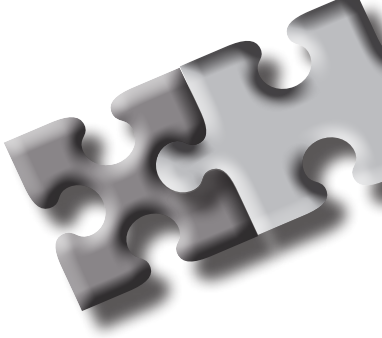

Samenvatting

Nederlands

Mis het niet!
Incomplete data kan waardevolle informatie bevatten

In epidemiologisch onderzoek wordt veel gebruik gemaakt van vragenlijsten om data te verzamelen. Deze vragenlijsten meten vaak een bepaald onderliggend construct door de scores op meerdere losse items (i.e., vragen) op te tellen tot een totaal score. Doordat een of meerdere vragen niet zijn ingevuld of doordat de gehele vragenlijst niet is ingevuld, kunnen de vragenlijst gegevens missende waarden bevatten. Missende scores op vragen (i.e., missende item scores) vereisen mogelijk andere statistische methoden dan missende totaal scores.

De onderliggende redenen van missende data kunnen onderverdeeld worden in verschillende mechanismes. De data kan *missing completely at random* (MCAR) zijn, wanneer het missende deel van de data een geheel random sub-sample van de data is. Een voorbeeld hiervan is dat een vragenlijst mist doordat deze in de post is kwijtgeraakt. Het is ook mogelijk dat de kans op missende data gerelateerd is aan andere variabelen die in de studie zijn gemeten. Dit mechanisme heet *missing at random* (MAR). Bijvoorbeeld wanneer fysieke activiteit scores vaker missen voor oudere mensen, dan is de missende data voor fysieke activiteit gerelateerd aan leeftijd. Missende data kan ook *missing not at random* (MNAR) zijn, wanneer de missende data gerelateerd is aan de missende score zelf. Bijvoorbeeld als de mensen met een lage score op fysieke activiteit hun fysieke activiteit score missen. De werking van methoden om met missende data om te gaan is afhankelijk van het onderliggende missing data mechanisme. Daarom is het belangrijk om een valide assumptie over het meest waarschijnlijke missing data mechanisme te maken. Dit kan gedaan worden door de data te onderzoeken en goed over de meest waarschijnlijke redenen voor de missende data na te denken.

Incomplete data in epidemiologische studies worden meestal simpelweg niet gebruikt in de analyse, oftewel een complete-case analyse. Daarnaast adviseren veel handleidingen van vragenlijsten om de missende waarden te vervangen voor een bepaalde waarde, bijvoorbeeld de gemiddelde score. Echter, deze methoden werken niet goed en veroorzaken bias in onderzoeksresultaten. Multipele imputatie en f*ull information maximum likelihood* schatting (FIML) zijn geavanceerde methoden om met missende data om te gaan. Beide methoden werken goed in MCAR en MAR data. In multipele imputatie worden de missende waarden vervangen door meerdere plausibele waarden, waardoor er meerdere kopieën van de dataset ontstaan met in iedere dataset andere geïmputeerde waarden. De plausibele waarden worden geschat met behulp van regressie technieken uit de geobserveerde data. De item scores in een vragenlijst worden vaak gemeten met een Likert schaal. Hierdoor zijn de item variabelen ordinaal en vaak niet normaal verdeeld. De regressie technieken om plausibele geïmputeerde waarden te schatten, zoals lineaire regressie, zijn dan niet altijd optimaal. Een procedure die robuust is tegen afwijkingen

van de normaal verdeling is *predictive mean matching.* Hierbij wordt er een random waarde getrokken uit de geobserveerde data waarden die het dichtst bij de voorspelde waarde uit de regressie schatting ligt. Deze methode gebruikt dus de geobserveerde data en imputeert daardoor meer realistische waarden. Multipele imputatie levert meerdere datasets op en deze worden ieder geanalyseerd met het analysemodel dat zou worden gebruikt als de data compleet was geweest. Vervolgens worden de resultaten van deze analyses gecombineerd voor het eindresultaat van de analyse. In FIML worden de populatie parameters geschat die meest waarschijnlijk het datasample zouden kunnen produceren. In deze methode worden geen waarden geïmputeerd of vervangen, maar alle geobserveerde data wordt gebruikt om de parameter schattingen te verkrijgen. Beide geavanceerde methoden, multipele imputatie en FIML, worden beschouwd als de state-of-the-art missing data methoden.

