ziektes die de meest negatieve impact hebben op het aantal gezonde levensjaren zet de Wereld Gezondheids Organisatie (WHO) depressie op de $4^e$ plaats. Naar verwachting zal depressie in 2020 zelfs de $2^e$ plaats innemen.

De behandeling van depressie kan globaal worden onderverdeeld in psychotherapie en farmacotherapie. In dit proefschrift wordt alleen de farmacotherapeutische behandeling van depressie met antidepressiva nader onderzocht. Deze worden volgens hun werkingsmechanismen in therapeutische categorieën onderverdeeld. De twee belangrijkste typen antidepressiva zijn de zogenaamde tricyclische antidepressiva (TCA's) en de serotonine-specifieke heropname remmers (SSRI's). Vanwege hun verbeterde veiligheidsprofiel worden de SSRI's het meest voorgeschreven. Echter, ruim 30% van de patiënten heeft geen baat bij het gebruik van de geneesmiddelen die op dit moment op de markt zijn. Daarbij komt dat de antidepressiva allerlei ongewenste bijwerkingen hebben, zoals xerostomie (droge mond), obstipatie en seksuele stoornissen. Het behoeft dan ook geen nadere uitleg dat de ontwikkeling van nieuwe en betere geneesmiddelen voor de behandeling van depressie een noodzaak is.

## Ontwikkeling nieuwe antidepressiva

Bij het ontwikkelen van nieuwe geneesmiddelen dienen eerst de mogelijke onderliggende pathofysiologische mechanismen te worden opgehelderd. Vervolgens worden stoffen ontwikkeld met relevante aangrijpingspunten die deel uitmaken van het causale pad van de ziekte. Pas wanneer de veiligheid en tot op zekere hoogte de werkzaamheid ervan zijn aangetoond in diermodellen kunnen deze potentiële geneesmiddelen bij mensen worden getest. Om de veiligheid zeker te stellen wordt dit type onderzoek eerst uitgevoerd bij gezonde vrijwilligers (fase I onderzoek). Vervolgens zijn kleine groepen depressieve patiënten aan de beurt (fase II onderzoek), waarbij extra aandacht wordt besteed aan het bepalen van de juiste dosering. Tot slot worden zowel de werkzaamheid als de veiligheid bewezen in grote studies (fase III onderzoek). Dit proefschrift richt zich op de laatste twee fasen van het klinische onderzoek.

Ondanks de ontwikkeling en toelating van verschillende geneesmiddelen gedurende de laatste 20 jaar, heeft een meta-analyse van studies met geregistreerde antidepressiva in de database van de Amerikaanse registratie autoriteit (FDA) aangetoond dat bijna de helft van deze studies geen bewijs liet zien voor de werking van deze stoffen. Aangezien de meeste psychiaters overtuigd zijn van het nut en de werking van de geregistreerde antidepressiva, kan hieruit de conclusie getrokken worden dat zelfs wanneer een *effectief* antidepressivum wordt getest in een klinische studie, de kans van slagen van zo'n studie slechts 50% is. Dit is volstrekt onacceptabel, zowel vanuit ethisch oogpunt (patiënten hebben voor niets meegedaan aan klinisch onderzoek) als vanuit bedrijfskundig oogpunt

(de hoge kosten van een onderzoek dat niets oplevert). Deze situatie leidt ertoe dat het onderscheiden van onwerkzame van werkzame geneesmiddelen en dus de ontwikkeling van nieuwe antidepressiva onnodig vertraagd wordt.

