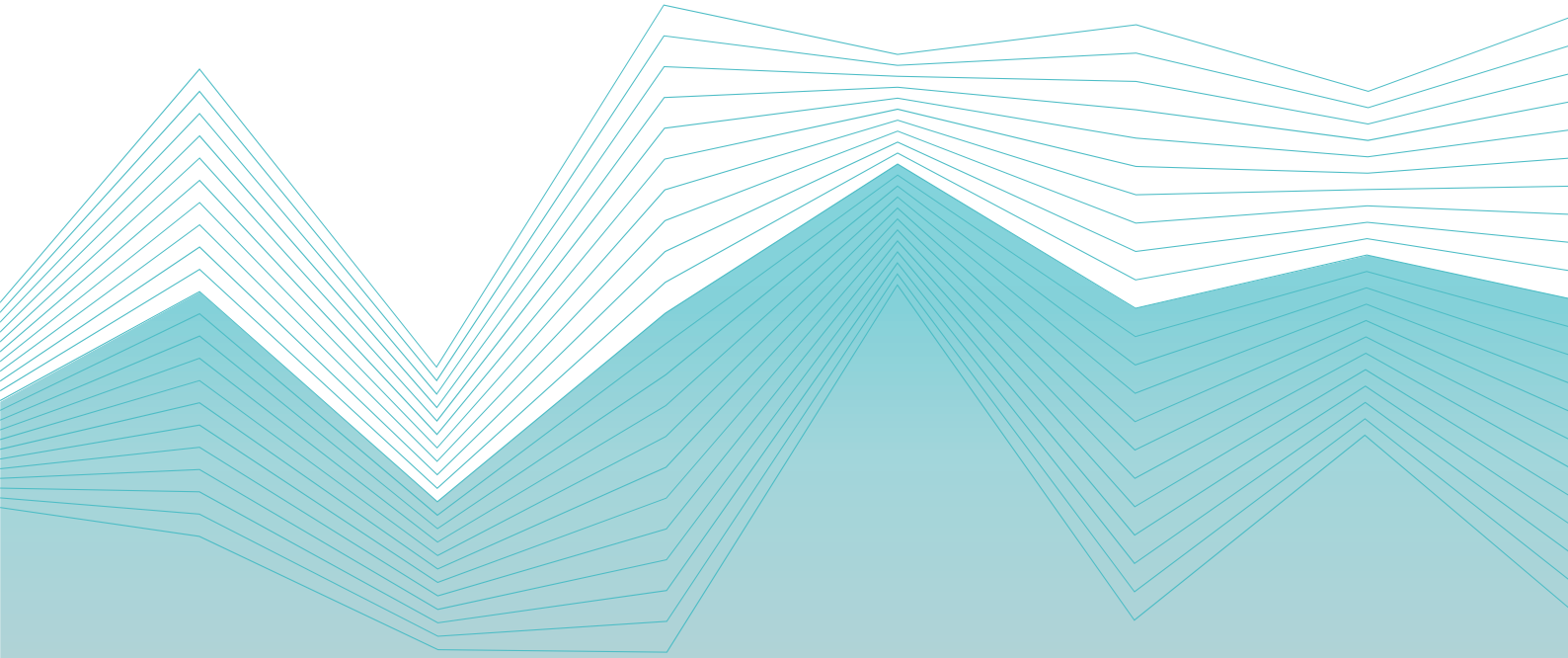
# Vroegopsporing hoogrisico-patiënten voor hart- en vaatziekten

## Een studie naar de potentie van machine learning op basis van het elektronisch patiëntendossier van de huisarts



Joke Korevaar
Claire Aussems
Marianne Heins
Rune Poortvliet
Mark Nielen
René Eijkemans
Derek de Beurs
Monika Hollander

Januari 2020

NIVEL
Kennis voor betere zorg

Het Nivel levert kennis om de gezondheidszorg in Nederland beter te maken. Dat doen we met hoogwaardig, betrouwbaar en onafhankelijk wetenschappelijk onderzoek naar thema's met een groot maatschappelijk belang. 'Kennis voor betere zorg' is onze missie. Met onze kennis dragen we bij aan het continu verbeteren en vernieuwen van de gezondheidszorg. We vinden het belangrijk dat mensen in staat zijn om deel te nemen aan de samenleving. Ons onderzoek draait uiteindelijk om de vraag hoe we de zorg voor de patiënt kunnen verbeteren. Alle onderzoeken publiceert het Nivel openbaar, dat is statutair vastgelegd.

