



Universiteit
Leiden

The Netherlands

Handling missing data, selection bias, and measurement error in observational studies

Choi, J.

Citation

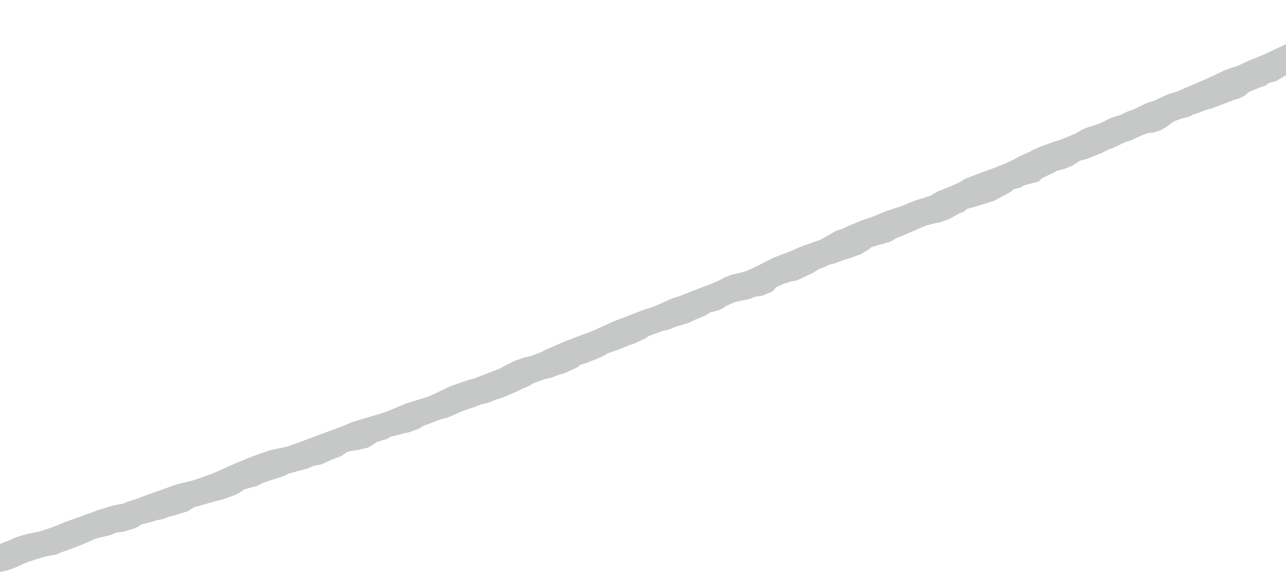
Choi, J. (2023, June 22). *Handling missing data, selection bias, and measurement error in observational studies*. Retrieved from <https://hdl.handle.net/1887/3626684>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/3626684>

Note: To cite this publication please use the final published version (if applicable).



Appendix

Dutch Summary

In dit proefschrift onderzoeken wij de potentiële uitdagingen die de validiteit van observationeel epidemiologisch onderzoek kunnen aantasten. Deze potentiële bronnen van vertekening, zoals confounding, ontbrekende gegevens, selectiebias en meetfouten, vormen aanzienlijke obstakels voor een accurate interpretatie van onderzoeksresultaten. Hoewel talrijke methoden zijn ontwikkeld om het effect van vertekeningen te verminderen, blijft het bepalen van de meest geschikte benadering voor specifieke empirische contexten een complexe taak. Bovendien krijgen problemen die worden besproken in de methodologische literatuur, in klinisch onderzoek vaak beperkte aandacht. Daarom analyseren we de problemen omtrent ontbrekende gegevens, selectiebias en meetfouten die optreden in diverse specifieke observationele situaties, en bespreken we de meest geschikte methoden om deze kwesties te behandelen.

In **Hoofdstuk 2** hebben wij systematisch onderzocht hoe covariabelen met ontbrekende gegevens op de meest optimale manier in een propensity score-analyse kunnen worden behandeld door verschillende simulatiescenario's te genereren. Hierbij hebben we rekening gehouden met diverse mechanismen voor ontbrekende gegevens en de aanwezigheid van modificatie van het behandelingeffect. Onze bevindingen tonen aan dat er geen universeel optimale aanpak is, aangezien de geschikte methode afhangt van de datastructuur, zoals het ontbrekende mechanisme, effectheterogeniteit en ongemeten confounding. Opmerkelijk is dat een complete case-analyse of het toevoegen van ontbrekende indicatoren aan een model, die doorgaans als 'naïef' en ongeschikt voor het omgaan met ontbrekende gegevens beschouwd worden, beter presteerden dan meervoudige imputatie bij niet-random ontbrekende gegevens. Meervoudige imputatie was het meest effectief wanneer gegevens willekeurig ontbraken, maar alleen wanneer het imputatiemodel correct gespecificeerd was. Dit impliceert dat de uitkomstvariabele in het imputatiemodel moet worden opgenomen en, indien van toepassing, een interactieterm om rekening te houden met heterogeniteit in het behandelings-effect.

In **Hoofdstuk 3** hebben we methoden onderzocht voor het opsporen van meetfouten die mogelijk te wijten zijn aan het verdunnen van monsters in tijdreeksen hormoongegevens. We vergeleken vier methoden: i) beoordeling door fysiologische experts, dat als gouden standaard gezien kan worden, ii) de stapsgewijze aanpak, die fysiologische kennis integreert met op standaarddeviatie gebaseerde detectie, iii) de methode van Tukey's fences, die fouten identificeert op basis van interkwartielafstanden, en iv) het verwachting-maximalisatie (EM) algoritme, dat de verdelingen van hormoonspiegels met en zonder fout wiskundig onderscheidt. Op basis van de prestaties in de praktijk en gesimuleerde gegevens, concludeerden we dat de stapsgewijze methode, die gebruikmaakt van fysiologische kennis, beter presteerde dan volledig geautomatiseerde datagedreven methoden zoals Tukey's fences en het EM-algoritme.