De redenen voor het mislukken van klinische studies met antidepressiva zijn onder te verdelen in ziekte-, stof- en onderzoeksgerelateerde factoren. Onder de ziektegerelateerde factoren kan men de moeilijkheden bij het objectief meten van de ernst van depressie scharen, evenals de grote verschillen tussen patiënten met depressie en de verschillen tussen patiënten in hun respons op een behandeling. Stofgerelateerde factoren omvatten bijvoorbeeld de variabiliteit in de farmacokinetiek (en daardoor de blootstelling) van geneesmiddelen, de afwezigheid van een concentratie-effect relatie op basis van preklinisch onderzoek en gebrekkige therapietrouw ten gevolge van bijwerkingen. Voorbeelden van onderzoeksgerelateerde factoren zijn onvoldoende grote steekproeven, verkeerde inclusiecriteria, een te korte studieduur, de toepassing van niet-optimale statistische methoden en het gebruik van ongevoelige klinische eindpunten.

Op dit moment wordt de PKPD-benadering steeds vaker toegepast, welke de interactie tussen het biologische systeem,het geneesmiddel en toevalsfactoren beschrijft, die van belang is zowel bij het voorspellen van de farmacologische werking als bij het opzetten van een klinische studie. Met een farmacostatistische benadering worden in dit proefschrift onderzoeksgerelateerde factoren en andere methodologische aspecten onderzocht waardoor de slagingskans van een klinische studie met een effectief geneesmiddel vergroot kan worden. Hiervoor worden gegevens van reeds uitgevoerde klinische studies gebruikt.

## KLINISCHE EINDPUNTEN VOOR DEPRESSIE

In de hoofdstukken 3-5 wordt het onderzoek naar de klinische eindpunten voor depressie beschreven. **Hoofdstuk 3** laat zien dat wanneer gekeken wordt naar de 17 individuele items van de Hamilton schaal (HAMD) er een groot verschil is in de gevoeligheid voor het behandelingseffect. Deze items pogen alle facetten van de ziekte depressie te meten, zoals bijvoorbeeld stemmings- en angststoornissen, eetlust en slaapproblematiek. De gevoeligheid wordt onderzocht door met grafische en statistische methoden het verschil in tijdsverloop van de individuele items tussen responders (patiënten die een duidelijke verbetering vertonen tijdens de behandelingsperiode) en niet-responders (patiënten zonder verbetering) te bekijken. Een gevoelig item is een item dat een duidelijk verschillend tijdsverloop laat zien tussen responders en niet-responders. Aangezien er binnen de groepen responders/niet-responders geen verschillen werden gevonden in het tijdsverloop tussen placebo en actieve behandeling is hiervoor geen onderscheid gemaakt. Sommige items blijken, zoals verwacht, een zeer verschillend tijdsverloop te vertonen tussen responders en niet-responders. Andere items laten echter geen enkel verschil zien en missen gevoeligheid voor respons. Wanneer een subset van de HAMD gemaakt wordt met alleen de gevoelige items, blijkt dat een dergelijke subschaal vaak een groter verschil tussen een actieve behandeling en placebo vindt.

In **hoofdstuk 4** wordt een vergelijking gemaakt tussen de HAMD en de Montgomery-Asberg (MADRS) schaal, welke eveneens gebruikt wordt om de ernst van depressie te meten. Voor deze vergelijking zijn twee studies gevonden waarbij beide schalen toegepast zijn. Eerst zijn de 11 items van de MADRS vergeleken op hun gevoeligheid voor respons, op dezelfde manier als in hoofdstuk 3 voor de HAMD gedaan is. In tegenstelling tot de HAMD blijkt de MADRS slechts één ongevoelig item te bevatten, waardoor besloten is geen subschaal van dit eindpunt voor te stellen. Ook is het tijdsverloop van de items van de HAMD en de items van de MADRS die hetzelfde symptoomdomein beschrijven vergeleken. Hier werden geen verschillen aangetroffen, hoewel opviel dat alle subschalen van de HAMD het symptoomdomein bevatten dat 'algemene lichamelijke symptomen' beschrijft, terwijl een dergelijk item niet in de MADRS aanwezig is.