Januari 2020

ISBN 978-94-6122-606-8

030 272 97 00
nivel@nivel.nl
www.nivel.nl

# Voorwoord

Voor u ligt het rapport ‘Vroegopsporing hoogrisico-patiënten voor hart- en vaatziekten. Een studie naar de potentie van machine learning op basis van het elektronisch patiëntendossier van de huisarts’. Hierin beschrijven we de mogelijkheden van machine learning voor ondersteuning bij het vroeg opsporen van hoogrisico-patiënten voor hart- en vaatziekten en geven we een overzicht van mogelijke vervolgstappen. Het Nivel heeft dit onderzoek uitgevoerd met subsidie van de Hartstichting.

De auteurs
Utrecht, januari 2020

# Inhoud

# Samenvatting

Hoe eerder patiënten met een hoog risico op hart- en vaatziekten (HVZ) herkend worden, hoe beter. Door vroege opsporing kan er tijdig begonnen worden met leefstijladviezen en/of medicamenteuze behandeling, waardoor het ontstaan van HVZ uitgesteld of zelfs voorkómen kan worden. Echter, het identificeren van patiënten met een (potentieel) hoog risico is een complex proces; veel patiënten moeten gescreend worden om enkele cases te vinden. Bovendien is het moeilijk de doelgroep te bereiken, omdat het merendeel zich, terecht, niet aangesproken voelt. Het gevolg hiervan is dat huisartsen vaker personen screenen die zich zorgen maken, maar geen hoog risico hebben, dan personen met een daadwerkelijk hoog risico. Als opsporing efficiënter kan door, op basis van routinematig verzamelde gegevens in de huisartsenpraktijk, alleen de 'echte' hoogrisico-patiënten uit te nodigen voor consult en screening, kan dit preventie van HVZ ten goede komen.

Machinelearning-technieken zijn in opkomst doordat deze goed om kunnen gaan met grote hoeveelheden routinematig verzamelde data. Anders dan traditionele technieken nemen machinelearning-technieken de complexe niet-lineaire verbanden tussen voorspellers mee in het optimaliseren van de voorspelling. Ook kan machine learning beter omgaan met grote hoeveelheden voorspellers. In deze studie onderzochten we de potentie van machine learning om hoogrisico-patiënten voor (risicofactoren voor) HVZ op te sporen. We maakten daarbij gebruik van routinematig verzamelde gegevens uit huisartsenpraktijken die deelnemen aan Nivel Zorgregistraties, waarbij informatie over zorgconsulten en medicatie uit het Huisarts Informatie Systeem (HIS) werd gebruikt. Waar mogelijk verrijkten we de data met meetwaarden (BMI, roken en bloeddruk). We maakten gebruik van twee veel veelgebruikte machinelearning-algoritmen: LASSO regressie en random forest.

De resultaten lieten zien dat de uitkomsten van de machinelearning-technieken op de beschikbare gegevens uit het HIS niet goed genoeg waren om de huisarts te ondersteunen bij de voorspelling van patiënten met een hoog risico op HVZ. De algoritmes selecteerden weliswaar een kleine groep patiënten, waarvan een groter deel daadwerkelijk een hoger risico op HVZ had dan in de algemene huisartsenpopulatie, maar de algoritmes misten ook veel cases, zodat de algehele voorspellende waarde van de modellen tegenvalt. Aangezien de huisarts nooit alle patiënten met een mogelijk risico op HVZ kan uitnodigen voor consult, valt deze uitkomst als winst te zien, maar voor een daadwerkelijke toepassing moet het aantal gemiste cases verkleind worden. Wanneer routinematig verzamelde gegevens uit de huisartsenpraktijk verrijkt zouden worden met gegevens over bijvoorbeeld de persoonlijke gezondheid en omgeving van de patiënt, of wanneer er gebruik wordt gemaakt van nieuwe generatie machinelearning-algoritmes (deep learning), kan de voorspelling wellicht verbeterd worden. Ook moet er worden nagedacht over de implementatie van machinelearning-algoritmes in het HIS van de huisarts. Tot die tijd hebben machinelearning-algoritmes, voor de voorspelling van (risicofactoren van) HVZ in de huisartsenpraktijk, op basis van beschikbare gegevens uit het HIS geen meerwaarde.

