Wat is een goed biomedisch

citation and similar papers at core.ac.uk

brought

provided by Ghent University A

James Lind Revisited

Erik Weber & Leen De Vreese[1]

Abstract – The Scottish physician James Lind is famous for the experiments on potential cures of scurvy which he performed in 1747 on board of the British navy vessel HMS Salisbury. In this paper we use Lind's experiments and their shortcomings to explain important contemporary methodological standards for biomedical research and the rationale behind these standards. We also describe the genesis of these methodological standards.

✧ ✧ ✧

Inleiding

De Schotse arts James Lind was een sleutelfiguur in het onderzoek naar de oorzaken van, en remedies tegen, scheurbuik. Zijn bekendheid heeft hij vooral te danken aan het experimenteel onderzoek dat hij in 1747 uitvoerde aan boord van het Britse marineschip HMS Salisbury. In dat onderzoek vergeleek hij de werking van zes – in die tijd populaire – behandelingen van scheurbuik. Over dit experimenteel onderzoek en allerlei andere aspecten van scheurbuik (o.a. zijn theorieën over de oorzaak ervan) publiceerde Lind in 1753 het boek *A Treatise of the Scurvy*. Ter gelegenheid van de 200ste verjaardag van de eerste publicatie werd het boek opnieuw uitgegeven onder de titel '*Lind's Treatise on Scurvy. A Bicentenary Volume Containing a Reprint of the First Edition of* A Treatise of the Scurvy *by James Lind*' (Stewart & Guthrie 1953).

Het onderzoek van Lind is om een aantal redenen een interessant onderwerp voor een gevalstudie. Ten eerste is het materiaal, dankzij de herdruk van het boek, nog steeds gemakkelijk beschikbaar. Ten tweede is het zo dat Lind zes biomedische experimenten deed (één per onderzochte behandeling) met een basisstructuur zoals we die nu nog steeds zien in biomedische experimenten waarin de

1. Erik Weber is als gewoon hoogleraar verbonden aan het Centrum voor Logica en Wetenschapsfilosofie van de Universiteit Gent. Leen De Vreese is postdoctoraal onderzoeker in hetzelfde centrum. De auteurs danken Roxan Degeyter voor het nalezen van deze tekst. Dit artikel is vrij beschikbaar onder de Creative Commons licentie CC-BY-NC-ND.

efficiëntie van therapieën onderzocht wordt. Maar – en dit is de derde en belangrijkste reden – de experimenten zijn vanuit hedendaags perspectief heel pover uitgevoerd.

Die tekortkomingen maken het onderzoek van Lind een prima uitgangspunt voor wat we in dit artikel willen doen: uiteenzetten aan welke methodologische standaarden biomedische experimenten op dit moment moeten voldoen om als adequaat aanzien te worden, en de motivering voor deze standaarden verduidelijken. We gebruiken het experimenteel onderzoek van Lind dus als contrastvoorbeeld, waarin allerlei cruciale eigenschappen afwezig zijn.

De structuur van dit artikel is als volgt. In de tweede sectie laten we Lind zelf aan het woord over zijn experimenten. In de derde sectie leggen we uit waarom zijn onderzoek wel degelijk experimenteel mag genoemd worden: het omvat zes biomedische experimenten. In de vierde en vijfde sectie behandelen we de gebreken die naar boven komen als we de experimenten van Lind toetsen aan hedendaagse methodologische standaarden. We leggen uit wat die hedendaagse standaarden inhouden en wat de redenen zijn om ze te hanteren.

Aangezien de werkwijze van Lind in 1747 sterk verschilt van de hedendaagse werkwijze, kunnen we de vraag stellen: wanneer zijn de huidige methodologische standaarden ontstaan en algemeen aanvaard geraakt? In de laatste sectie laten we zien dat het gaat om een proces dat zich in de loop van de 20ste eeuw geleidelijk voltrokken heeft.

James Lind aan het woord

Van 1746 tot 1748 was James Lind scheepsarts op de Salisbury. In de zomer van 1746 maakte hij een eerste uitbraak van scheurbuik mee, waarbij 80 van de 350 bemanningsleden ziek werden. Tijdens een tweede uitbraak in 1747 deed Lind zijn beroemde experiment.

Door het eerder aangehaalde boek hebben we een goed beeld van wat hij deed en waarom. Lind omschrijft het doel van zijn experiment als volgt:

> [...] to relate the effects of several medicines tried at sea in this disease, on purpose to discover what might promise the most certain protection against it upon that element. (Stewart & Guthrie 1953, p. 144)

De in het experiment gevolgde werkwijze omschrijft hij als volgt:

> On the 20th of May 1747, I took twelve patients in the scurvy, on board the Salisbury at sea. Their cases were as similar as I could have them. They all in general had putrid gums, the spots and lassitude, with weakness of their knees. They lay together in one place, being a

> proper apartment for the sick in the fore-hold; and had one diet common to all, ... Two of these were ordered each a quart of cider a-day. Two others took twenty-five gutts of elixir vitriol three times a-day, upon an empty stomach; using a gargle strongly acidulated with it for their mouths. Two others took two spoonfuls of vinegar three times a-day, upon an empty stomach; having their gruels and their other food well acidulated with it, as also the gargle for their mouth. Two of the worst patients, with the tendons in the ham rigid, (a symptom none of the rest had), were put under a course of sea-water. Of this they drank half a pint every day, and sometimes more or less as it operated, by way of gentle physic. Two others had each two oranges and one lemon given them every day. These they eat with greediness, at different times, upon an empty stomach. They continued but six days under this course, having consumed the quantity that could be spared. The two remaining patients, took the bigness of a nutmeg three times a-day, of an electuary recommended by an hospital-surgeon [...]; using for common drink [a tamarind decoction] [...] (Stewart & Guthrie 1953, pp. 145-146)

De geteste therapieën zijn dus: (1) cider, (2) een elixir met zwavelzuur ('vitriol'), (3) azijn, (4) zeewater, (5) citrusvruchten (sinaasappel en citroen) en (6) een specifieke medicinale pasta (die o.a. look, mosterdzaad en mirrehars bevatte).