De beste missing data methode om met missende data om te gaan is afhankelijk van de analyse methode die wordt toegepast om de data te analyseren (i.e., longitudinaal of niet), het type variabele in de analyse dat missende waarden bevat (i.e., de predictor/ covariaat of de uitkomst), het missing data mechanisme (i.e., MCAR, MAR, MNAR), het percentage respondenten in de data met missende waarden en het niveau van de missende data in de vragenlijst (i.e., item scores of totaal score niveau).

Missende data in een vragenlijst moet worden behandeld op het item niveau van de vragenlijst. Als de uitkomst in een studie op één tijdspunt is gemeten, en er dus geen longitudinale analyse wordt uitgevoerd, moet multipele imputatie op de item scores worden toegepast. Dit houdt in dat de incomplete item variabelen worden geïmputeerd en dat na de imputatie de totaal scores van de vragenlijsten worden berekend en gebruikt voor analyse.

In studies met heel veel vragenlijsten of extreem lange vragenlijsten kan het aantal item variabelen te groot worden om betrouwbare imputaties te schatten. Een oplossing hiervoor is passieve imputatie. Passieve imputatie methoden combineren de variabelen in het imputatie model om het aantal variabelen in het model te reduceren. De item scores van een vragenlijst worden geïmputeerd, waarbij de totaal scores van de andere vragenlijsten worden gebruikt als predictor. Deze totaal scores kunnen ook missende waarden bevatten die veroorzaakt zijn door missende item scores, en deze worden dan ook op dezelfde manier geïmputeerd. De totaal scores worden tussen elke imputatie herhaling (i.e., iteratie) geüpdate door de geïmputeerde item scores.

Wanneer de uitkomst in een studie op meerdere tijdspunten wordt gemeten, moet er in de analyse methode rekening gehouden worden met de correlatie tussen de meerdere meetmomenten. Longitudinale analyses maken vaak gebruik van FIML procedures om parameter schattingen te verkrijgen en deze procedures behandelen de missende data in de analyse. Wanneer de uitkomst variabele wordt gemeten aan de hand van een vragenlijst, wordt over het algemeen alleen de totaal score van de vragenlijst in de longitudinale

analyse gebruikt. Desalniettemin moet de missing data op het item niveau aangepakt worden, wanneer de totaal scores incompleet zijn doordat de item scores missende waarden bevatten. De informatie uit de items kan in de analyse worden toegevoegd door de geobserveerde item scores te includeren als hulpvariabelen (i.e., *auxiliary variables*). Op die manier zijn de parameter schattingen meer precies en bevatten ze minder bias. De missende data in de predictor of covariaten in een longitudinale analyse moeten worden behandeld met multipele imputatie.

Het advies met betrekking tot vragenlijsten kan ook worden gebruikt voor andere situaties. Bijvoorbeeld bij kosten-data, waar de totale kosten worden gebruikt in een kosten-effectiviteitsanalyse. Deze totale kosten kunnen incompleet zijn door missende sub-kosten. Het is hier wederom het beste om op het sub-kosten niveau met de missende data om te gaan. Ook is de verdeling van kosten data bijna nooit normaal. Kosten zijn vrijwel altijd positief en vaak scheef naar rechts verdeeld met een overmaat aan nullen. De imputatie strategie kan worden aangepast door het gebruik van *predictive mean matching* op de log-getransformeerde data. Na de imputatie kan de data dan weer terug-getransformeerd worden voor de data analyse.

De belangrijkste conclusies van dit proefschrift zijn dat de methoden die in handleidingen voor vragenlijsten geadviseerd worden vaak niet optimaal zijn en moeten worden genegeerd. Missende waarden in vragenlijsten moeten worden behandeld op item niveau. De informatie uit de geobserveerde item scores verhoogt de accuraatheid en precisie in onderzoeksresultaten. Daarnaast vergroot de inclusie van geobserveerde item informatie als hulpvariabelen in een longitudinaal model om de totaal scores te analyseren de precisie en power van parameter schattingen. Passieve imputatie om missende item scores te imputeren in een dataset met een extreem groot aantal items is een valide methode om met missende item scores om te gaan.

Over het algemeen is het belangrijk om alle beschikbare informatie uit de geobserveerde item scores te betrekken bij het omgaan met missende item scores in vragenlijsten. Dit zorgt voor een optimaal niveau van accuraatheid en precisie in parameter schattingen.