CODE ZWART

crisismanagement met slimme modellen

Er komen twee doodzieke patiënten binnen met COVID-19 op jouw intensive care. Alle overplaatsingslocaties zitten vol, dus je moet een keuze maken. Wie ga je helpen? Hoe maak je een keuze tussen mensen met een ziekte die je niet begrijpt? Voor veel artsen wereldwijd is dit dilemma werkelijkheid geweest, in Nederland waren we heel dichtbij. Om dit dilemma het hoofd te bieden heeft een multidisciplinair team van clinici en datawetenschappers de handen in elkaar geslagen om een data-gedreven model te ontwikkelen, zodat deze kritieke keuze niet zonder informatie gemaakt hoeft te worden. We nemen je mee door een beschrijving van keuzes en dilemma's om van idee naar resultaat te komen in onzekere tijden.

Maarten Ottenhoff

Nu het grootste deel van de bevolking is gevaccineerd lijken de strenge lockdowns van afgelopen jaar verleden tijd. Alle maatregelen waren essentieel om de instroom van COVID-19-patiënten in het ziekenhuis te beperken, zodat iedereen de zorg kan krijgen die zij nodig hebben. Het punt waarop zorg niet voor iedereen gegarandeerd kon worden, ook wel code zwart genoemd, was meerdere keren dichtbij. Met overplaatsingen naar onze Duitse buren kon deze situatie op het nippertje voorkomen worden.

Tijdens de eerste infectiegolf werden al snel ad-hoc protocollen gemaakt voor code zwart. In een ideale wereld zou je een afweging maken aan de hand van de risicofactoren van de ziekte en de huidige gezondheid van de patiënt. Echter, hiervoor moet er wel een duidelijke wetenschappelijke basis zijn van de risicofactoren. Voor het nieuwe coronavirus was die basis er nog lang niet, dus al snel werd er gekeken naar bredere risicofactoren, zoals leeftijd en bepaalde co-morbiditeiten zoals obesitas, diabetes of longaandoeningen. Toen de besmettingscijfers een paar maanden later waren gedaald en de collectieve paniek was gaan liggen, laaide de discussie op of het grondwettelijk was toegestaan om te selecteren op leeftijd.

Kortom, er was een grote vraag naar medische informatie over het coronavirus, maar voor zorgvuldig onderzoek was er tijd nodig. Om een betere keuze te kunnen maken tijdens code zwart moest er acuut informatie komen. Deze vraag naar informatie was het startpunt van een multidisciplinair onderzoek door een team van artsen en data-wetenschappers van het MUMC+ Maastricht en AMC Amsterdam. Het idee was simpel: verzamel de beschikbare data van opgenomen COVID-19-patiënten in Nederlandse ziekenhuizen en ontwikkel een model dat een inschatting maakt van de kans op overlijden. Bij een succesvol model heeft de arts tijdens code zwart extra informatie over de risico's per patiënt, gebaseerd op medische data. In dit artikel neem ik u graag mee in de verschillende ontwikkelingsstappen van dit project. Voor een volledige inhoudelijke beschrijving verwijs ik u graag door naar het inmiddels gepubliceerde artikel in het wetenschappelijke tijdschrift BMJ open (Ottenhoff et al. 2021).

Stap 1: Dataverzameling

Een week na het begin van de eerste lockdown in maart 2020 ging het project van start. Er waren zoveel mogelijk data nodig, dus werden alle mogelijke kanalen ingezet, van persoonlijke netwerken tot een oproep op televisieprogramma *Op1*. Uiteindelijk waren tien ziekenhuizen bereid om data aan te leveren. Ook bedrijven droegen graag op vrijwillige basis bij, zo konden we bijvoorbeeld de databasestructuur van CASTOR (data-management-platform voor klinische trials) kosteloos gebruiken. Ondanks dat projecten vaak gestroomlijnd lijken te verlopen op papier, zo anders was het in de praktijk: het online invoerformulier werd meerdere malen aangepast toen de data invoer al begonnen was; vanuit statistisch oogpunt een slecht idee. Ook werd er en passant hard gewerkt om alles juridisch in orde te krijgen. Ethische commissies moesten snel een voorstel krijgen om goed te keuren, ziekenhuizen moesten besluiten of er wel of geen informed consent nodig was. Desondanks stond er binnen een maand een database met data van meer dan tweeduizend patiënten van tien Nederlandse ziekenhuizen.