Na twee weken konden de volgende effecten geobserveerd worden:

> [...] the most sudden and visible good effects were perceived from the use of the oranges and lemons; one of those who had taken them, being at the end of six days fit for duty. [...] The other was the best recovered of any in his condition [...] Next to the oranges, I thought the cider had the best effects. It was indeed not very sound. However, those who had taken it, were in a fairer way of recovery than the others at the end of the fortnight [...] As to the elixir of vitriol, I observed that the mouths of those who had used it by way of gargarism, were in a much cleaner and better condition than many of the rest, especially those who used the vinegar; but perceived otherwise no good effects from its internal use upon the other symptoms [...] There was no remarkable alteration upon those who took the electuary and tamarind decoction, the sea-water, or vinegar, upon comparing their condition, at the end of the fortnight, with others who had taken nothing but a little lenitive electuary and *cremor tartar*, at times, in order to keep their belly open; or a gentle pectoral in the evening, for relief of their breast. (Stewart & Guthrie 1953, 146-148)

Nu we weten hoe Lind zijn onderzoek heeft uitgevoerd, kunnen we de volgende stap zetten: uitleggen dat dit zes biomedische experimenten zijn met een basisstructuur zoals we die nu nog steeds zien in biomedische experimenten waarin de efficiëntie van therapieën onderzocht wordt.

De zes experimenten van Lind

Een biomedisch experiment wordt steeds uitgevoerd op een bepaalde *proefgroep* van mensen die allemaal de medische aandoening hebben waarvoor de geteste behandeling verondersteld wordt een mogelijke oplossing te bieden. Over hoe die proefgroep idealiter tot stand komt hebben we het later (in de vijfde sectie van dit artikel). Belangrijk is nu dat in een biomedisch experiment met betrekking tot de werking van een mogelijke therapie de proefgroep steeds wordt opgedeeld in een experimentele groep (die de geteste behandeling ondergaat) en een controlegroep (die de geteste behandeling niet ondergaat).

Stel dat je een nieuw geneesmiddel wil testen, nl. een oogzalf waarvan je vermoedt dat die bepaalde oogontstekingen kan verhelpen. Een mogelijke werkwijze bestaat er dan in dat je (i) een proefgroep samenstelt van 2000 personen die allemaal kampen met de bedoelde vorm van oogontsteking, en (ii) dat je daarvan 1000 personen in de experimentele groep plaatst en de andere 1000 in de controlegroep. Als experimentator moet je er vervolgens voor zorgen dat de 1000 mensen in de experimentele groep gedurende een aantal dagen regelmatig (volgens het voorgeschreven ritme en dosis) de zalf aanbrengen. De 1000 mensen in de controlegroep mogen in geen geval de geteste zalf gebruiken (over wat die dan wel mogen/moeten doen, hebben we het in de volgende sectie).

In elk experiment is er *experimentele manipulatie*: we zorgen ervoor dat de experimentele groep de therapie ondergaat, en de controlegroep niet. Dit is het tweede cruciale kenmerk. In ons voorbeeld is het feit dat we erop toezien dat mensen in de experimentele groep de zalf aanbrengen en die in de controlegroep juist niet, de specifieke vorm van experimentele manipulatie die bij het verondersteld experiment wordt toegepast.

Een laatste kenmerk van een experiment is dat er na verloop van tijd *observaties* met betrekking tot het gewenste effect worden gedaan, zowel in de experimentele groep als in de controlegroep. In het oogzalf-experiment kunnen we bijvoorbeeld na een week nagaan bij hoeveel van de 1000 personen in de experimentele groep de oogontsteking verdwenen is. Een analoge telling kan gebeuren in de contolegroep.

Samengevat heeft een biomedisch experiment drie cruciale kenmerken: (a) er is een proefgroep die onderverdeeld wordt in een experimentele groep en in een controlegroep, (ii) er is experimentele manipulatie, en (iii) er is na een bepaalde periode een vergelijkende observatie van de verhoopte resultaten.

Als we met deze achtergrondinformatie terugkijken naar wat Lind gedaan heeft, dan kunnen we stellen dat hij *zes* experimenten heeft gedaan. In het eerste experiment (het 'cider-experiment') bestaat de proefgroep uit de twee zieke bemanningsleden die cider als therapie toegewezen kregen (experimentele groep) plus alle niet-behandelde scheurbuiklijders aan boord van het schip (controlegroep). In het tweede experiment (het 'vitriool-experiment') bestaat de proefgroep uit de twee zieke bemanningsleden die het elixir met zwavelzuur toegewezen kregen (experimentele groep) plus weer alle niet-behandelde scheurbuiklijders aan boord van het schip (controlegroep). Hetzelfde geldt voor de vier andere geteste behandelingen. Er is dus telkens een experimentele groep met twee zieke bemanningsleden en een controlegroep (de niet behandelde scheurbuiklijders aan boord van de Salisbury) die voor elk van de zes experimenten dezelfde is.

Er is ook experimentele manipulatie: Lind zorgt ervoor dat de personen in de experimentele groep eten wat ze moeten eten. En de controlegroep krijgt het gewone dieet, waarin geen enkele van de geteste voedingssupplementen aanwezig is. Wat de bemanningsleden in de controlegroep wel kregen, was een pijnstillende pasta ("lenitive electuary"), een laxeermiddel ("cremor tartar") en/of een hoestsiroop ("pectoral"). Deze producten kunnen een effect hebben op bepaalde symptomen (pijn, constipatie) maar zeker de scheurbuik niet genezen.

