How to produce bullshit about diabetes detection with heartrate sensors on smartwatches

De afgelopen paar dagen is er in de Amerikaanse media, die publiceren over de zegeningen van de "wearables", zoals smartwatches en fitnesstrackers a la Fitbit, fors uitgepakt over het kunnen detecteren van diabetes met smartwatches. Ook Nederlandse media doken erop, zoals onder andere de website Smarthealth en de iCulture-website. De aanleiding is de presentatie van een onderzoeksgroep van de universiteit van San Francisco samen met de mensen achter de medische startup Cardiogram, (klik hier op de link Recent Claim) die de gelijknamige app maakt, op een congres van de Association for the Advancement of Artificial Intelligence (AAAI) in New Orleans van 2 tot 7 februari. Het onderzoek dat nog niet in vakliteratuur gepubliceerd is pretendeert met behulp van gebruik van hartfrequentie-data van smartwatches(Apple), geanalyseerd met een diep neuraal netwerk genaamd DeepHeart, diabetes mellitus met een accuratesse van 85 procent vast te stellen. Eerder publiceerden dezelfde mensen resultaten van onderzoek waarbij met dezelfde methodiek slaapapneu en hypertensie vastgesteld zouden kunnen worden. Dat onderzoek kon op zijn zachtst gezegd "sloppy science" genoemd worden, omdat de beschreven nauwkeurigheid op de keper beschouwd nogal fors tegenvalt en tot forse aantallen fout-positieven en fout-

negatieven leidt. Ik publiceerde daar op 24 november 2017 over.

Achtergrond

Nog steeds zijn wearables in de ogen van techno-optimisten de heilige graal waarmee men vele ziekten denkt te kunnen voorspellen. Het probleem is dat sensors in wearables een aantal zeer specifieke lichaamsfuncties kunnen volgen, zoals hartslag, temperatuur, bewegingen. Wat men met gebruikmaking van kunstmatige intelligentie probeert is om op indirecte wijze uit de meetwaarden van die sensors conclusies te trekken ten aanzien van ziekten waarvan de sensors niet direct de kenmerken meten. Zo is de achterliggende gedachte bij het proberen diabetes te vast te stellen met gebruikmaking van hartslagmetingen dat er bij diabetes vaak sprake is van bepaalde patronen(andere hartfrequentie in rust en andere variabiliteit in het hartritme) die anders zijn dan bij niet diabeten.

Wat deed men?

Bij de melding van een accuratesse van 85 moet men zich altijd afvragen wat daarmee nu precies bedoeld wordt. Bij diagnostische apparatuur dient men ook altijd de sensitiviteit en de specificiteit van de metingen te weten om een enigszins gefundeerd oordeel te kunnen vellen. In dit geval voedde men het neurale netwerk DeepHeart 70 procent van de metingen van in totaal 14000 Apple Watch-gebruikers en ontwikkelde een algoritme dat bij de overige 30 procent uitgetest werd. Daarbij slaagde het algoritme erin om bij de mensen waarvan eerder diabetes was vastgesteld 85 procent te kunnen identificeren. Dat klinkt allemaal heel aardig, maar men miste toch blijkbaar 15 procent van de anderszins bewezen diabeten. Men mat dus uitsluitend de sensibiliteit, het percentage terecht positieve uitslagen onder de zieke personen. Met 15 procent niet gediagnostiseerd en het volledig ontbreken van gegevens over de specificiteit(het percentage terecht

negatieve testuitslagen onder de niet-zieke personen) zegt deze methode om diabetes vast te stellen niets, omdat die in wezen te onnauwkeurig is.

Geen FDA-goedkeuring aangevraagd

Om in de Verenigde Staten eendiagnostische test te mogen verkopen, heeft men goedkeuring van de Food and Drug Administration nodig, zulks om de betrouwbaarheid van die tests te borgen. De oprichter van Cardiogram en drijvende kracht achter alle publiciteit, Brendan Ballinger(geen arts, maar ICT-er) heeft echter geenszins de bedoeling om voor zijn methodiek een FDA-goedkeuring aan te vragen. Hij weet ook dat de Cardiogram-app nooit als een standalone diagnostische test gebruikt kan worden, maar als een soort vriendelijk advies, en zegt:

"To stay on the right side of the US Food and Drug Administration, the app can't function as a standalone diagnostic, more like some friendly advice. But the kind of advice an insurer might cover if they thought it would get people into treatment earlier and save healthcare costs ."