Stap 2: Data schoonmaken

Een goede datakwaliteit is cruciaal, dus werd de eerste maand na dataverzameling besteed aan het schoonmaken van de data. Vaak zijn dit vrij simpele bewerkingen zoals alle binaire variabelen dezelfde kant op laten wijzen (nee = 0, ja =1), alle waarden naar dezelfde eenheid omrekenen of tijdstippen omrekenen naar een relatieve tijd. Ook variabelen waarbij de vraagstelling tijdens data-in-

voer veranderde moesten worden aangepast . Niet alleen moet alles numeriek kloppen, maar ook vanuit medisch perspectief moet het logisch en verklaarbaar zijn. Staan er geen vreemde waarden in? Zijn gerelateerde medische variabelen op de juiste manier gecombineerd?

Hierdoor zijn er potentieel invloedrijke fouten hersteld, wat het belang van een goede samenwerking tussen de datawetenschapper en clinici onderstreept.

Stap 3: Modelontwikkeling

Een machine-learning algoritme genereert een model aan de hand van voorbeelden. Dat wil zeggen dat je het algoritme een set aan datapunten geeft met een bijbehorende uitkomst: overleden of niet. Het model leert dan welke onderliggende patronen bij welke uitkomst horen. Vervolgens kun je een vergelijkbare set aan datapunten invoeren in het model, die vervolgens een voorspelling teruggeeft over de uitkomst. Essentieel is een duidelijke definitie van de te voorspellen uitkomst. Wat valt er precies onder het overlijdensrisico en over welke tijdspanne gaat het? Dat laatste is een belangrijke afweging: een lange tijdspanne geeft genoeg tijd om het hele ziektebeloop te vangen, maar het kost meer tijd en de onzekerheid wordt groter. Er kunnen immers meer onverwachte gebeurtenissen voorkomen in zes weken dan in een week. Een te korte tijdspanne zorgt weer voor een zekere voorspelling, maar de kans is groter dat de patiënt daarna nog overlijdt. Het gevolg is dat een patiënt ten onrechte wordt gelabeld als 'overleefd'. Uiteindelijk is er gekozen voor de balans van 3 weken, wat resulteerde in de volgende definitie: De kans op overlijden binnen 21 dagen na ziekenhuisopname. Onder de kwalificatie 'overlijden' vallen ook patiënten die worden ontslagen naar palliatieve zorg. Van de initiële 2527 patiënten bleven er nu 2273 over.

Om meer informatie te krijgen over de risicofactoren zijn de meer dan 400 variabelen teruggebracht aan de hand van selectiecriteria opgezet door een team van clinici. Niet alleen de potentieel voorspellende waarde werd meegewogen, maar ook de uitvoerbaarheid werd bediscussieerd. Voor sommige voorspellende variabelen was er bijvoorbeeld een CT-scan nodig. Theoretisch waardevol, maar vanuit praktisch oogpunt is een CT-scan tijdrovend, duur en er is weinig capaciteit. Het heeft daarom weinig nut om deze variabele mee te nemen.

Uiteindelijk bleven er 80 variabelen over, die vervolgens weer zijn weer opgesplitst in drie categorieën:

1. Premorbiditeiten; leeftijd, geslacht, beroep en medische geschiedenis;
2. Klinische karakteristieken; bestaande uit simpele medische testen zoals hartslag, bloeddruk en ademhalingsritme;
3. Laboratorium en radiologie variabelen: bloedtesten en scans.