Tot slot is er ook vergelijkende observatie. Na twee weken stelt Lind vast dat de personen in de experimentele groep van het 'azijn-experiment' (#3), het 'zeewater-experiment' (#4) en het medicinale-pasta-experiment (#6) er helemaal niet beter aan toe zijn dan de bemanningsleden in de controlegroep. Bij de personen in de experimentele groep van het 'vitriool-experiment' (#2) is de toestand van de mond beter, maar is er geen verbetering van de andere symptomen. Bij de personen in de experimentele groep van het 'citrusvruchten-experiment' (#5) rapporteert Lind een substantiële verbetering van de gezondheidstoestand: één persoon is terug geschikt voor dienst, de andere is diegene die het dichtst bij herstel staat van alle andere proefpersonen. Bij het 'cider-experiment' (#1) rapporteert hij dat de twee personen in de experimentele groep er beter aan toe zijn dan die in de controlegroep (hierover is Lind eerder vaag).

Samengevat kunnen we stellen dat Lind zes biomedische experimenten deed, waarbij drie belangrijke kenmerken (opdeling in experimentele groep en controlegroep, experimentele manipulatie en vergelijkende observatie) aanwezig zijn. In de twee volgende secties behandelen we de tekortkomingen vanuit hedendaags perspectief.

Een belangrijk verschil tussen de werkwijze van Lind en de hedendaagse onderzoekspraktijk is dat Lind geen *inferentiële statistiek* gebruikte om uit zijn observaties besluiten te trekken. Dat kon hij ook helemaal niet, want de inferentiële statistiek bestond nog niet in zijn tijd (meer hierover in de laatste sectie van

dit artikel). Een aantal van de tekortkomingen in de werkwijze van Lind zijn van die aard dat ze de toepassing van methodes uit de inferentiële statistiek problematisch maken. Maar er zijn ook problemen die daar niets mee te maken hebben. Die problemen behandelen we in de volgende sectie. In de vijfde sectie bespreken we de tekortkomingen die wel gelinkt zijn aan het gebruik van statistische methodes.

Toevalsverdeling en dubbelblindheid

We behandelen achtereenvolgens drie gebreken in Linds experimenten: (i) afwezigheid van willekeurige indeling, (ii) geen neutralisatie van placebo-effecten en (iii) geen neutralisatie van onderzoekerseffecten. Telkens leggen we uit wat de hedendaagse standaard inhoudt, hoe Lind ervan afweek en waarom de standaard in huidig biomedisch onderzoek wordt gehanteerd.

Geen willekeurige indeling

Een eerste gebrek is dat de experimenten van Lind geen *toevals*experimenten zijn. Een toevalsexperiment is een experiment zoals beschreven in de vorige sectie, waarbij de indeling in experimentele en controlegroep *willekeurig* gebeurt (d.w.z. aan de hand van een toevalsprocedure, bijvoorbeeld een lottrekking). Wat Lind had kunnen doen om een toevalsexperiment te creëren, is papiertjes met de namen van alle zieke bemanningsleden in een bokaal stoppen en er daar telkens (voor elk experiment) twee blindelings uithalen. Er is geen reden om aan te nemen dat hij een dergelijke toevalsmatige selectie heeft georganiseerd, wel integendeel. Lind zegt dat hij twaalf patiënten nam die sterk gelijkend waren (cfr. “as similar as I could have them”, in het tweede citaat in de sectie over Lind). En twee van de patiënten die er het slechtste aan toe waren, werden aan een zeewater-kuur onderworpen. Dit zijn aanwijzingen dat Lind zelf heeft gekozen wie de twaalf uitverkorenen waren, en wie in welke experimentele groep terechtkwam.

Willekeurige indeling is belangrijk omdat we ervoor moeten zorgen dat de geobserveerde verschillen (wat betreft het beoogde effect) tussen experimentele en controlegroep enkel kunnen verklaard worden door de behandeling die de experimentele groep gekregen heeft. Om dit te illustreren werken we verder met het voorbeeld uit de voorgaande sectie. Stel dat we na een week vaststellen dat in de experimentele groep (de 1000 personen die oogzalf aangebracht hebben) 800 mensen genezen zijn, d.w.z. de oogontsteking is bij hen verdwenen. In de controlegroep is de ontsteking slechts bij 300 personen verdwenen. Als de groepen *niet* via een toevalsprocedure samengesteld zijn, dan zijn er tal van alternatieve verklaringen voor dit nochtans grote verschil mogelijk. Stel dat je bijvoorbeeld de 1000

jongste deelnemers in de experimentele groep hebt ingedeeld en de 1000 oudste in de controlegroep. Dan kan het verschil verklaard worden door het feit dat de graad van spontaan herstel bij jongeren hoger is dan bij ouderen. Of stel dat je de proefpersonen naar hun beroep hebt gevraagd, en degenen die het meest buiten werken in de experimentele groep hebt gezet. Dan kan het verschil in genezingsgraad mogelijk verklaard worden door een positieve invloed van zonlicht. Om alle denkbare alternatieve verklaringen uit te sluiten, moeten we een toevalsprocedure toepassen: *alle* andere mogelijke beïnvloedende factoren (jong/oud; veel/weinig zon; …) zijn dan in principe in gelijke mate aanwezig in beide groepen, en kunnen dus geen verklaring vormen voor het verschil.

Geen neutralisatie van placebo-effecten

Een tweede gebrek is dat de experimenten van Lind niet *blind* zijn. Een blind experiment is een experiment waarbij de proefpersonen niet weten of ze in de experimentele groep dan wel in de controlegroep zitten. In de experimenten van Lind is duidelijk dat de proefpersonen weten of ze in een experimentele groep zitten (nl. wanneer ze iets speciaals krijgen bovenop hun normale dieet dat vroeger niemand kreeg) en wanneer niet (wanneer ze niets nieuws krijgen; de hoestsiroop en pijnstillers werden al bedeeld voor het experiment startte).