Vervolgens is gekeken of er een verschil bestaat in gevoeligheid tussen de HAMD, de MADRS en de subschalen van de HAMD, om het effect van een actieve behandeling van placebo te kunnen onderscheiden. Hieruit bleek dat de MADRS altijd gevoeliger was dan de HAMD, maar dat de subschalen van de HAMD op hun beurt gevoeliger waren dan de MADRS. Dit verschil zou verklaard kunnen worden door de afwezigheid van het symptoomdomein '*algehele lichamelijke symptomen*' in de MADRS.

Aangezien in de eerste fasen van de ontwikkeling van geneesmiddelen veel aandacht wordt besteed aan de selectiviteit van potentiële geneesmiddelen voor receptoren zou het interessant zijn om te onderzoeken of deze selectiviteit is terug te vinden in het klinische eindpunt. Dergelijke relaties tussen receptorsystemen en klinische eindpunten zijn bijvoorbeeld gevonden voor de GABA-receptor en EEG, en de dopamine D2-receptor en de ziekte van Parkinson. In **hoofdstuk 5** is onderzocht of voor de verschillende groepen antidepressiva (TCA's, SSRI's en het anti-epilepticum lamotrigine) specifieke responspatronen konden worden gevonden in de HAMD. Eerst is het tijdsverloop van de 17 items van de HAMD vergeleken voor de verschillende groepen antidepressiva, voor zowel responders als niet-responders. Vervolgens is gekeken naar de bijdrage van ieder individueel item van de HAMD aan de totale verandering in de HAMD van het begin van de studie tot de laatste meting bij een patiënt. Met beide methoden werd geen verschil gevonden tussen antidepressiva, hetgeen erop wijst dat de HAMD weliswaar geschikt is als maat voor de globale ernst van de ziekte, maar niet om de specifieke verschillen tussen antidepressiva te onderzoeken. Tevens betekent dit dat de receptorsystemen die door de onderzochte antidepressiva aangegrepen worden niet specifiek aan symptoomdomeinen van depressie gekoppeld kunnen worden. Wanneer een nieuwe stof wordt ontwikkeld die aangrijpt op één van deze systemen en bedoeld is om een symptoomdomein specifiek te verbeteren, is het dus onwaarschijnlijk dat een dergelijk effect onderscheiden kan worden door de HAMD. Hiervoor zijn andere schalen, zoals de zogenaamde multi-componentiële aanpak van Katz *et al.* beter geschikt. Bij het toepassen van deze schaal zijn immers wel verschillen gevonden tussen antidepressiva.

## STATISTISCHE METHODEN VOOR DE ANALYSE VAN KLINISCHE DATA

Voor de analyse van de klinische eindpunten die beschreven zijn in de voorgaande hoofdstukken zijn verschillende statistische methoden beschikbaar. Dit deel van het proefschrift behandelt een aantal van deze methoden.

**Hoofdstuk 6** laat de toepassing zien van een Bayesiaans parametrisch 'cure-rate' model (CRM), dat niet alleen de tijd tot respons, maar ook de fracties van genezen en ongenezen patiënten (niet-responders) kan bepalen. Dit model is een statistisch overlevingsmodel waarbij, in tegenstelling tot een standaard model, een deel van de patiënten niet 'dood gaat' of in dit geval 'respons vertoont'. Bovendien is het CRM toegepast in een Bayesiaans kader, waardoor de interpretatie van kansen en betrouwbaarheidsintervallen eenvoudiger is. Voorts is het mogelijk om reeds beschikbare informatie op een formele wijze in de statistische analyse mee te nemen. In dit hoofdstuk worden het gemiddelde en de standaarddeviatie van de verdeling van de responstijden in grote mate beïnvloed door historische data. In de ontwikkelingsfase van dit model is aangetoond dat het behandelingseffect het best beschreven kan worden door middel van een effect op het percentage niet-responders. Een analyse van twee studies met het CRM en een vergelijking met andere statistische modellen laat zien dat het CRM ongeveer even gevoelig is voor het vinden van een geneesmiddeleffect. Bovendien bleek dat het CRM in staat was een nauwkeurige voorspelling te doen van het geneesmiddeleffect nog voor het eind van de studie bereikt was. Helaas bleek het niet mogelijk om objectieve criteria vast te stellen waarmee bepaald kon worden *wanneer* tijdens een studie met het CRM een adequate uitspraak te doen is over het behandelingseffect. Bij het toepassen van een interim analyse staan dergelijke criteria centraal.