Wonderlijk

Hoewel de Cardiogram-app gratis is, ziet Ballinger wel een businessmodel in zijn black-box-intelligentie, dus nadere analyse tegen betaling. Maar als er dan een afwijkende uitslag komt zoals diabetes is zijn advies:

"Right now, we'd always use an existing FDA-cleared (or CLIA-waved) test to confirm the diagnosis."

Dus, even diep ademhalen, hij adviseert zijn uitslag te laten bevestigen door een test te doen met het gebruikelijk onderzoek van diabetes in bloed met gevalideerde apparatuur en zeer betrouwbare media. Dan is het toch veel simpeler voor een persoon die zich afvraagt of hij/zij diabetes heeft, om de te betalen service van Cardiogram over te slaan en meteen een

gevalideerde bloedtest te laten doen bij de eigen arts.

Ongerustheid

In feite is men met dit soort zogenaamde doorbraken alleen maar bezig de ongerustheid bij mensen over hun gezondheid aan te wakkeren en te exploiteren. Het lijkt allemaal heel mooi en voedt de gedachte dat met hightech-toepassingen van alles gemeten kan worden, maar niets is minder waar. Dit soort propaganda van hightech wordt klakkeloos in de media overgenomen in de vorm van copy-paste-journalistiek zonder het afvragen wat de basale vereisten zijn van betrouwbaar te achten diagnostisch onderzoek. Het nu gepubliceerde verhaal van Cardiogram kan ik slechts als “bullshit” karakteriseren.

W.J. Jongejan

Sloppy science met Apple Watch samen met Cardiogram-app

Op meerdere websites was rond 14 november 2017 te lezen dat hoge bloeddruk en het slaap-apneu-syndroom verrassend goed vast te stellen zou zijn met de app Cardiogram op een Apple Watch. De publicatie waarop dit verhaal gebaseerd is, werd als abstract op die datum gepubliceerd in het magazine Circulation van de American Heart Association. Het artikel oogt indrukwekkend

met als titel: “Cardiovascular Risk Stratification Using Off-the-Shelf Wearables and a Multi-Task Deep Learning Algorithm “. Geschermd wordt met het gebruik van een “deep neural network” waarbij door middel van algoritmen computerprogramma’s voorspellingen kunnen doen over medische diagnose op basis van data afkomstig van “wearables”. Dat zijn draagbare sensors van lichaamsfuncties, zoals die zitten in diverse smart-watches en fitness-trackers(FitBit bijv.). Bij nadere beschouwing van de vermelde gegevens blijkt de accuratesse van de gebruikte methode nogal tegen te vallen en zullen de resultaten eerder ongerustheid bij smartwatch-dragers vergroten en leiden tot medische overconsumptie.

Hypertensie en slaap-apneu

Het doel van het onderzoek aan de universiteit van San Francisco was om te evalueren of een “deep neural network” cardiovasculaire risicofactoren kon voorspellen met “wearables”. Men koos voor de Apple Watch en de app Cardiogram. Op de achterzijde van dit smartwatch zitten twee lichtbronnen met bijbehorende sensoren. Eén voor gewoon licht en één voor infraroodlicht. Op basis van photoplethysmografie wordt hiermee dan het hartritme gevolg. In het onderzoek gebruikte men ook de versnellingsmeter die in elke smartphone of smartwatch zit om de lichaamsactiviteit te meten. De reden dat men hypertensie, slaap-apneu koos was dat veranderingen van het hartritme en lichaamsbeweging in de tijd in het verleden daarmee al geassocieerd waren.