In de hedendaagse biomedische methodologie wordt blindheid bewerkstelligd door gebruik te maken van zogenaamde *placebo's*. Een placebo-middel is een inactieve imitatie van de onderzochte substantie of behandeling. Het kan een injectie, pil, operatie of een andere behandeling zijn. Placebo-injecties bestaan meestal uit een zoutoplossing. Een placebo-pil kan o.a. gemaakt worden door een capsule te vullen met aardappelzetmeel, maïszetmeel of suiker. In ons oogzalf-voorbeeld kan aan de controlegroep als placebo-middel een zogenaamde medicinale 'basiszalf' gegeven worden die normaal gezien gebruikt wordt om werkzame bestanddelen in te mengen.

De eis om experimenten blind te maken kan – net als de eis van toevalsverdeling – begrepen worden vanuit het feit dat alternatieve verklaringen voor het geobserveerde verschil tussen experimentele en controlegroep moeten uitgesloten worden. Door een toevalsverdeling kan je alle denkbare externe factoren als verklaring elimineren, maar niet de *verwachtingen* van de proefpersonen. Als de controlegroep helemaal geen behandeling zou krijgen, dan zou er een verschil in verwachtingen zijn: de personen die behandeld worden, zouden verwachten om snel te genezen, de andere niet. Die verwachtingen kunnen een positief effect op de gezondheidstoestand hebben. Dit fenomeen is bekend als het *placebo-effect*. In een blind experiment worden placebo-effecten niet uitgeschakeld, maar gelijk verdeeld. Bij een goed functionerend placebo is er geen verschil in verwachtingen, en

dus kan het placebo-effect geen verklaring zijn voor de geobserveerde verschillen in de mate van genezing. Zoals de wetenschapsfilosoof David Teira het uitdrukt:

> The aim of a randomized allocation of concealed treatments is to distribute evenly these expectations across the arms of the trial to prevent them from having an effect on the outcome. (2013, p. 361)

'Concealed treatments' verwijst hier naar blindheid van experimenten.

Geen neutralisatie van onderzoekerseffecten

Een derde gebrek is dat de experimenten van Lind niet *dubbelblind* zijn. Een dubbelblind experiment is een blind experiment waarbij ook de onderzoeker niet weet wie in de experimentele dan wel in de controlegroep zit. Dubbelblinde experimenten zijn de norm in biomedisch onderzoek naar de effectiviteit van behandelingen:

> The necessity of keeping experimenters blind is well recognized in randomized drug trials, for example. In fact, no drug trial is taken completely seriously unless it has followed elaborate 'double-blind procedures,' in which neither the subjects nor the experimenters know who is in the experimental and who is in the control group. (Rosnow 2001, p. 5122)

Een handige manier om een onderzoek dubbelblind te maken, is taakverdeling. Daarbij is er één onderzoeker (de 'behandelaar') die instaat voor de toevalsmatige opdeling en de experimentele manipulatie. Die onderzoeker weet wie in welke groep zit, maar registreert geen resultaten. Een tweede onderzoeker (de 'evaluator') weet niets af van de opdeling en registreert na de vooropgestelde tijd (bv. 1 week in ons oogzalf-voorbeeld, 2 weken in de experimenten van Lind) de resultaten. Van dit soort opdeling was geen sprake bij de experimenten van Lind: hij deed alles zelf.

Dubbelblinde experimenten zijn noodzakelijk door een combinatie van twee factoren. Ten eerste worden mogelijke therapieën niet zomaar op mensen getest. Er moeten aanwijzingen zijn (op basis van bv. experimenten met dieren of theoretische kennis uit de biochemie) dat de therapie kan werken. Dat betekent dat de onderzoekers bepaalde vermoedens hebben: ze vermoeden dat de therapie werkt, anders zouden ze die niet testen. Ten tweede is het zo dat de resultaten die geobserveerd en genoteerd worden, meestal geen zwart-wit karakter hebben. Er komt classificatie en interpretatie aan te pas. Zo moest Lind zijn patiënten onderverdelen in "geschikt voor dienst", "nog niet genoeg hersteld" en "nog even slecht

als twee weken geleden". In ons oogzalf-voorbeeld moet elke patiënt ondergebracht worden in de categorie "ontsteking verdwenen" of "ontsteking niet verdwenen" (eventueel kunnen er fijnere categorieën zijn, maar dat verandert niets aan het feit dat classificatie van resultaten noodzakelijk is). Om zeker te zijn dat de verwachtingen van onderzoekers geen invloed kunnen hebben op de classificatie en rapportering van de resultaten, moet je ervoor zorgen dat de personen die de resultaten observeren en verwerken geen gerichte verwachtingen kunnen hebben ten aanzien van specifieke patiënten. Dat kan door de hierboven beschreven taakverdeling: de rapporteurs weten dan niet wie in de experimentele en de controlegroep zit. Zelfs als ze vermoeden dat de therapie werkt, hebben ze ten aanzien van geen enkele specifieke patiënt een verwachting, omdat ze niet weten of een specifieke patiënt al dan niet echt behandeld is.

De onderliggende logica is dezelfde als die bij toevalsverdeling en placebomiddelen. We moeten ervoor zorgen dat er voor het geobserveerde verschil tussen experimentele en controlegroep slechts één mogelijke verklaring overblijft, nl. de aanname dat de therapie werkt. Alternatieve verklaringen moeten uitgesloten worden. Door een experiment dubbelblind te maken, sluiten we (na de twee hierboven besproken verklaringen) ook een derde mogelijke verklaring uit: een vertekende rapportering van de resultaten.