Variabiliteit bevat niet alleen ruis maar ook informatie. Bij de analyse van klinische data wordt vaak alleen gekeken naar het gemiddelde tijdsverloop per behandelingsgroep van het klinische eindpunt. Er kan echter ook veel informatie gehaald worden uit de verschillen tussen patiënten, waarmee vervolgens betere statistische modellen ontwikkeld kunnen worden. Deze zogenaamde 'functionele data analyse' wordt beschreven in **hoofdstuk 7**. Hierin wordt gekeken naar de vorm van de afwijking van patiënten van het gemiddelde. Deze afwijkingen zijn consequent terug te vinden bij verschillende klinische studies. De belangrijkste afwijking is een verticale afwijking, waarbij patiënten meer of minder depressief zijn dan de gemiddelde populatie tijdens het gehele verloop van een studie. De tweede afwijking is een hellingseffect, het gaat hierbij om patiënten die in het begin van een studie depressiever zijn dan de gemiddelde populatie, maar aan het eind juist minder depressief zijn (of andersom). Deze informatie zal in hoofdstuk 8 gebruikt worden bij de ontwikkeling van een nieuw statistisch model. Een andere bevindingen van de functionele data analyse is dat de verschillen tussen patiënten dezelfde vorm hebben bij responders enerzijds en niet-responders anderzijds. Dit is een aanwijzing dat responders en niet-responders wellicht geen volstrekt verschillende populaties zijn, maar eerder de uiteinden van een continu spectrum, waarbij de blootstelling aan een geneesmiddel

(en dus de farmacokinetische variabiliteit) een bepalende factor is. De functionele data analyse liet ook zien dat de vorm van de variabiliteit onafhankelijk is van het gebruikte klinische eindpunten, in dit geval de volledige HAMD en een subschaal van de HAMD. Hieruit kan afgeleid worden dat de belangrijke karakteristieken van patiënten behouden zijn in de subschaal van de HAMD, en dat deze schaal dus zonder problemen toegepast kan worden in de toekomstige analyse van klinische studies.

In **hoofdstuk 8** wordt de ontwikkeling van een duaal hiërarchisch random effecten model (DREM) beschreven. De keuze van parameters om de data te beschrijven (parameterisatie) van de toevalsfactoren (random effecten) is in hoge mate bepaald door de resultaten van hoofdstuk 7. Het eerste random effect is additioneel (verticale variatie) en het $2^e$ random effect wordt vermenigvuldigd met de tijd, waardoor het relatieve stijgings- en dalingsverloop beschreven wordt. Eerst wordt door middel van normalisatie van de discrepantie tussen gesimuleerde en geobserveerde data, oftewel de NPDE-methode, aangetoond dat data die gesimuleerd zijn met het DREM meer lijken op data van depressieve patiënten dan data die gesimuleerd zijn met de reeds beschikbare modellen. Omdat in veel studies ruim 30% van de patiënten de behandeling niet voltooit (drop-out), is vervolgens onderzocht in hoeverre de resultaten van een analyse met de DREM worden beïnvloed door het achterliggende mechanisme van dit verschijnsel. De invloed van verschillende soorten drop-out patronen op de schatting van het behandelingseffect wordt vergeleken voor (1) het DREM, (2) de zogenaamde 'last observation carried forward' methode (LOCF), waarbij op het moment dat een patiënt uit de studie verdwijnt of verwijderd wordt, de laatste waarneming wordt beschouwd als ware deze daadwerkelijk aan het eind van de studie gemeten, en (3) het 'mixed model for repeated measures' (MMRM), een ander veelgebruikt model. Deze analyse laat zien dat wanneer het drop-out percentage even hoog is voor de placebo als voor de actieve behandeling, resultaten van een studie weinig vertekend worden (bias), ongeacht het statistische model. Wanneer het drop-out percentage echter verschillend is, is de LOCF methode zelfs onder waarschijnlijke drop-out mechanismen niet in staat een goede schatting van het behandelingseffect te geven. De andere methoden ontlopen elkaar weinig en geven alleen verkeerde schattingen van het behandelingseffect wanneer sprake is van drop-out tengevolge van extreme, onwaarschijnlijke mechanismen.