Ook de combinaties van categorie 1 en 2 en alle drie de categorieën werden meegenomen. Als laatste maken we ook gebruik van een data-gedreven test om de tien 'meest voorspellende' variabelen mee te nemen. Hierbij selecteerde we de tien variabelen met de hoogste F-waarde uit een Analysis of Variance (ANOVA). In totaal zijn er dus zes verschillende groepen van variabelen meegenomen in het uiteindelijke model.

Voor het modelleren hebben we gekozen voor twee fundamenteel verschillende soorten modellen: een lineaire logistische regressie (LR), en een non-lineair *tree-based gradient boosting* algoritme (TBGB).

Een LR is eenvoudig model dat goed te interpreteren is. In essentie het een binaire versie van een lineaire regressie. Bij een lineaire regressie wordt een lijn gezocht met de kleinste afstand tot alle datapunten. Om uitkomst te transformeren naar een waarde tussen 0 en 1 wordt deze in een logistische functie gegoten: $y = 1 / (1+\exp(-ax+b))$ waarbij ax+b de lijn is van de lineaire regressie. Vervolgens wordt er een grens gesteld om een voorspelling te maken: alle waardes boven die grens is overlijden, alles daaronder overleven.

Een TBGB-model is gebaseerd op een keuzeboom. Het model genereert een serie van keuzes met een uiteindelijke voorspelling: 'Als variabel groter dan 1, dan A, anders B'. Het TBGB-model controleert vervolgens welke voorspellingen hij fout heeft, en geeft deze datapunten een bepaalde weging aan de hand van de grootte van de fout. Hierdoor krijgen de moeilijke datapunten meer aandacht in het leerproces, waardoor de fout kleiner wordt. Dit proces wordt herhaald totdat de foutmarge geminimaliseerd is, en de optimale voorspelling gemaakt kan worden.

Idealiter gebruikt men twee onafhankelijke datasets om het model te trainen en te testen. Hiermee voorkomt je dat de modellen 'overfitten' op de data, dat wil zeggen dat de modellen irrelevante of dataset specifieke patronen vinden die niet generaliseerbaar zijn. Om dat te voorkomen hebben we gebruik gemaakt van een techniek genaamd cross-validatie. Daarbij splitsen we de data op per centrum. Vervolgens zijn negen ziekenhuizen gebruikt als 'training set' en een ziekenhuis als 'test set'. Dit herhaalde we dit tien keer, waarbij elk ziekenhuis een keer als test set gebruikt wordt. Uiteindelijk wordt de gemiddelde score gebruikt als eindscore.

Stap 4: Model validatie

Samenvattend hebben we twee modellen, LR en TBGB, gemodelleerd op 6 verschillende datasets, gebruikmakende van *leave-one-hospital-out cross-validation*. De uitkomst meten we onder meer met de *area under the receiver operator curve* (AUC).

De receiver operator curve is een plot waarbij de sensitiviteit (aantal correct positieve voorspellingen in verhouding met alle daadwerkelijk positieve gevallen plus alle fout negatieve voorspellingen) en specificiteit (aantal correct negatieve voorspellingen in verhouding met alle daadwerkelijke negatieve gevallen plus het aantal fout positieven) tegen elkaar uitgezet worden. Vervolgens wordt de beslisgrens (bijv. een geval is positief bij een waarde boven de 0,5, anders negatief) met kleine stapjes van 0 naar 1 verplaatst en wordt telkens de sensitiviteit en specificiteit berekend. Als je deze collectie aan punten tegen elkaar plot en de oppervlakte onder die lijn berekent, dan krijg je de AUC, een waarde tussen 0 en 1. Bij een AUC van 1 is de voorspelling perfect, bij een AUC van 0 is de voorspelling perfect fout, ofwel precies andersom. Een AUC van 0,5 betekent dat de voorspelling op kansniveau is. Je kunt dan net zo goed het model achterwege laten en gokken wat het juiste antwoord is.

Uiteindelijk haalde LR en TBGB een score van 0,82 AUC, een goede score. Ter vergelijking, als een simpel model wordt gehanteerd zoals 'iedereen ouder dan 70 of 80 zal overlijden', dan krijgt je een AUC van respectievelijk 0,69 of 0,61. Het werkt dus al beter dan simpele op leeftijd gebaseerde regels die in andere landen in de praktijk zijn toegepast.