Inferentiële statistiek en de correcte toepassing ervan

We leggen eerst uit wat inferentiële statistiek is en welke rol deze speelt in wetenschappelijk onderzoek. Daarna bespreken we meer specifiek de rol van inferentiële statistiek in biomedisch onderzoek naar de efficiëntie van therapieën. Dit laat ons toe om twee bijkomende gebreken van Linds onderzoek te identificeren. Tot slot vestigen we de aandacht op enkele methodologisch interessante aspecten van Linds werk.

De rol van inferentiële statistiek in wetenschappelijk onderzoek

Inferentiële statistiek is een verzameling van statistische technieken waarbij je schattingen over populatiekenmerken doet op basis van steekproefgegevens. Stel dat je wil weten hoeveel procent van de Vlaamse huisartsen rookt. In principe is het mogelijk om alle Vlaamse huisartsen te ondervragen, maar dat kost veel tijd en geld. Hetzelfde probleem stelt zich wanneer je bv. wil weten hoeveel procent van de Vlamingen een smartphone heeft (het probleem is hier nog sterker aanwezig: er zijn meer Vlamingen dan Vlaamse huisartsen). Deze voorbeelden illustreren een situatie die typerend is voor veel onderzoek in de biomedische wetenschappen, de gedragswetenschappen en de sociale wetenschappen: we zijn

geïnteresseerd in een kenmerk van een (relatief grote) *populatie*, maar we kunnen niet alle individuen in die populatie observeren omdat dit te veel middelen en tijd vergt. In principe kan het wel, want de populaties zijn eindig; maar in de praktijk kan het niet. Daarom onderzoeken/bevragen we een beperkte deelverzameling van de populatie: de steekproef.

Het wetenschappelijke standaardproces om met deze situatie om te gaan bestaat uit drie stappen:

a. Neem een representatieve steekproef uit de populatie waarin je geïnteresseerd bent.
b. Doe de nodige observaties van de kenmerken van de individuen in die steekproef.
c. Gebruik een gepaste methode uit de inferentiële statistiek om uit de observaties in de steekproef een verantwoorde conclusie te trekken over de eigenschap van de populatie.

Dit proces is alomtegenwoordig in de sociale wetenschappen, de gedragswetenschappen en ook in de biomedische wetenschappen. Het onderzoek naar het percentage rokers onder de Vlaamse huisartsen kan bijvoorbeeld als volgt aangepakt worden:

a'. Stel een representatieve steekproef van 1000 Vlaamse artsen samen.
b'. Ga na hoeveel van die 1000 huisartsen roken.
c'. Bereken een confidentie-interval rond de steekproefproportie die je bekomen hebt in (b').

Het berekenen van confidentie-intervallen is een belangrijke techniek uit de inferentiële statistiek. Voor we die uitleggen, gaan we eerst in op een ander belangrijk concept: representativiteit (cfr. stap (a)).

Een steekproef is *representatief* als en alleen als elk element uit de populatie waaruit getrokken is, evenveel kans had om in de steekproef terecht te komen. Een representatieve steekproef kan je bekomen door willekeurig (dus aan de hand van een toevalsprocedure) te selecteren uit de *volledige* populatie waarover je iets wil te weten komen. Een klassiek voorbeeld waarin het belang van representativiteit geïllustreerd wordt, is de opiniepeiling uitgevoerd in 1936 in de Verenigde Staten door het magazine *Literary Digest*. Ongeveer tien miljoen mensen werden willekeurig geselecteerd uit telefoonboeken en registratielijsten van auto's uit alle delen van het land. Ze kregen een formulier toegestuurd waarin gevraagd werd voor welke presidentskandidaat ze zouden stemmen, de republikein Alfred Landon of de democraat Franklin D. Roosevelt. Meer dan twee miljoen formulieren werden ingevuld terugbezorgd. De telling leverde een duidelijke meerder-

heid voor Landon op (55% tegenover 41% voor Roosevelt). Toch won Roosevelt de verkiezingen met een score van bijna 61%. De steekproef van *Literary Digest* was niet representatief, omdat enkel telefoon- en autobezitters erin terecht konden komen.

De werkwijze van *Literary Digest* is een voorbeeld van wat men 'convenience sampling' noemt (Forster 2001, p. 13467). Daarbij wordt een steekproef samengesteld uit een deelverzameling van de totale populatie die voor de onderzoeker gemakkelijk toegankelijk is. Een typisch voorbeeld is een onderzoeker aan een universiteit die zijn/haar steekproef selecteert uit de aan de universiteit ingeschreven studenten, terwijl hij/zij iets wil weten over alle mensen in een bepaalde gemeenschap. Dergelijke steekproeven zijn niet representatief, omdat er willekeurig geselecteerd wordt uit een specifieke deelverzameling van de populatie in plaats van uit de populatie als geheel. Dit was ook het geval bij de *Literary Digest* poll. Als er een correlatie is tussen lidmaatschap van de deelverzameling (in dit geval: die van auto- of telefoonbezitters) en de onderzochte eigenschap (in dit geval: politieke voorkeur) dan geeft een steekproef die tot stand gekomen is door 'convenience sampling' een vertekend beeld. Dat was het geval in de poll: mensen die in 1936 in de VS een telefoon en/of een auto bezaten, behoorden tot de hogere inkomenscategorieën en waren eerder geneigd om op Landon te stemmen dan mensen die geen auto en geen telefoon bezaten.[2]

Nu we weten wat representativiteit is, bekijken we het andere concept dat we hierboven gebruikt hebben: confidentie-interval. Stel dat 160 artsen in de steekproef roken. Dan hebben we een *steekproefproportie* van 0.16. De steekproefproportie is de verhouding tussen het aantal individuen in de steekproef dat de onderzochte eigenschap heeft (hier: 160 rokers) en het totale aantal mensen in de steekproef (hier: 1000). De vraag is nu: wat kunnen we uit deze steekproefproportie besluiten over de populatie als geheel (de verzameling van alle Vlaamse huisartsen) waarin we eigenlijk geïnteresseerd zijn? Het antwoord is dat we onder meer:

- met 95% zekerheid kunnen stellen dat de verhouding in de populatie als geheel in het interval [0.14, 0.18] ligt; en
- met 99% zekerheid kunnen zeggen dat de verhouding in de populatie als geheel in het interval [0.13, 0.19] ligt.