Uitkomsten

Laten we eens zien wat men vertelt:

> *6,115 active users of the Cardiogram app for Apple Watch…Mean age was 42.3 ± 12.1, 69% male. 2,230 (36.5%) of participants had hypertension, 1,016 (16.6%) had sleep apnea, and 462 (7.6%) had diabetes. In the validation set, the DNN outperformed a baseline logistic regression model*

> *incorporating age, sex, and beta blocker use, predicting prevalent hypertension with a c-statistic of 0.819 (95% CI 0.76-0.88; with an optimal operating point yielding **84.8% sensitivity and 63.6% specificity)** vs a baseline c-statistic of 0.682 (95% CI 0.60-0.76), and prevalent sleep apnea with a c-statistic of 0.902 (95% CI 0.85-0.95; with an optimal operating point yielding **90.4% sensitivity and 59.8% specificity**) vs a baseline c-statistic of 0.459 (95% CI 0.39-0.53). Results were not statistically significant for diabetes.*

Zowel de onderzoekers zelf als meerdere publicaties in de pers(A, B, C, D, E) over dit onderzoek geven aan dat er sprake is van een behoorlijke accuratesse, maar dat valt nogal tegen.

Commentaar

In de eerste plaats blijkt er geen sprake van een normale representatie van de bevolking. Het gaat om mensen die een smartwatch al bezitten, waardoor een selectie op sociaaleconomische status erg waarschijnlijk is. Mannen zijn over-vertegenwoordigd. Het percentage mensen met slaapapneu is veel hoger dan in de normale bevolking. In Nederland schat men het voorkomen van het obstructieve slaapapneu syndroom(OSAS) bij mannen tussen de 0,45 en 4 %, bij vrouwen lager. In een kleine vijver met veel bekende vissen is het makkelijker een te voren gewenste vis te vangen dan in een grote vijver met een lager percentage vissen van dezelfde soort. Niets staat vermeldt over subgroepen van de onderzoekspopulatie(blank, latijn-Amerikaans, afro-Amerikaans etc.. Dat is onder andere van belang voor de nauwkeurigheid van de Apple Watch. Bij personen met een donkere huidskleur is de accuraatheid van de hartritmemeting met photoplethysmografie beduidend lager dan bij een lichtere huidskleur. De getallen over de sensitiviteit en de specificiteit van de test laten het duidelijkst het falen van de methodiek zien. Schrijvers van artikelen in de lekenpers spreken vaak over accuratesse van een test als ze

het over de sensitiviteit hebben van die test, maar alleen de sensitiviteit en specificiteit tezamen zeggen daar iets over.

Sensitiviteit en specificiteit

Met sensitiviteit wordt de kans op een positieve uitslag bij aanwezigheid van de ziekte bedoeld, met specificiteit de kans dat de test negatief is bij afwezigheid van de ziekte. Hoe hoger de sensitiviteit van een test, hoe groter de kans dat iemand die daadwerkelijk de ziekte heeft, een positieve testuitslag. De vermelde percentages voor de sensitiviteit van de test voor hypertensie en slaapapneu zijn respectievelijk 84,8 en 90,4. Dat lijkt heel mooi maar zegt niets zonder de samenhang met de specificiteit. Die is voor 63,6 en 59,8. Dat laatste valt heel erg tegen. Het betekent dat bij de onderzochte personen rond de veertig procent ten onrechte te horen krijgt dat ze de ziekte hebben zonder dat zulks het geval is. Idealiter moeten diagnostische tests een sensitiviteit en specificiteit van 100 % hebben. Dat wordt vrijwel nooit behaald maar een specificiteit van rond de 60% is beslist veel te laag. Er vindt dus evidente over-diagnostisering plaats.

Heilige graal

De over-diagnostisering leidt alleen maar tot meer ongerustheid en meer medische consumptie bijv. om met betere onderzoeksmethoden, die niet met een omweg iets meten, vast te stellen of de testuitslag met de smartwatch wel klopt. Fabrikanten van wearables en adepten ervan zijn op zoek naar hun “heilige graal” om met relatief simpele sensoren met gebruik van computerbeoordeling iets over ziekten te zeggen waar de sensor niet rechtstreeks iets van meet. In een tijd waarin de verkoop van wearables over haar hoogtepunt heen is, lijkt dat voor de fabrikanten aantrekkelijk maar het is een doodlopende weg. Dat kan alleen maar leiden tot veel fout-positieven, dus veel ongeruste mensen.

W.J. Jongejan

Met dank aan Johan Goris voor discussie over dit onderwerp.