STAND VAN ZAKEN METHODOLOGIE VAN ONDERZOEK

Rekenen met ontbrekende gegevens

Ralph C.A. Rippe, Martin den Heijer en Saskia le Cessie

Ontbrekende gegevens in medisch-wetenschappelijk onderzoek zijn soms onvermijdelijk.

Er zijn verschillende soorten ontbrekende gegevens: (a) 'missing completely at random'; (b) 'missing by design'; (c) 'missing at random' en (d) 'missing not at random'.

Als men deelnemers met ontbrekende gegevens buiten de statistische analyses laat ('complete case'-analyse), kan dit leiden tot vertekende onderzoeksresultaten en verlies van statistische 'power'.

Imputatiemethoden kunnen worden toegepast om ontbrekende waarden te schatten; door meervoudig te imputeren wordt een goed beeld verkregen van de onnauwkeurigheid van de gereconstrueerde metingen.

De meest gangbare imputatiemethoden veronderstellen dat ontbrekende gegevens 'missing at random' zijn.

Imputatie levert een grote bijdrage aan de efficiëntie en de betrouwbaarheid van schattingen, omdat maximaal gebruik wordt gemaakt van de verzamelde data.

Imputatie is zeker niet bedoeld om de lage kwaliteit van data te ondervangen.

Bij medisch-wetenschappelijk onderzoek ontbreken regelmatig gegevens van deelnemers, bijvoorbeeld als gevolg van onvolledig ingevulde vragenlijsten, of doordat bepaalde metingen niet gedaan of niet gelukt zijn. Dit maakt het analyseren van data lastig. Om die reden verwijderen veel onderzoekers de deelnemers met ontbrekende gegevens uit hun statistische analyses. Dit kan echter leiden tot vertekende onderzoeksresultaten, omdat het ontbreken van gegevens zelden willekeurig plaats vindt.

Als de bloeddruk hoofdzakelijk gemeten is bij oudere en zwaardere patiënten, en deze meting veelal ontbreekt bij jonge en slanke patiënten, dan zal de berekende gemiddelde bloeddruk niet representatief zijn voor de bloeddruk in de gehele studiepopulatie. Bovendien leidt het resoluut verwijderen van alle gegevens van een bepaalde deelnemer al snel tot verlies van veel data. Het is daarom wenselijker ontbrekende gegevens op een of andere manier te reconstrueren uit gegevens die wel bekend zijn. Zo zou de bloeddruk geschat kunnen worden op basis van het geslacht, de leeftijd, de lengte en het gewicht van de patiënt. In een dataset kunnen deze schattingen ingevuld ('geïmputeerd') worden op plaatsen waar gegevens ontbreken.

Imputatie van ontbrekende gegevens wordt steeds vaker toegepast in medisch-wetenschappelijk onderzoek. In dit artikel bespreken we in grote lijnen de principes van deze statistische methode.

LUMC, afd. Klinische Epidemiologie, Leiden.
Dr. R.C.A. Rippe, statisticus; prof.dr. M. den Heijer, internist-endocrinoloog (tevens: VUmc, afd. Endocrinologie);
dr. S. le Cessie, statisticus (tevens: afd. Medische Statistiek).
Contactpersoon: dr. S. le Cessie (cessie@lumc.nl).

SOORTEN ONTBREKENDE GEGEVENS

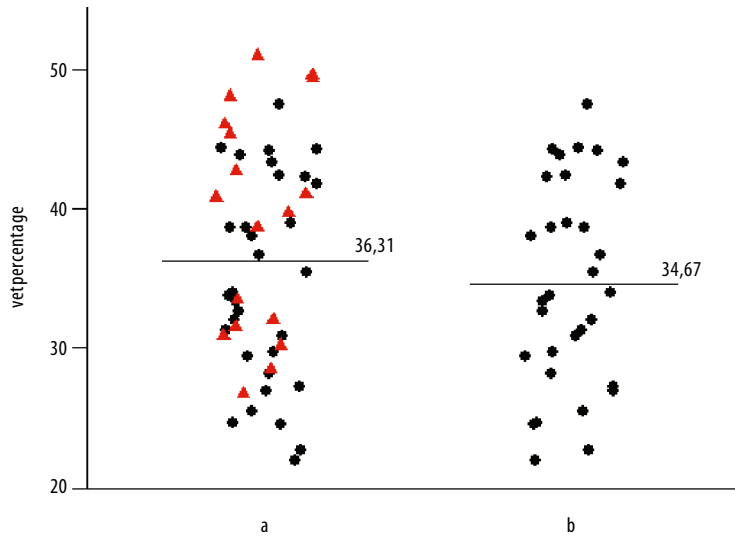
Ontbrekende gegevens kunnen het gevolg zijn van toeval of van specifieke oorzaken. Men onderscheidt 3 soorten ontbrekende gegevens,[1] die hieronder kort worden toegelicht.