Met andere woorden: we kunnen er 95% zeker van zijn dat tussen de 14% en 18% van de Vlaamse huisartsen rookt; en we kunnen er 99% zeker van zijn dat tussen de 13% en 19% van de Vlaamse huisartsen rookt.

2. Zie Squire 1988 voor een gedetailleerde uiteenzetting van de problemen met de poll van *Literary Digest*.

We kunnen hier niet verder ingaan op de manier waarop deze intervallen berekend worden.[3] Belangrijk is dat we op basis van een exacte steekproefproportie (0.16) met een bepaalde zekerheidsgraad (95%, 99%) kunnen besluiten dat de verhouding in de populatie als geheel in een bepaald interval rond die exacte waarde ligt: 0.16 is het middelpunt van de beide intervallen. Bij het eerste interval is er een foutenmarge van 0.02, bij het tweede interval een foutenmarge van 0.03. Merk op dat het besluit altijd een interval betreft, nooit een exacte waarde (we mogen zeker niet uit onze steekproef besluiten dat 16% van alle Vlaamse huisartsen rookt).

Inferentiële statistiek in biomedisch onderzoek naar de efficiëntie van therapieën

In de derde sectie van dit artikel hebben we gezien dat er bij een biomedisch experiment steeds een vergelijkende observatie gebeurt: er worden op een bepaald moment vaststellingen gedaan met betrekking tot de toestand van de patiënten in zowel de experimentele groep als de controlegroep. Die vaststellingen worden steeds met elkaar vergeleken en er wordt een verschil in proportie berekend. In ons oogzalf-voorbeeld hebben we verondersteld dat er in de experimentele groep na een week 800 mensen genezen zijn (een proportie van 0.80) en in de controlegroep 300 (een proportie van 0.30). Het *verschil* tussen deze twee proporties is 0.50.

De vraag is nu: wat kunnen we uit dit verschil in de steekproef besluiten over het al dan niet bestaan van een verschil in de populatie als geheel? Kunnen we op basis van het (in dit geval relatief grote) verschil besluiten dat, als we iedereen met de aandoening oogzalf zouden laten smeren, er meer mensen genezen zouden zijn (nl. 80% in plaats van 30%) na een week? Inferentiële statistiek laat ons toe om die vraag te beantwoorden, omdat we confidentie-intervallen voor het waargenomen verschil kunnen berekenen. Net als in de voorgaande sectie gaan we niet dieper in op de berekeningswijze, maar kijken we naar een aantal mogelijke resultaten. Op basis van de gegevens van het experiment (2 keer 1000 mensen, proporties van 0.80 en 0.30 en dus een verschil van 0.50) kunnen we onder meer:

- met 95% zekerheid stellen dat het verschil in de populatie als geheel in het interval [0.46, 0.54] ligt; en
- met 99% zekerheid stellen dat het verschil in de populatie als geheel in het interval [0.45, 0.55] ligt.

3. Lezers die meer willen weten kunnen hoofdstuk 2 van Weber et al. 2016 lezen. De berekeningswijze komt ook aan bod in alle inleidende handboeken statistiek.

Twee bijkomende gebreken van de experimenten van Lind

Als Lind in deze tijd biomedisch onderzoeker zou zijn, dan zou hij tijdens zijn academische opleiding geleerd hebben om relevante technieken uit de inferentiële statistiek toe te passen. Hij zou ook geleerd hebben aan welke voorwaarden een experiment moet voldoen om de resultaten van statistische berekeningen betrouwbaar te maken.

Een eerste voorwaarde hebben we al gezien: representativiteit. Confidentie-intervallen en andere technieken uit de inferentiële statistiek leveren enkel betrouwbare resultaten op indien de steekproef waarvan vertrokken wordt, representatief is voor de populatie waarin we geïnteresseerd zijn. Laten we even aannemen dat de ambities van Lind beperkt waren en dat hij enkel wilde nagaan hoe scheurbuik kan genezen worden bij zeelieden van de Britse marine (en niet bij andere mensen). Dan moest hij als goede onderzoeker zijn steekproef zodanig samenstellen dat elke Britse marine-zeeman die aan scheurbuik leed, evenveel kans had om erin terecht te komen. Dat was niet het geval: zijn steekproef bestond uit alle scheurbuiklijders op het marineschip waar hij toevallig werkte. Om een representatieve steekproef te bekomen had hij via een toevalsprocedure een proefgroep moeten selecteren uit de verzameling van alle scheurbuiklijders in de Britse marine.

Als we aannemen dat Lind iets wil aantonen over alle mensen, dan is het probleem nog sterker. Dan had hij een willekeurige steekproef moeten nemen uit de verzameling van alle scheurbuiklijders in 1747. Scheurbuik kwam ook voor op het land (bv. in langdurig belegerde steden), op marineschepen van andere landen en op koopvaardijschepen.

Een tweede voorwaarde betreft het aantal elementen in de proefgroep en in de twee delen ervan. In een experiment moet de experimentele groep minstens 10 elementen bevatten, de controlegroep ook. Dit heeft te maken met de zogenaamde centrale limietstelling uit de statistiek.[4] Het heeft geen zin om confidentie-intervallen te berekenen voor steekproeven met minder dan 10 elementen, omdat de methode in die gevallen onbetrouwbaar is. De zes experimentele groepen van Lind zijn dan ook vanuit hedendaags perspectief veel te klein.