Om de voorspellende variabelen zo inzichtelijk mogelijk te maken, hebben we gebruik gemaakt van SHAP-waarden. Deze methode berekent de gemiddelde relatieve bijdrage van een variabel ten opzichte van de voorspelling. Hierdoor krijgt men inzicht in a. de relatieve bijdrage ten opzichte van van de andere variabelen en b. de richting van de variabele. Met andere woorden: zorgen hogere waardes vaker voor een positieve uitkomst? Leeftijd leverde met afstand de belangrijkste bijdrage. Niet onverwacht zorgde een hogere leeftijd voor een hoger risico op overlijden. Verder vonden we ook dat het aantal verschillende medicaties, het stikstofgehalte in het bloedplasma, lactaatwaarden en het zuurstofgehalte in het bloed belangrijk waren voor een goede voorspelling. Een belangrijke kanttekening van contributie-analyses voor machine-learning modellen is dat je er geen conclusies uit kan trekken over het daadwerkelijk risico. Correlatie is immers geen causatie. Deze variabelen kunnen echter wel een goed startpunt vormen voor vervolgonderzoek.

In de zomer van 2020 barste er een flinke discussie los in de maatschappij over of leeftijd als selectiecriterium gebruikt mag worden indien er een gedwongen keuze gemaakt moet worden tussen patiënten. Veel medici zeggen van wel; leeftijd is een medisch relevant risico voor veel ziektes en aandoeningen. Denk bijvoorbeeld aan Parkinson of dementie. De politiek is het hier niet mee eens, omdat de grondwet voorschrijft dat er niet gediscrimineerd mag worden op leeftijd, geslacht, huidskleur of religie. Vanwege deze discussie hebben we het best presenterende model nog eens getraind, maar dan zonder leeftijd. Opvallend was dat de AUC slechts met 0,04 afnam naar 0,78, ondanks dat leeftijd met afstand het meeste bijdroeg aan de voorspelling. Dit is een indicatie dat de voorspellende waarde indirect ook aanwezig is in andere variabelen. Uit onze analyse blijkt dus dat het gebruik van leeftijd lijdt tot een betere voorspelling, maar dat het vanuit medisch perspectief te verantwoorden is om leeftijd niet meer te nemen.

Uiteindelijk is code zwart nooit bereikt. Er hoefde geen keuzes gemaakt te worden tussen twee patiënten die acuut hulp nodig hebben. Wel zijn we meerdere keren dichtbij geweest, en mocht dat nog voorkomen, dan hebben we in ieder geval een extra hulpmiddel. In de tussentijd rest het nog om het model in de praktijk te valideren. Bijvoorbeeld door schaduw te draaien en de uitkomst te vergelijken met de keuze die een arts zou maken. Ook is het belangrijk om te kijken of het model genoeg toevoegt; wellicht voorspelt het model uitstekend, maar is de keuze op basis van ervaring van de arts simpelweg beter. De beste uitkomst is in ieder geval als we altijd in het ongewis blijven, dat het model verstoft op de plank, zonder ooit gebruikt te worden.

Literatuur

Ottenhoff, M. C., Ramos, L. A., Potters, W. et al. (2021). Predicting mortality of individual patients with COVID-19: a multicentre Dutch cohort. *BMJ Open 11.* (e047347). doi: 10.1136/bmjopen-2020-047347

Maarten Ottenhoff is promovendus aan de Universiteit van Maastricht in de vakgroep Neurochirurgie. In zijn onderzoek binnen de neurotechnologie zoekt hij naar nieuwe signaalanalysemethoden voor de ontwikkeling van neuromodulators en brain-computerinterfaces. Door de kritieke COVID-toestand in het ziekenhuis heeft hij tijdelijk zijn aandacht gericht op acute problemen in de COVID-zorg.
E-mail: m.ottenhoff@maastrichtuniversity.nl