Missing completely at random Een meting kan totaal willekeurig ontbreken, dus onafhankelijk van de waarde zelf en van andere wel of niet gemeten patiëntkarakteristieken. Dit heet 'missing completely at random' (MCAR). Denk bijvoorbeeld aan reageerbuisjes met bloed die op de grond kapot zijn gevallen, of aan een vragenlijst die verloren is geraakt. Een nauw aan MCAR gerelateerde variant is het opzettelijk ontbreken van metingen vanwege de studieopzet ('missing by design'). Voorbeelden hiervan zijn DNA-bepalingen en MRI-opnamen die vanwege hoge kosten alleen bij een representatieve subgroep worden uitgevoerd. Voor de overige deelnemers ontbreken dan de gegevens van deze metingen. Zowel bij MCAR als bij missing-by-design geldt dat de deelnemers van wie alle gegevens bekend zijn, representatief zijn voor de gehele studiepopulatie.

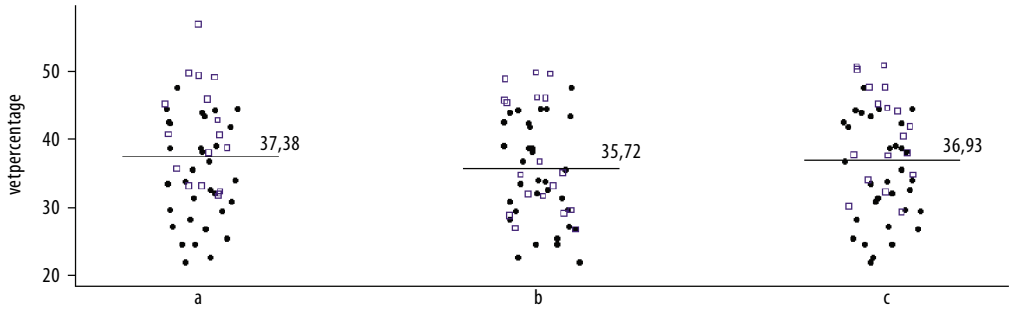
Missing at random Wanneer het ontbreken van gegevens samenhangt met andere gemeten patiëntkarakteristieken, maar daarnaast niet met de uitkomst van de variabele zelf, dan spreekt men van 'missing at random'. Dit is een verwarrende term, omdat deze suggereert dat er geen systematiek mag bestaan in het ontbreken van de gegevens, terwijl die er wel is. Bij een onderzoek in een huisartspraktijk naar risicofactoren voor hart- en vaatziekten ontbreekt de cholesterolwaarde bij jongeren vaker dan bij ouderen, omdat er bij jongeren veelal geen reden is voor een cholesterolbepaling. In dit voorbeeld bepaalt de leeftijd het ontbreken van de cholesterolwaarden bij jongeren. Op grond van variabelen als leeftijd, geslacht, lengte, gewicht en een positieve familieanamnese voor hart- en vaatziekten kan de cholesterolwaarde gereconstrueerd worden.

Not missing at random Tenslotte kan de reden voor het ontbreken van een gegeven afhangen van de uitkomst van de betreffende variabele. Dergelijke ontbrekende gegevens zijn dan 'not missing at random' (NMAR). In een dataset kunnen gegevens over inkomen ontbreken, omdat deelnemers deze niet wilden delen vanwege hun lage inkomen. In dit voorbeeld bepaalt de hoogte van het inkomen het al dan niet ontbreken van inkomensgegevens.

Onderscheid in de praktijk Bij MCAR zullen de deelnemers met en zonder ontbrekende gegevens vergelijkbaar zijn voor alle variabelen, terwijl er bij MAR en MNAR verschillen gevonden kunnen worden. Het is niet mogelijk om na te gaan of het ontbreken van gegevens at random (MAR) of niet at random is (NMAR).[2] De uitkomst van de variabele is immers niet bekend en men kan dus niet nagaan of de uitkomst samenhangt met het ontbreken ervan. In de praktijk veronderstelt men meestal dat de gegevens MAR zijn.



FIGUUR 1 (a) Vetpercentages van 50 deelnemers, waarbij de rode driehoekjes staan voor 18 deelnemers met een BMI > 31 kg/m^2. (b) Vetpercentage van 32 deelnemers met een BMI ≤ 31 kg/m^2. Van beide groepen zijn de gemiddelde vetpercentages weergegeven.