## SIMULATIES VAN KLINISCHE STUDIES

In het vorige deel van het proefschrift is een nieuw model (DREM) ontwikkeld dat in het bijzonder geschikt is om data voor klinische studies te simuleren. In dit deel van het proefschrift wordt dit model gebruikt om te onderzoeken welke onderzoeksfactoren de nauwkeurigheid van de schatting van een behandelingseffect bepalen en hoe een klinische studie prospectief kan worden geoptimaliseerd teneinde fout-negatieve resultaten te verminderen.

In **hoofdstuk 9** wordt de invloed op het bepalen van een behandelingseffect van ver-

schillende onderzoeksfactoren onderzocht. Hierbij wordt gebruik gemaakt van simulatie scenario's waarbij één of meerdere studieprotocollen worden nagebootst (clinical trial simulations, CTS), een methode die de mogelijkheid biedt om de invloed van de opzet van klinische studies te bepalen *zonder* dat deze eerst uitgevoerd hoeven te worden, zoals bij meta-analyses het geval is. Hierbij is gekeken naar (a) de grootte van de steekproef (het aantal patiënten), (b) de invloed van ongelijke randomisatie over de verschillende behandelingen, (c) het aantal keren dat de ernst van depressie gemeten wordt per patiënt, (d) de invloed van verschillende realistische drop-out scenario's, (e) de gekozen klinische eindpunten, (f) statistische methoden voor het analyseren van de studies en (g) het invoeren van een interim analyse. De belangrijkste conclusies van dit onderzoek zijn de volgende:

1. Het is mogelijk het behandelingseffect nauwkeurig te schatten, zelfs wanneer de ernst van depressie slechts 2-3 keer gemeten wordt in plaats van de gebruikelijke 6-7 keer
2. Ongelijke verdeling van patiënten over de behandelingsgroepen (bijvoorbeeld minder patiënten in de placebogroep) kan het effect van de actieve behandeling aanzienlijk vertekenen. Hierdoor kan de kans groter worden dat men de onjuiste conclusie trekt uit een studie met een effectief antidepressivum. Wanneer ongelijke randomisatie wordt overwogen moet door middel van simulaties goed onderzocht worden wat de mogelijke gevolgen hiervan zijn.
3. Statistische analyse van afgeleiden van het klinische eindpunt, zoals het percentage responders of remitters (patiënten die als genezen beschouwd worden), leidt tot een veel lagere statistische kracht dan analyses van het klinische eindpunt zelf. Daarom zouden bij voorkeur de klinische schalen zelf geanalyseerd moeten worden.
4. Statistische analyse met de veelgebruikte LOCF methode leidt, ondanks dat dit op statistische gronden af te raden is, als realistische drop-out scenario's een rol spelen tot redelijke schattingen van het behandelingseffect.
5. Het is mogelijk een interim analyse op te nemen in de studieopzet die in staat is een ineffectieve behandeling vroegtijdig te stoppen zonder dat een effectieve behandeling een onacceptabel risico loopt vroegtijdig gestopt te worden.

