SAMENVATTING

INLEIDING

Als je geïnteresseerd bent in de vraag welke van twee behandelingen of geneesmiddelen het beste werkt, zijn er grofweg twee manieren om dat te onderzoeken: *experimenteel* en *observationeel*. Bij een experiment heeft de onderzoeker controle over wie welke behandeling krijgt. Vaak zal de onderzoeker dat *door het toeval* laten bepalen (randomiseren), zodat er gemiddeld genomen geen andere factoren van invloed zijn op de relatie tussen behandeling en uitkomst. Dit houdt in dat de behandelingsgroepen gemiddeld genomen 'gelijk' zijn en dus een directe relatie kan worden gelegd tussen behandeling en uitkomst. Deze manier van onderzoek is over het algemeen geaccepteerd als de beste methode om het effect van een behandeling of geneesmiddel vast te stellen.

Bij een observationele studie heeft de onderzoeker *geen invloed* op het toewijzen van behandelingen aan de personen in het onderzoek. Het directe gevolg hiervan is dat de behandelingsgroepen niet alleen 'behandeling' als verschil hebben, maar ook op *andere kenmerken* zullen verschillen. Hierdoor is het veel lastiger om het directe effect van de behandeling te bepalen, met name als deze kenmerken ook invloed hebben op de uitkomst (*prognostische factoren*). Bijvoorbeeld: behandeling A wordt vaker aan jongeren gegeven en behandeling B wat vaker aan ouderen. Als blijkt dat behandeling A over het geheel beter werkt, hoeft dat niet alleen aan de behandeling te liggen, maar kan dat ook komen doordat jongeren over het algemeen gezonder zijn dan ouderen. In zo'n geval is het effect van de behandeling op de uitkomst *verstrengeld* of *verward* met het effect van leeftijd op de uitkomst. De Engelse term hiervoor is *confounding*.

Ondanks de *verstrengeling* of *verwarring* met mogelijke andere factoren, kunnen observationele studies van grote waarde zijn om het effect van een behandeling te schatten, zeker als experimentele studies niet mogelijk zijn. Dit is bijvoorbeeld bij de lange-termijneffecten van medicijnen of bij patiënten die bij de experimentele studies zijn uitgesloten van deelname. Als observationele studies inderdaad gebruikt worden om het effect van een behandeling te schatten, zal de belangrijkste aandacht moeten uitgaan naar het *ontwarren* van behandelingseffect en effecten van verstorende factoren. Hiervoor zijn verschillende *correctie-* of *ontwarringsmethoden* beschikbaar die in Hoofdstuk 2 zijn besproken. Twee daarvan zijn in de andere hoofdstukken nader aan de orde geweest.

EEN OVERZICHT VAN CORRECTIEMETHODEN

Deze correctiemethoden kunnen in twee groepen worden verdeeld: methoden waarbij *voorafgaand aan de studie* wordt gecorrigeerd en methoden die corrigeren *nadat de data is verzameld*. *Restrictie* en specifieke onderzoeksdesigns zoals *case-crossover* en *case-time-control* designs zijn voorbeelden van correctiemethoden voorafgaand aan de studie. Tot de traditionele (analytische) methoden die achteraf corrigeren behoren naast standaardmethoden als *stratificatie*

en *matching*, ook de multivariabele statistische methoden *logistische regressie* en *Cox proportional hazards regressie*. Bij deze methoden worden geobserveerde variabelen toegevoegd aan een model met alleen behandeling als verklarende factor, met als doel het geschatte behandelingseffect te corrigeren voor de verstorende invloed van de toegevoegde variabelen (covariaten). Dit betekent dat hierbij geen speciale aandacht wordt besteed aan de ongelijke verdelingen van covariaten tussen behandelingsgroepen. Een methode die zich wel richt op de balans in covariaten tussen behandelingsgroepen is de methode van *propensity scores*. Eerst wordt een propensity score model gemaakt om de conditionele kans te schatten dat een individu tot de behandelde groep behoort (t.o.v. onbehandelde groep), gegeven een set covariaten. Vervolgens wordt dan een gecorrigeerd behandelingseffect geschat waarbij de propensity score wordt gebruikt als matchingvariabele, stratificatievariabele, covariaat of als gewicht (inverse probability weight).