FIGUUR 2 (a-c) 3 datasets waarin telkens de ontbrekende vetpercentages van 18 deelnemers met een BMI > 31 kg/m^2 geïmputeerd zijn (blauwe vierkantjes).

EEN VOORBEELD

REKENEN ZONDER ONTBREKENDE GEGEVENS

Als men in de statistische analyse alleen de deelnemers meeneemt van wie alle gegevens compleet zijn ('complete case'-analyse), dan kunnen de resultaten van die berekeningen leiden tot een vertekend beeld van de studiepopulatie. Dit kunnen we laten zien middels een eenvoudig voorbeeld (figuur 1).

Stel dat van 50 patiënten het gemiddelde percentage lichaamsvet 36,31 bedraagt. Wanneer de vetpercentages van 18 deelnemers met een BMI > 31 kg/m^2 buiten de analyses worden gelaten, daalt het gemiddelde vetpercentage tot 34,67; dit is een onderschatting van het werkelijke gemiddelde in de studiepopulatie. Om het gemiddelde nauwkeuriger te berekenen, kunnen de ontbrekende vetpercentages worden geschat aan de hand van gegevens die wel bekend zijn, zoals de middel-heupratio, het BMI en het geslacht van de betreffende deelnemers. Het schatten en invullen van ontbrekende gegevens met behulp van regressietechnieken wordt 'imputeren' genoemd.

IMPUTATIE VAN ONTBREKENDE GEGEVENS

Imputatie van ontbrekende gegevens kan op verschillende manieren plaatsvinden. Eenvoudige methoden zijn bijvoorbeeld gebruik maken van het gemiddelde, de mediaan of de modus.[3] In bovengenoemd voorbeeld is dit geen passende oplossing: als voor alle ontbrekende vetpercentages het geobserveerde gemiddelde van 34,67 ingevuld zou worden, dan blijft het geschatte gemiddelde onveranderd.

Het is wenselijker om het vetpercentage te voorspellen op grond van bijvoorbeeld de BMI en de middel-heupratio. Voor individuele deelnemers zullen deze voorspellingen altijd een zekere mate van onnauwkeurigheid bevatten. Moderne imputatiemethoden houden hier rekening mee, waardoor voor 2 deelnemers met eenzelfde BMI en middel-heupratio toch enigszins verschillende waarden geïmputeerd kunnen worden. Dit gebeurt door een aantal keren achter elkaar de ontbrekende waarden te imputeren ('meervoudige imputatie'). Hierdoor wordt een goed beeld verkregen van de onnauwkeurigheid van de gereconstrueerde metingen. Ook dit kunnen we inzichtelijk maken aan de hand van een voorbeeld (figuur 2).

In de figuur zijn 3 keer de vetpercentages geïmputeerd van de 18 deelnemers van wie deze percentages ontbraken, op grond van hun BMI en middel-heupratio. Door rekening te houden met de onzekerheid van de schattingen, worden telkens iets andere waarden gegenereerd. Als de gemiddelden van de 3 geïmputeerde datasets gecombineerd worden, bedraagt de uiteindelijke schatting van het gemiddelde vetpercentage: (37,38 + 35,72 + 36,93)/3 = 36,67. Door imputatiemethoden toe te passen, kunnen vervolgens reëlere betrouwbaarheidsintervallen en p-waarden berekend worden.[4]

THEORETISCHE VERONDERSTELLINGEN

In de praktijk volstaat het gebruik van 5-10 geïmputeerde datasets, tenzij er veel ontbrekende gegevens zijn. Exacte grenzen waarbinnen nog betrouwbaar geïmputeerd kan worden, zijn niet te geven. In de literatuur wordt gesproken over 20-50% van de gehele steekproef, afhankelijk van de relaties tussen de metingen.[5] Met sterk gerelateerde metingen kunnen immers vrij nauwkeurig grote hoeveelheden ontbrekende gegevens worden geschat. Dankzij de moderne techniek – snelle processoren en grote werkgeheugens – is de bovengrens van het aantal geïmputeerde datasets steeds minder relevant; steeds vaker imputeren geeft steeds nauwkeurigere schattingen, terwijl nauwelijks meer rekentijd nodig is.