Bovendien is het zo dat experimenten die net deze drempel halen, weinig nut hebben. Wanneer we confidentie-intervallen berekenen volgens de geëigende for-

4. Voor een technische uiteenzetting van deze stelling (Central Limit Theorem in het Engels) kan de online tutorial van de School of Public Health van Boston University geraadpleegd worden (http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Probability/BS704_Probability12.html). Belangrijk voor ons is dat deze stelling zegt dat je bij een 'voldoende grote steekproef' een aantal assumpties kan maken. Het berekenen van confidentie-intervallen berust op die assumpties. Voor dichotome variabelen (zoals genezen / niet genezen) betekent 'voldoende groot' in elk geval minstens 10 (voor andere soorten variabelen zijn er andere drempels).

mules, worden de intervallen kleiner naarmate het aantal elementen in de steekproef groter wordt. Kleine steekproeven (van bv. 10 of 20 elementen) leiden tot grote intervallen, en dus tot een weinig informatieve conclusie. Bijvoorbeeld: als in een kleine steekproef van 20 Vlaamse huisartsen 10 respondenten aangeven dat ze overwerkt zijn (een proportie van 0.50), dan geeft dat met 95% zekerheidsgraad het interval [0.28, 0.72] als conclusie voor de populatie als geheel. Dat is niet heel informatief.

Linds retoriek en zijn latere onderzoek

We vestigen hier de aandacht op enkele methodologisch interessante aspecten van Linds werk.

Ten eerste lijkt het erop dat Lind een retorische truc wilde toepassen, maar dat die mislukt is. In zijn boek over de geschiedenis van scheurbuik en het onderzoek ernaar schrijft Kenneth Carpenter:

> It is interesting that the two subjects in worst condition both received the seawater treatment. Was this chance, or had Lind perhaps been a believer in it and expected that they would give a dramatic response from "worst" to "best"? (1986, p. 53)

Carpenter suggereert dat Lind een vorm van retoriek voor ogen had: hij wilde aantonen dat zeewater werkt als therapie door middel van een spectaculaire verbetering van de gezondheidstoestand van de twee patiënten die er het slechtst aan toe waren. Maar dat lukte niet: bij de zeewater-drinkers was er geen verbetering merkbaar. De patiënten die citroenen en sinaasappelen kregen, waren in de beste conditie na twee weken. Deze evolutie (van "bad but not worst" naar "best") was niet zo dramatisch als de evolutie waarop Lind wellicht hoopte bij zijn zeewaterpatiënten.

Ten tweede heeft Lind later in zijn carrière betere experimenten gedaan die veel minder bekend zijn. In 1758 werd hij arts in Haslar Hospital, een ziekenhuis van de Britse marine in de omgeving van Portsmouth. Daar deed hij een experiment waarin hij aan 130 scheurbeukpatiënten gedurende twee weken een kuur van wort oplegde (Carpenter 1986, p. 65). Dit experiment (en de andere die hij uitvoerde in Haslar) was in twee opzichten beter dan het originele experiment: de experimentele groep was veel groter (en niet onbelangrijk: groot genoeg volgens hedendaagse standaarden) en de patiënten kwamen van verschillende Britse marineschepen. Dit laatste betekent dat de proefgroep representatief kon zijn voor de onderzochte populatie (indien zijn ambitie beperkt was tot Britse zeelui, zie boven). Deze experimenten zijn minder bekend omdat er geen positieve resultaten waren.

Methodologische standaarden voor biomedisch onderzoek: een product van de 20ste eeuw

De huidige methodologische standaarden voor biomedisch onderzoek zijn geleidelijk tot stand gekomen in de 20ste eeuw. We bespreken eerst de opgang van dubbelblinde studies en vervolgens de ontwikkeling van de inferentiële statistiek.

De opgang van dubbelblinde studies

We kunnen ruwweg drie fasen onderscheiden in de opgang van dubbelblinde studies: pionierswerk vanaf 1930, algemene verspreiding vanaf 1950 en tenslotte institutionele verankering vanaf 1970.

Over het pionierswerk schrijft Elaine Shapiro:

> A major milestone was a single-blind study in 1932, when Harry Gold tested the use of xanthine against placebo (lactose) for cardiac pain. He realized that physicians were asking the patients leading questions and prejudicing answers, and thereafter, tried to blind the physicians, as well. In 1935, Hediger and Gold compared two forms of ether in a 'blind test' and legitimized the use of placebos. (2001, p. 11457)

Gold deed dus eerst een blind experiment en besefte vrij snel dat een dubbelblinde onderzoeksopzet nodig is. In de daaropvolgende decennia trokken deze pioniers onderzoekers aan die de nieuwe methodologie wilden toepassen. Ze superviseerden dus heel wat dubbelblinde studies.

De grote doorbraak kwam echter pas later:

> It was not until the 1950s, when antibiotics were discovered and mechanisms of several metabolic diseases understood that modern medicine began, and clinical research could continue to make inroads. Through the 1960s and 1970s, the scientific method superseded authority and tradition. Treatments had to show sensitivity, specificity, and predictability through statistically sound techniques, randomization, the double-blind method, and placebo controls. (2001, pp. 11457-11458)

Het is dus pas na de Tweede Wereldoorlog dat geleidelijk de verwachting groeit dat biomedische onderzoekers dubbelblinde studies doen. Dit is een tweede fase: algemene verspreiding van het idee.

In de laatste fase is er institutionele verankering:

> By 1970 the Food and Drug Administration was empowered to monitor the safety and effectiveness of new drugs based on adequate, well-controlled investigations using appropriate statistical methods [.] (p. 11457)

Het idee van dubbelblind onderzoek wordt ingebouwd in de eisen die bevoegde instituten (o.a. de Food and Drug Administration in de Verenigde Staten) opleggen wat betreft de wetenschappelijke onderbouwing van nieuwe geneesmiddelen. Deze instituten beslissen over het toelaten van nieuwe producten op de markt.

