Marcel Levi is internist en bestuursvoorzitter van het AMC

Marcel Levi

Hoe groter de data-verzameling, hoe minder vergelijkbaar de gegevens

VINCENT BOON

Kleren van de keizer

Een van de nieuwe buzz-termen in de geneeskunde is big data. Door koppeling van toenemend beschikbare kolossale databestanden vol gegevens over individuele patiënten, bijvoorbeeld van zorgverleners, zorgverzekeraars en andere instanties, kan meer inzicht worden verkregen over preventie, diagnostiek en behandeling van ziekten. Profeten van de bigdatabeweging beweren vol vuur dat analyse van deze enorme hoeveelheid informatie de geneeskunde en gezondheidszorg gigantisch vooruit zal helpen. Zoals zo vaak bij een hype, spreken veel mensen anderen met enthousiasme na over de mogelijkheden die het linken van grote datasets op kan leveren zonder dat ze precies weten wat er nu werkelijk mogelijk is en wat de beperkingen zijn. Big data lijken een soort kleren van de keizer, door velen bewonderd of vol verwachting tegemoetgezien, maar eigenlijk is er op dit moment ternauwernood een draadje realiteit aan te ontwaren.

Zo ontbreekt tot op heden elk bewijs voor de werkelijke effectiviteit van big data voor gezondheidszorg. In 2009 publiceerden onderzoekers van Google in Nature triomfantelijk dat door trendanalyse van tientallen miljoenen zoekopdrachten op internet naar griepsymptomen en antigriepmiddelen zij veel sneller dan het Center for Disease Control influenza-epidemieën op het spoor konden komen. Na enkele echte influenza-epidemieën bleek echter dat deze bigdata-analyse totaal onnauwkeurig was en werd *Google Flu Trends* roemloos van internet verwijderd.

Waarom is de toekomst van big data in de gezondheidszorg zo onzeker? Onze kennis is voor een belangrijk deel gebaseerd op extrapolatie van waarnemingen bij een beperkte groep patiënten naar de gehele populatie. Het principe van big data is dat hoe groter de groep, hoe groter de kans is dat zelfs hele kleine of onverwachte effecten kunnen worden waargenomen. Echter, tegelijkertijd geldt dat hoe groter de dataverzameling, hoe groter de kans dat gegevens niet goed vergelijkbaar zijn. Bijvoorbeeld omdat verschillende waarnemers data op een andere manier interpreteren of op een andere manier rapporteren. Dat geldt zeker voor de informatie in verschillende databestanden die allemaal voor heel uiteenlopende doelen zijn opgezet, en waarin de gegevens op uiterst variabele wijze zijn verzameld en opgeslagen. Als de input in de databestanden zo ongelijkvormig is, dan kan de output ook nooit veel betekenis hebben, of minder diplomatiek geformuleerd: *garbage in, garbage out*.

Een ander probleem van big data is dat de uitkomsten gebaseerd zijn op correlaties, dikwijls zonder onderliggende hypothese of begrip van de gegevens. Zo kan een mogelijke uitkomst van bigdata-analyse zijn dat gebruikers van middeltjes tegen acne minder vaak een hartinfarct krijgen als niet in de beschouwing wordt betrokken dat op de leeftijd van jeugdpuisten coronaire hartziekten tamelijk zeldzaam zijn. En ten slotte zullen big data in de geneeskunde flink worden gehinderd doordat individuele medische gegevens verspreid zijn over een groot aantal verschillende databases van ziekenhuizen, verzekeraars en anderen en dat – als deze bestanden al te koppelen zijn – velen grote waarde hechten aan de bescherming van privacy, zeker wat medische gegevens betreft. Voor het delen van deze tot op het individu herleidbare data, zelfs tussen zorgverleners, bestaat minimale tolerantie in onze samenleving.

Big data lijken vooralsnog de zoveelste modegril in de gezondheidszorg te zijn. ■