In **Hoofdstuk 4** zijn strategieën voor het omgaan met variabelen die worden beïnvloed door medicijngebruik onderzocht, voor de situatie waarbij interesse is effecten zonder behandeling. Een voorbeeld is het onderzoeken van de invloed van een genetische factor op de bloeddruk op een bepaalde leeftijd of de invloed van bloeddruk op het risico op hart- en vaatziekten zonder gebruik van antihypertensiva bij personen met hypertensie. Uit simulaties bleek dat de keuze van de methode afhangt van of de beïnvloede variabele de blootstelling, de uitkomst of een confounder is. Als de blootstelling wordt beïnvloed, kan het reduceren van de onderzoekspopulatie tot onbehandelde individuen tot een valide resultaat leiden. Echter, als er effectheterogeniteit aanwezig is, kunnen de resultaten mogelijk niet naar de algemene populatie worden geëxtrapoleerd. Als externe kennis over het gemiddelde en de standaardafwijking van het medicijneffect bekend is, kan regressiecalibratie met het toevoegen van het gemiddelde medicijneffect aan de behandelde waarden worden gebruikt. Als de uitkomst wordt beïnvloed, leiden simpele gangbare methoden tot vertekening. In plaats daarvan is het optellen van het gemiddelde medicijneffect of het gebruik van gecensureerde normale regressie geschikt. Op basis van de resultaten adviseren we onderzoekers om kritisch na te denken over het proces van medicijnvoorschrijving, de aanwezigheid van effectheterogeniteit en welke informatie over medicijneffecten beschikbaar is.

Verschillende methoden die in **Hoofdstuk 4** worden besproken, vereisen externe kennis over het geschatte effect van medicatie en in sommige gevallen de standaardafwijking van het medicijneffect. Deze gegevens kunnen uit gerandomiseerde gecontroleerde onderzoeken (RCT's) verkregen worden. Populaties in dergelijke onderzoeken zijn vaak niet representatief voor de populaties in observationeel onderzoek. Door verschillen tussen klinische settings in RCT's en praktijksettings, kan het toepassen van het medicijneffect verkregen in RCT's in observationele settings vertekening geven.

In **Hoofdstuk 5** hebben we veranderingen in glucose- en HbA1c-niveaus na het gebruik van glucoseverlagende medicatie beschreven met behulp van routinematig verzamelde gegevens van deelnemers aan de Nederlandse Epidemiologie van Obesitas (NEO) studie. Elektronische Patiëntendossiers werden gebruikt om nieuwe gevallen van diabetes en herhaalde metingen van glucose en HbA1c te identificeren. Lineaire gemengde modellen werden toegepast, waarbij tijd als een categorische variabele werd gebruikt voor zowel vaste als willekeurige effecten. Om regressie naar het gemiddelde te beperken, werd het referentiepunt ingesteld op 6 tot 12 maanden vóór het voorschrijven van medicatie. De resultaten toonden aan dat de meest aanzienlijke medicatie-effecten optraden 6 tot 12 maanden na de start van de behandeling, en dat de medicatie ook na twee jaar effectief bleef. Deze effecten waren echter kleiner dan die waargenomen in RCT's. We zagen verhoogde glucose- en HbA1c-waarden kort voor het gebruik van medicatie, wat mogelijk wijst op willekeurig hoge metingen die leiden tot behandelingsbeslissingen en daaropvolgende regressie naar het gemiddelde.

In **Hoofdstuk 6** is systematisch in de wetenschappelijke literatuur onderzocht hoe variabelen die door medicatie worden beïnvloed, worden behandeld in klinisch onderzoek. Hierbij kwamen we ambiguïteit in onderzoeksdoelen en het frequente gebruik van ongeldige methoden tegen. Geavanceerde methodologieën werden zelden toegepast, wat wijst op een behoefte aan meer kennis onder klinische onderzoekers en de noodzaak van een duidelijke onderzoeksvraag met betrekking tot medicijngebruik.

In **Hoofdstuk 7** zijn verschillende manieren onderzocht om medicijngebruik op te nemen in onderzoeksvragen wanneer blootstelling- of uitkomstvariabelen worden beïnvloed door medicatie bij sommige individuen. We bespreken aannames over de relaties tussen variabelen van belang en de mogelijke klinische relevantie daarvan. Sommige vragen passen in een causaal kader, waar het nabootsen van gerandomiseerd onderzoek de onderzoeksvraag kan verhelderen. Andere vragen zijn mogelijk niet causaal, maar kunnen wel etiologische inzichten bieden. Medicijngebruik moet voor elk type vraag op een andere manier worden behandeld en de overwegingen voor het omgaan met medicatie en methodologische kwesties variëren per vraag.

In **Hoofdstuk 8** concludeerden wij dat er geen universeel toepasbare methode bestaat om de problemen van confounding, ontbrekende gegevens, selectiebias en meetfouten in observationele onderzoekssettings te verminderen. Analytische beslissingen zouden niet als vanzelfsprekend moeten worden beschouwd en elke bron van vertekening moet worden behandeld op basis van contextspecifieke kennis. Een constant streven naar het verbinden van de methodologische en klinische werelden en het verbreden van de perspectieven op het omgaan met biases zal bijdragen aan de validiteit van observationeel onderzoek.