Voor het imputeren van ontbrekende gegevens kunnen verschillende modellen worden gebruikt. In het statistische softwareprogramma SPSS wordt bijvoorbeeld lineaire regressie gebruikt met normaal verdeelde uitkomsten om continue waarden te imputeren, en logistische regressie voor een dichotome variabele. Het is belangrijk

dat het imputatiemodel correct gespecificeerd is, omdat anders de verkeerde onderlinge relaties tussen variabelen worden gebruikt voor de schattingen. Als het verband tussen middel-heupratio en vetpercentage niet lineair maar kwadratisch zou zijn, dan zouden de geïmputeerde waarden systematisch afwijken van de werkelijke ontbrekende waarden. Verder moet de uitkomst ook meegenomen worden in het imputatiemodel, zodat de relatie tussen de geïmputeerde waarden en de uitkomst dezelfde is als tussen geobserveerde waarden en de uitkomst.[6]

Een andere aanname is dat de ontbrekende waarden MAR zijn. Als de missende waarden NMAR zijn, zoals bij gevoelige vragen over inkomen en seksuele geaardheid, leidt imputeren tot foute schattingen.

PRAKTISCHE TOEPASSING

Van 5384 deelnemers is de BMI, het geslacht, de middel-heupratio en het vetpercentage bepaald (tabel). Stel dat men wil weten hoe het vetpercentage afhangt van de middel-heupratio en het geslacht, dan kan dit bepaald worden met een lineair regressiemodel met vetpercentage als uitkomst. Uit de tabel blijkt dat als de middel-heupratio toeneemt met 0,1, het gemiddelde vetpercentage stijgt met 4,39 bij gelijkblijvend geslacht. Ook blijkt dat vrouwen gemiddeld 17,6% meer lichaamsvet hebben dan mannen, bij een gelijkblijvende middel-heupratio.

Om te kijken hoe de regressiecoëfficiënten beïnvloedt worden door ontbrekende gegevens verwijderen we geheel willekeurig (MCAR) de helft van de vetpercentagemetingen. Als we de gegevens van de subgroep zonder ontbrekende gegevens (n = 2692) gebruiken, krijgen we vrijwel dezelfde regressiecoëfficiënten. Deze waarden zijn echter onnauwkeuriger, vanwege de kleinere steekproef. Het gevolg hiervan is dat de standaardfouten groter zijn. Als we de ontbrekende gegevens meervoudig imputeren met behulp van BMI, geslacht en vetpercentagemetingen blijkt dat de regressiecoëfficiënten veel nauwkeuriger geschat kunnen worden; de standaardfouten worden immers kleiner.

Het nut van imputeren wordt nog duidelijker als de ontbrekende gegevens MAR zijn. Dit is nagebootst door de vetpercentages van personen met een BMI > 31 kg/m^2 buiten beschouwing te laten. Als enkel de deelnemers van wie alle gegevens bekend zijn (n = 4387) worden meegenomen in de statistische analyses, worden sterker afwijkende regressiecoëfficiënten gevonden; zo zakt de regressiecoëfficiënt voor de middel-heupratio van 0,439 naar 0,389. Meervoudige imputatie leidt opnieuw tot waarden die dichter bij de echte waarden liggen. We zien kleine afwijkingen die waarschijnlijk veroorzaakt worden doordat de ontbrekende waarden voorspeld zijn met een imputatiemodel dat mogelijk niet perfect is.

ANDERE METHODEN

Er zijn verschillende manieren om met ontbrekende gegevens om te gaan. Van een aantal eenvoudige methoden – imputeren van het gemiddelde, gebruik maken van de laatst gemeten waarde, of het aanmaken van een aparte categorie voor ontbrekende gegevens – is bekend dat deze tot vertekende onderzoeksresultaten kunnen leiden.[3] In specifieke situaties kan men gebruik maken van statistische modellen en 'maximum likelihood'-methoden om tot de juiste schattingen komen. Een voor-

TABEL Relatie tussen de onafhankelijke voorspellers 'middel-heupratio' en 'geslacht' en de uitkomst 'vetpercentage' op basis van een lineair regressiemodel.

voorspeller	regressiecoëfficiënt (SE)				
	alle vetpercentages (n = 5384)	50% van vetpercentages ontbreekt volledig willekeurig (MCAR)		vetpercentages van deelnemers met BMI > 31 kg/m^2 ontbreken (MAR)	
		'complete case'-analyse (n = 2692)	meervoudige imputatie	'complete case'-analyse (n = 4387)	meervoudige imputatie
middel-heupratio*	4,39 (0,11)	4,36 (0,15)	4,39 (0,13)	3,89 (0,11)	4,34 (0,13)
♀	17,6 (0,2)	17,5 (0,3)	17,5 (0,2)	16,5 (0,2)	18,2 (0,2)

SE = standaardfout; MCAR = 'missing completely at random'; MAR = 'missing at random'.