Bij al deze methoden is een belangrijke beperking dat correctie alleen plaatsvindt voor variabelen die ook daadwerkelijk zijn gemeten en dat voor andere belangrijke ongemeten covariaten niet kan worden gecorrigeerd. Een methode die deze beperking niet kent, is de methode van *instrumentele variabelen*. Bij deze methode ligt het accent op het vinden van een variabele (het instrument) die samenhangt met de variabele 'behandeling' en die alleen aan de uitkomstvariabele gerelateerd is via de samenhang met 'behandeling'. Deze methode kan hetzelfde effect bereiken als het randomiseren in een experiment, maar de sterke aannames beperken het gebruik in de praktijk. Gerelateerde methoden met een soortgelijke beperking zijn *two-stage least squares* and *grouped-treatment approach*.

In de rest van de hoofdstukken hebben we ons beperkt tot de methoden van propensity scores en instrumentele variabelen. Hiervan hebben we de sterke en zwakke punten en speciale toepassingen besproken en hebben een balansmaat voorgesteld bij gebruik van propensity score methoden.

STERKE EN ZWAKKE PUNTEN VAN CORRECTIEMETHODEN

BELANGRIJK VOORDEEL VAN PROPENSITY SCORES

In Paragraaf 3.1 en Paragraaf 3.2 van Hoofdstuk 3 is het gebruik van logistische regressieanalyse vergeleken met propensity score methoden. In medisch onderzoek worden beide methoden toegepast om een gecorrigeerd behandelingseffect te schatten in observationele studies. Hoewel de effectschattingen van beide methoden in de literatuur als 'vergelijkbaar' en 'niet echt verschillend' worden bestempeld, is in deze paragrafen aangetoond dat de verschillen *systematisch* zijn en *wel substantieel verschillend* kunnen zijn. Met betrekking tot het doel om voor ongelijkheid in de verdelingen van covariaten te corrigeren, is laten zien dat het geschatte behandelingseffect bij propensity score methoden over het algemeen *dichter bij het werkelijke te schatten effect* ligt dan bij logistische regressie-analyse. Het verschil kan groot zijn, vooral

als het aantal prognostische factoren groter is dan 5, het behandelingseffect groter is dan een odds ratio van 1.25 (of kleiner dan 0.8) of als de incidentie groter is dan 5%. Dit betekent dat er een voordeel is van propensity score methoden vergeleken met logistische regressiemodellen, hetgeen door onderzoekers en in de literatuur vaak over het hoofd wordt gezien.

Deze overschatting van het behandelingseffect in logistische regressie-analyse komt door het gebruik van de *odds ratio* als effectmaat. Hetzelfde geldt voor de *hazard ratio* zoals deze bijvoorbeeld gebruikt wordt in Cox proportional hazards regresie; het geldt niet voor minder vaak gebruikte effectmaten zoals het risicoverschil. Het voordeel van propensity score methoden is dus dat de odds ratio kan worden gebruikt zonder dat daarbij het werkelijke gemiddelde behandelingseffect ernstig wordt overschat, zoals bij logistische regressie het geval kan zijn. Deze bevinding is onafhankelijk van de manier waarop de propensity score wordt gebruikt (voor stratificatie, voor matching of als covariaat). Dit voordeel verdwijnt echter als naast correctie via de propensity score ook nog extra gecorrigeerd wordt door covariaten apart in het uitkomstmodel op te nemen.

We concluderen dat de methode van propensity scores over het algemeen *beter geschikt is* om een gemiddeld behandelingseffect te schatten dan logistische regressie-analyse en Cox proportional hazards regressie.

BEPERKING VAN DE METHODE VAN INSTRUMENTELE VARIABELEN

In Paragraaf 3.3 is gekeken naar de mogelijkheden om via de methode van instrumentele variabelen te corrigeren voor vertekening in niet-gerandomiseerde studies. Het sterke punt van deze methode is de potentie om voor *alle verstorende invloeden* te corrigeren, zowel geobserveerde als niet-geobserveerde. Het zwakke punt van deze methode ligt in de *belangrijkste voorwaarde* die er op neerkomt dat de instrumentele variabele geen directe invloed mag hebben op de uitkomstvariabele, en ook geen indirecte invloed via de relatie met andere variabelen behalve behandeling. Of aan deze voorwaarde is voldaan kan in de praktijk alleen theoretisch worden beoordeeld en kan niet empirisch getest worden.