**Hoofdstuk 10** gaat verder in op deze laatste conclusie. In dit hoofdstuk wordt nauwkeurig beschreven hoe een interim analyse geïmplementeerd kan worden die de mogelijkheid biedt een studie te stoppen wanneer er geen behandelingseffect is, of wanneer er voldoende bewijs is voor een behandelingseffect. Het is mogelijk zo'n interim analyse uit te voeren zonder dat er onacceptabel veel onjuiste beslissingen worden genomen. Dit wordt bereikt door gebruik te maken van gegevens over de inclusiesnelheid van patiënten van een lopende studie, alsmede van gesimuleerde data gebaseerd op parameters van reeds uitgevoerde studies. Door hypothetische interim analyses uit te voeren op deze gesimuleerde datasets met verschillende starttijden van de analyse en verschillende beslissingscriteria, kan ervoor gezorgd worden dat de uiteindelijke interim analyse de juiste statistische eigenschappen heeft.

De voorgestelde methoden zijn met de 're-enrolment test' getest op twee datasets van reeds uitgevoerde studies. Hierbij wordt gebruik gemaakt van het feit dat de volgorde van inclusie van patiënten willekeurig geacht wordt. Door deze volgorde te randomiseren ontstaan nieuwe interim-datasets, die geschikt zijn om de voorgestelde methodologie te testen. Hierdoor kan een gegeven methode rigoureuzer getest worden dan door deze alleen op de oorspronkelijke dataset toe te passen. Wij adviseren dan ook om in de toekomst bij de ontwikkeling van een nieuwe interim analyse deze te testen met de 're-enrolment test'. Uit de analyses blijkt dat de twee studies een grote kans maakten voortijdig gestopt te worden met de juiste redenen, waardoor tot 50% minder patiënten aan de studie deel hadden hoeven te nemen.

## SAMENVATTING EN CONCLUSIES

Depressie is een ziekte die veel mensen treft en die nog niet optimaal behandeld kan worden. De ontwikkeling van nieuwe geneesmiddelen wordt bemoeilijkt doordat bij de helft van de studies met werkende geneesmiddelen een onjuiste conclusie wordt getrokken over de effectiviteit van de behandeling. Enerzijds is in dit proefschrift onderzocht op welke manier de kans om een effectieve behandeling te vinden in een klinische studie vergroot kan worden. Anderzijds is de relevantie van een modelmatige analyse van historische data uit bestaande klinische studies aangetoond. Het gebruik van gevoeligere klinische eindpunten, zoals de MADRS of subschalen van de HAMD, is daarbij een eerste stap. Analyse van de klinische data met geschikte statistische modellen speelt ook een belangrijke rol. Ook laten wij zien hoe simulaties gebruikt kunnen worden om verschillende studieopzetten te vergelijken en te bepalen welke factoren van invloed zijn bij het schatten van de behandelingseffecten van nieuwe antidepressiva. Tot slot stellen we een adaptieve studieopzet voor waarbij een interim analyse gebruikt wordt die aangepast wordt aan de informatiedichtheid van de vergaarde data. Door deze interim analyse toe te passen kunnen klinische studies voortijdig afgebroken worden wanneer nieuwe informatie weinig toevoegt aan de kennis die reeds verzameld is.

In **hoofdstuk 11** worden de bevindingen uit dit proefschrift kritisch bediscussieerd en worden ook andere factoren die niet aan bod zijn gekomen besproken, evenals de toepasbaarheid van de gebruikte methoden voor andere ziektes. Lopende het onderzoek werden wij ons steeds meer bewust van de gebrekkige informatie die bestaat over de relatie tussen de blootstelling en de werking van antidepressiva. Naast de ontwikkeling van betere klinische eindpunten en geavanceerdere experimentele technieken, zou de toepassing van farmacokinetische-farmacodynamische (PKPD) modellen een belangrijke plaats kunnen innemen bij het ophelderen van deze relaties. De huidige onderzoekspraktijk, waarbij er stilzwijgend van wordt uitgegaan dat de blootstelling aan een antidepressivum niet medebepalend is voor de werkzaamheid hiervan, zou moeten worden verlaten voor een meer farmacologisch georiënteerde benadering, waarbij de concentratie-werkingsrelatie een centrale rol inneemt.