* De regressiecoëfficiënt geeft de gemiddelde stijging van het vetpercentage aan als de middel-heupratio met 0,1 toeneemt (bij gelijkblijvend geslacht). In de volledige dataset (n = 5384) stijgt het gemiddelde vetpercentage dus met 4,39 als de middel-heupratio met 0,1 toeneemt (bij gelijkblijvend geslacht) en is het vetpercentage bij vrouwen 17,6 % hoger, bij een gelijkblijvende middel-heupratio.

beeld hiervan zijn de 'mixed effect'-modellen die veel gebruikt worden bij het modelleren van herhaalde metingen met ontbrekende gegevens.

Het voordeel van meervoudige imputatiemethoden is dat ze breed toepasbaar en relatief eenvoudig uit te voeren zijn. Ook kan hiermee een mengsel van ontbrekende gegevens in continue en categorische variabelen gehanteerd worden.[7,8] Imputatiemethoden zijn aanvankelijk bedoeld als reparatie achteraf om 'accidenteel' ontbrekende waarden zo goed mogelijk te schatten, maar deze methoden werken ook als metingen vanuit financiële overwegingen niet zijn gedaan.

Hoewel imputatie theoretisch goed onderbouwd is, moet men zich wel bewust zijn van de aannames. De meest gangbare imputatiemethoden veronderstellen dat ontbrekende gegevens MAR zijn. Verder moeten de imputatiemodellen correct zijn; men moet voldoende informatie hebben om de ontbrekende gegevens met redelijke zekerheid te kunnen imputeren. Imputatie is zeker niet bedoeld om de lage kwaliteit van data te ondervangen. Mits goed gebruikt, levert imputeren een grote bijdrage aan de efficiëntie en de betrouwbaarheid van schattingen, omdat maximaal gebruik wordt gemaakt van de verzamelde data.

LEERPUNTEN

- **Er zijn verschillende soorten ontbrekende gegevens: (a) 'missing completely at random'; (b) 'missing by design'; (c) 'missing at random' en (d) 'missing not at random'.**
- **Statistische analyses waarin deelnemers met ontbrekende gegevens zijn weggelaten, leiden tot vertekening van onderzoeksresultaten en verlies van statistische 'power'.**
- **Imputatiemethoden kunnen worden toegepast om ontbrekende waarden te schatten; meervoudig imputeren geeft een goed beeld van de onnauwkeurigheid van de gereconstrueerde metingen.**
- **Imputatie levert een grote bijdrage aan de efficiëntie en de betrouwbaarheid van schattingen, omdat maximaal gebruik wordt gemaakt van de verzamelde data.**
- **Imputatie is niet bedoeld om de lage kwaliteit van data te ondervangen.**

CONCLUSIE

De meeste medisch-wetenschappelijke studies hebben te maken met ontbrekende waarnemingen, bijvoorbeeld doordat een meting bij een deelnemer mislukt is of om bepaalde redenen niet uitgevoerd kon worden. Het beperken van de analyses tot deelnemers van wie de waarnemingen volledig zijn, kan tot een aanzienlijk verlies van data leiden als er veel variabelen gemeten zijn. Dit zou een inefficiënt gebruik van middelen betekenen.

In dit artikel beschreven we het principe van meervoudige imputatie en gaven we aan onder welke veronderstellingen de uitkomsten correct geschat kunnen worden. We lieten zien dat het weglaten van deelnemers met ontbrekende gegevens kan leiden tot vertekende onderzoeksresultaten en conclusies, met name als de gegevens niet geheel willekeurig ontbraken. Imputatie van ontbrekende gegevens is daarom niet alleen nuttig voor het vergroten van statistische 'power', het kan bovendien vertekening in de schattingen voorkomen.

Belangenconflict en financiële ondersteuning: geen gemeld.

Aanvaard op 2 januari 2013

Citeer als: Ned Tijdschr Geneeskd. 2013;157:A5539

KIJK OOK OP WWW.NTVG.NL/KLINISCHEPRAKTIJK

LITERATUUR

1 Rubin DB. Multiple imputation for nonresponse in surveys. New York: John Wiley & Sons; 1987.
2 Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods. 2002;7:147-77.
3 Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59:1087-91.
4 Rubin D. Inference and missing data. Biometrika. 1976;63:581-92.
5 Graham JW, Olchoswksi AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prev Sci. 2007;8:206-13.
6 Moons KG, Donders AR, Stijnen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol. 2006;59:1092-101.
7 Van Buuren S, Boshuizen SC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999;18:681-94.
8 Van Buuren S. Flexible imputation of missing data. Boca Raton: Chapman & Hall/CRC Press; 2012.