Verder is ingegaan op de *correlatie* tussen de instrumentele variabele en behandeling (of blootstelling). Als deze correlatie erg klein is, zal de methode niet alleen tot een grote standaardfout van de effectschatting leiden, maar ook tot een aanzienlijke *onzuiverheid* van de schatter bij kleine steekproeven; ook bij grote steekproeven zal een dergelijke onzuiverheid optreden als niet volledig aan de belangrijkste voorwaarde is voldaan.

Daarnaast is een *bovengrens* aangetoond van de correlatie tussen de instrumentele variabele en de behandeling (of blootstelling). Deze bovengrens is echter geen praktische belemmering als er geringe of matige vertekening bestaat, omdat deze bovengrens behoorlijk hoog ligt. Als de mate van vertekening wel groot is, bijvoorbeeld in geval van *vertekening door indicatie*, dan zal de maximale correlatie tussen de potentiële instrumentele variabele en de behandeling erg laag zijn, hetgeen resulteert in een erg zwakke instrumentele variabele. Omdat er een wisselwerking is tussen het voldoen aan de voorwaarde en de sterkte van de instrumentele variabele,

zal het bestaan van sterke vertekening naast een sterk instrument zeer waarschijnlijk leiden tot schending van deze veronderstelling en een onzuivere schatting van het behandelingseffect geven.

We concluderen dat de methode van instrumentele variabelen bruikbaar kan zijn *bij matige vertekening*, maar dat deze methode minder bruikbaar is bij sterke vertekening (bijv. door indicatie) omdat sterke instrumenten niet gevonden kunnen worden en voorwaarden gemakkelijk worden geschonden.

TOEPASSING VAN CORRECTIEMETHODEN

Propensity score methoden en in mindere mate instrumentele variabele methoden, worden in toenemende mate toegepast in de medische literatuur. In Hoofdstuk 4 zijn beide methoden toegepast op *overlevingsdata*.

PROPENSITY SCORES EN OVERLEVINGSDATA

In Paragraaf 4.1 is de methode van propensity scores toegepast op een studie naar het effect van het behandelen van hypertensie voor het voorkomen van een hersenberoerte. Wat betreft de mogelijkheid om voor vertekening te corrigeren, zijn drie propensity score methoden met elkaar vergeleken en vergeleken met de traditionele Cox proportional hazards regressie. Het aantal personen met een hersenberoerte in onze dataset was gering hetgeen een voordeel is voor propensity score methoden omdat dan meer covariaten kunnen worden opgenomen in vergelijking met standaard regressiemethoden.

Uit de literatuur is bekend dat bij de meeste propensity score toepassingen de propensity score als covariaat wordt gebruikt (is niet aanbevolen), dat de selectie van het propensity score model *routinematig* wordt gedaan door alle covariaten in het model te stoppen en dat aandacht voor en het *rapporteren van de bereikte balans* over het algemeen onvoldoende is. We hebben enkele methoden laten zien om inzicht te krijgen in de balans en om deze te rapporteren. Het blijkt dat er geen uniforme manier is om dat te doen in de praktijk. Tot slot hebben we een variabele toegevoegd die niet met de behandeling samenhangt waaruit geconcludeerd kan worden dat het geschatte behandelingseffect kleiner is bij stratificatie op de propensity score dan bij Cox proportional hazards regressie. Dit bevestigt de resultaten in Paragraaf 3.2.

We kunnen concluderen dat propensity score methoden bruikbaar zijn in combinatie met overlevingsdata, maar dat er meer aandacht moet zijn voor het maken van het propensity score model en voor het checken en rapporteren van de balans.