De ontwikkeling van de inferentiële statistiek

Ook de inferentiële statistiek is een product van de 20ste eeuw. Dat kunnen we heel goed aantonen door te bekijken wanneer de pioniers ervan leefden:

> Current mathematical statistics proceed from the works of Karl Pearson and his successors: his son Egon Pearson (1895-1980), the Polish mathematician Jerzy Neyman (1894-1981), the statistician pioneering in agricultural experimentation Ronald Fisher (1890-1962), and finally the engineer and beer brewer William Gosset, alias Student (1876-1937). (Desrosières 2001, p. 15085)

Meer bepaald zijn er belangrijke nieuwe methodes ontwikkeld in ruwweg het eerste derde van de 20ste eeuw. Tore Schweder schrijft over deze periode:

> This period also saw the emergence of fundamentally new ideas and methodologies. Statistical inference, e.g., in the forms of hypothesis testing and confidence interval estimation, was identified as distinct from data description. (2001, p. 15031)

De statistische concepten en methodes die een cruciale functie hebben in de hedendaagse manier van werken in de biomedische wetenschappen, waren dus niet beschikbaar voor wetenschappers die actief waren voor pakweg 1900, waaronder Lind.

Besluit

In de literatuur wordt vaak naar James Linds studie uit 1747 verwezen als het eerste experimentele onderzoek in de geschiedenis van de biomedische wetenschappen (zie bv. Carpenter 1986, p. 52). In dit artikel gebruikten we zijn experiment als een gevalstudie op basis waarvan we de hedendaagse methodologische standaarden voor biomedische experimenten uiteenzetten.

We toonden aan dat de zes experimenten van Lind inderdaad voldeden aan drie belangrijke basiskenmerken van biomedische experimenten: er was een opdeling in experimentele groepen en een controlegroep, er was sprake van experimentele manipulatie en Lind deed een vergelijkende observatie van de effecten van de geteste therapieën.

Vanuit hedendaags perspectief konden we echter ook wijzen op een aantal tekortkomingen van zijn experimenten: er was geen willekeurige indeling in experimentele en controlegroep, en geen neutralisatie van placebo- en onderzoekerseffecten. Om deze tekortkomingen te vermijden moeten hedendaagse biomedische experimenten voldoen aan twee bijkomende standaarden: toevalsverdeling en dubbelblindheid.

Vervolgens hebben we gewezen op twee bijkomende problemen vanuit hedendaags perspectief voor het onderzoek van Lind: het probleem van representativiteit, en het probleem met betrekking tot de grootte van steekproeven. Deze twee problemen hangen samen met het toepassen van inferentiële statistiek om conclusies uit observaties te trekken.

Ten slotte hebben we uitgelegd dat de methodologische standaarden voor het biomedisch onderzoek zoals we ze nu kennen, een product zijn van de 20ste eeuw: pas in deze periode ontwikkelde zich de inferentiële statistiek en vond de eis van dubbelblinde studies langzaamaan ingang als algemene standaard in het biomedisch veld. Die technieken en methodologische inzichten waren er nog niet in de periode waarin Lind onderzoek deed.

De mogelijkheden van Lind waren beperkt, dus de tekortkomingen vanuit hedendaags perspectief mogen zeker niet gezien worden als tekortkomingen van Lind als wetenschapper in de 18de eeuw. Anderzijds is het ook te kort door de bocht om zijn werk te categoriseren als het 'eerste experimenteel onderzoek in de biomedische wetenschappen'. Daarvoor zij de verschillen met wat we nu onder een goed biomedisch experiment verstaan te groot.

Bibliografie

Carpenter Kenneth J. (1986), *The History of Scurvy and Vitamin C.* Cambridge: Cambridge University Press.

Desrosières Alain (2001), 'Statistics, History of', in N. Smelser & P. Baltes (eds.), *International Encyclopedia of the Social & Behavioral Sciences.* Amsterdam & New York: Elsevier, pp. 15080-15085.

Forster Jon (2001), 'Sample surveys: nonprobability sampling', in N. Smelser & P. Baltes (eds.), *International Encyclopedia of the Social & Behavioral Sciences.* Amsterdam & New York: Elsevier, pp. 13467-13470

Rosnow Ralph (2001), 'Experimenter and Subject Artifacts: Methodology', in N. Smelser & P. Baltes (eds.), *International Encyclopedia of the Social & Behavioral Sciences.* Amsterdam & New York: Elsevier, pp. 5120-5124.

Schweder Tore (2001), 'Statistical Methods, History of: Post-1900', in N. Smelser & P. Baltes (eds.), *International Encyclopedia of the Social & Behavioral Sciences.* Amsterdam & New York: Elsevier, pp. 15031-15037.

Shapiro Elaine (2001), 'Placebo Studies (Double-blind Studies)', in N. Smelser & P. Baltes (eds.), *International Encyclopedia of the Social & Behavioral Sciences.* Amsterdam & New York: Elsevier, pp. 11455-11460.

Squire Peverill (1988) 'Why the 1936 Literary Digest Poll Failed' in *The Public Opinion Quarterly* **52**, pp. 125-133.

Stewart C.P. & Guthrie Douglas (eds.) (1953), *Lind's Treatise on Scurvy. A Bicentenary Volume Containing a Reprint of the First Edition of* A Treatise of the Scurvy *by James Lind*. Edinburgh: Edinburgh University Press.

Teira David (2013), 'Blinding and the Non-interference Assumption in Medical and Social Trials', in *Philosophy of the Social Sciences* **43**, pp. 358-372.

Weber Erik, Leuridan Bert & Lefevere Merel (2016), *Wetenschap: wat, hoe en waarom? Systematische inleiding tot de wetenschapsfilosofie*. Antwerpen/Apeldoorn: Garant.