INSTRUMENTELE VARIABELEN EN OVERLEVINGSDATA

Hoewel toepassingen van instrumentele variabelen langzaamaan wat vaker in de medische literatuur verschijnen als methoden om voor vertekening te corrigeren, wordt deze methode nauwelijks toegepast bij gecensureerde overlevingsdata. In Paragraaf 4.2 is deze methode

toegepast met als behandelingseffect het verschil in overlevingskansen. Hiervoor hebben we gegevens gebruikt van type-1 diabetespatiënten die zijn verzameld om het verschil in overleving te bepalen tussen patiënten die alleen een niertransplantatie hebben ondergaan en patiënten bij wie daarnaast ook een alvleesklier is getransplanteerd. Als deze twee groepen patiënten direct met elkaar zouden worden vergeleken (*as-treated* of *zoals-behandeld*), zal het werkelijke verschil tussen beide behandelingen verkeerd worden geschat vanwege vertekening, ofwel omdat bij het bepalen van het type transplantatie selectieprocessen een rol spelen. Als daarentegen de oorspronkelijk gerapporteerde methode gebruikt zou worden om het behandelingseffect te schatten (*intention-to-treat* of *zoals-bedoeld*), dan worden in feite niet de behandelingen zelf vergeleken maar het beleid van ziekenhuizen om in principe de ene dan wel de andere methode te kiezen. Er is dan geen vertekening, maar het werkelijke verschil tussen de typen operaties wordt onderschat. Een derde mogelijkheid is de methode van instrumentele variabelen. Hiermee wordt het werkelijke effect van het tegelijkertijd transplanteren van de alvleesklier geschat, waarbij gecorrigeerd wordt voor vertekening. Het blijkt dan dat het effect veel groter is dan bij de methode van intention-to-treat, zij het dat de onbetrouwbaarheid van de schatting ook groter is.

Een nadeel van het gebruik van deze methode is de gevoeligheid van de effectschatting voor de onderliggende veronderstellingen. Een ander nadeel is dat bij weinig waarnemingen of bij een gering aantal gebeurtenissen de overlevingscurve die via de methode van instrumentele variabelen is geschat *onbetrouwbaarder* is dan bij andere schattingsmethoden. Dit geldt dan voornamelijk aan het einde van de curve, als er nog maar weinig personen tot de risicoset behoren.

We concluderen dat de methode van instrumentele variabelen ook gebruikt kan worden in geval van overlevingsdata. Aandacht voor de voorwaarden die ten grondslag liggen aan deze methode, is van groot belang. Standaardfouten van de schattingen moeten worden vermeld en kunnen erg groot zijn aan het einde van de overlevingscurve, zeker als het aantal gebeurtenissen gering is.

VERBETERING VAN PROPENSITY SCORE METHODEN

Zoals uit de literatuur en uit de toepassing in Paragraaf 4.1 is gebleken, is er behoefte aan informatie over hoe de methode van propensity scores precies moet worden toegepast, met name als het gaat om het maken van het propensity score model en het checken van de balans. In Hoofdstuk 5 zijn we nader ingegaan op balansmaten die kunnen helpen om uniformiteit te creëren bij het rapporteren van de balans die is bereikt en die kunnen helpen bij het selecteren van de variabelen voor het propensity score model om zo groot mogelijke gelijkheid op covariaten tussen behandelingsgroepen te verkrijgen.

DE OVERLAPPINGSCOËFFICIËNT

In Paragraaf 5.1 is voorgesteld om de *overlappingscoëfficiënt* te gebruiken bij propensity score toepassingen. Dit is een bestaande maat om de overlap tussen twee verdelingen te schatten. Als de propensity score in strata is verdeeld, kan met deze maat aangegeven worden hoe binnen deze strata de overlap is op de covariaten tussen beide behandelingsgroepen. Verder kan deze maat vergeleken worden met een waarde die men zou kunnen *verwachten* in geval van perfecte overlap in de populatie om zo de kwaliteit van de correctie voor vertekening te beoordelen. Deze informatie is van cruciaal belang om het gerapporteerde behandelingseffect te kunnen interpreteren. Vaak genoeg wordt alleen vermeld *dat er gecorrigeerd is* voor een set covariaten zonder daarbij te vermelden in welke mate dat ook daardwerkelijk *gelukt* is. Deze overlappingscoëfficiënt kan ook gebruikt worden om uit de vele mogelijke modellen een model te selecteren met een acceptabele balans; met een simulatiestudie hebben we namelijk aangetoond dat er een relatie is tussen deze maat en de zuiverheid van de effectschatter. De gewogen gemiddelde overlappingscoëfficiënt die berekend is op een set van beschikbare covariaten heeft de sterkste samenhang met de zuiverheid van de schatter in grotere datasets. Voor kleinere datasets hebben andere maten zoals de p-waarden uit significantietesten of de c-statistic een betere voorspellende kracht dan de overlappingscoëfficiënt.

We concluderen dat het gebruik van de overlappingscoëfficiënt nuttig kan zijn bij het toepassen van propensity score methoden. Bij kleinere steekproeven lijkt de methode minder goed te werken omdat dan het schatten van de overlap moeilijker is.

MEER BALANSMATEN

Een prettige eigenschap van de overlappingscoëfficiënt is dat het een directe maat is voor de overlap tussen twee verdelingen en daarmeee de balans in een propensity score toepassing. In Paragraaf 5.2 hebben we ook naar andere maten gekeken om deze balans vast te stellen. In een simulatiestudie (die anders is qua opzet dan in Paragraaf 5.1) hebben we de overlappingscoëfficiënt vergeleken met de *Kolmogorov-Smirnov afstand* en de *Lévy metriek*. Bij alle drie was een relatie aantoonbaar met de zuiverheid van de schatter; deze was het sterkste voor de gewogen overlappingscoëfficiënt en was sterker voor grotere steekproeven. Voor de overlappingscoëfficiënt was de samenhang $R = -0.06$ bij 400 waarnemingen en $R = -0.63$ bij 2000 waarnemingen. Een propensity score model selecteren met behulp van de overlappingscoëfficiënt lijkt het meest effectief te zijn, omdat de *gemiddelde gekwadrateerde fout* voor deze methode over het algemeen het kleinst was (variërend van 0.031 tot 0.197). De verschillen met de twee andere balansmaten waren echter niet groot. Een standaard propensity score model dat alle covariaten bevat, had een veel grotere gemiddelde gekwadrateerde fout. Ook als het propensity score model gekozen wordt met alle werkelijke (en in de praktijk altijd onbekende) verstorende factoren, dan was de gemiddelde gekwadrateerde fout groter.

We concluderen dat het gebruik van een van de gepresenteerde maten nuttig is bij het toepassen van propensity score methoden. Voorkeur gaat uit naar de overlappingscoëfficiënt

vanwege de directe interpretatie en de geringere gemiddelde gekwadrateerde fout bij het schatten van het behandelingseffect. Ondanks dat het schatten lastiger is bij kleinere steekproeven, levert de modelkeuze via de overlappingscoëfficiënt betere resultaten op dan de in de praktijk vaak gebruikte keuze voor het model met alle covariaten.

TOT SLOT

Propensity score methoden hebben verschillende voordelen ten opzichte van traditionele regressiemethoden als het gaat om het corrigeren voor vertekening (*confounding*). Een belangrijk voordeel is dat propensity score methoden behandelingseffecten geven die over het algemeen dichter bij de werkelijke gemiddelde behandelingseffecten liggen dan logistische of Cox proportional hazards regressie. Daarom zou deze methode vaker moeten worden toegepast als het gaat om correctie voor vertekening. Daarbij komt wel dat er meer aandacht voor het maken van het propensity score model moet zijn en voor het meten en rapporteren van de bereikte balans op covariaten, bijvoorbeeld door het gebruik van de overlappingscoëfficiënt.

De methode van instrumentele variabelen wordt veel minder toegepast in epidemiologisch onderzoek. Het zou wel meer aandacht verdienen omdat het een goede mogelijkheid is om volledig voor vertekening te corrigeren. Tegelijkertijd zullen de beperkingen van deze methode zeer algemeen gebruik ervan in de weg staan. Goede voorbeelden van toepassingen in de medische literatuur zullen ertoe bijdragen dat de methode op juiste waarde wordt geschat.