# 1 Inleiding

Hoe eerder patiënten met een hoog risico op hart- en vaatziekten (HVZ) herkend worden, hoe beter. Door vroege opsporing kan er tijdig begonnen worden met leefstijladviezen en/of medicamenteuze behandeling, waardoor het ontstaan van HVZ uitgesteld of zelfs voorkómen kan worden. Echter, het identificeren van patiënten met een (potentieel) hoog risico is een complex proces. De huidige methode, op basis van het PreventieConsult module cardiometabool, waarbij patiënten tussen de 45 en 70 jaar oud zonder bekende HVZ een uitnodiging krijgen voor het invullen van een (online) risicoschatting en zelf, op basis van het advies dat daaruit volgt, een afspraak met de huisartsenpraktijk moeten maken, is weinig efficiënt. Nivel onderzoek laat zien dat twee derde van de patiënten geen vragenlijst invult en dat twee derde van de hoogrisico-patiënten, die wel de vragenlijst heeft ingevuld, het advies om een afspraak met de huisartsenpraktijk te maken niet heeft opgevolgd [1]. Ook een recente Nederlandse studie laat een lage respons zien op de uitnodiging van de huisartsenpraktijk [2]. Resultaten van een Europese studie (SPIMEU) lieten zien dat in Nederland slechts 30% van de patiënten de vragenlijst teruggestuurde en dat bijna niemand van de hoogrisico-patiënten op consult kwam. Vergelijkbare percentages werden gevonden voor patiënten uit Zweden en Denemarken (data nog niet gepubliceerd). Dit terwijl relatief veel patiënten met een hoog risico baat zouden hebben bij eerdere opsporing en preventieve maatregelen. Nederlands onderzoek uit 2011 liet zien dat er in een gemiddelde huisartspraktijk rond de 100 patiënten tussen de 45-70 jaar een niet herkend hoog risico op HVZ hebben [3].
Een betere methode om hoogrisico-patiënten te identificeren dan het PreventieConsult cardiometabool, zou dan ook zeer welkom zijn. Zeker als het ook nog een methode betreft die weinig extra inspanning van de huisartsenpraktijk behoeft. Dit roept de vraag op in hoeverre het elektronische patiëntendossier (EPD) in de huisartsenpraktijk bruikbaar is om huisartsen meer te ondersteunen bij het identificeren van de groep patiënten met een hoog risico op HVZ.

> Wat is machine learning?
>
> Machine learning is een set statistische technieken die leert van data zonder daar expliciet voor geprogrammeerd te zijn. Dat is handig wanneer je met grote datasets met veel variabelen te maken hebt, omdat je als onderzoeker niet van te voren weet hoe alle variabelen met elkaar samenhangen. Traditionele regressie technieken hebben moeite wanneer grote sets variabelen op een complexe niet-lineaire manier met elkaar samenhangen. Machine learning leert zelf welke verbanden er zijn tussen de variabelen, op basis van de beschikbare data.

In tegenstelling tot de meer traditionele statistische modellen, zijn machinelearning-modellen in staat om met grote hoeveelheden ongestructureerde data om te gaan. Machinelearning-technieken kunnen niet-lineaire verbanden vinden in de data zonder dat de onderzoeker van te voren deze verbanden expliciet hoeft te maken. Dat maakt machine learning tot een flexibele set technieken om tot een zo goed mogelijke voorspelling te komen. Een recente studie uit Engeland onderzocht de inzet van vier verschillende machinelearning-algoritmes waaronder LASSO regressie en Random forest om patiënten met een verhoogd risico op HVZ op te sporen [4]. Vergeleken met de gouden

standaard (in dit geval de richtlijn van de Amerikaanse cardiology vereniging) was het algoritme in staat om 355 (7.6%) hoogrisico-patiënten extra te herkennen. Dergelijke technieken worden momenteel ook al ingezet om op basis van het EPD van de huisarts, patiënten met een verhoogd suïcidaal risico op te sporen [5] en om de kans op huiselijk geweld te voorspellen [6]. Naast een betere voorspelling is het grote voordeel dat machine learning de screening kan automatiseren. Zonder tussenkomst van de huisarts of patiënt kan het algoritme op basis van gegevens uit het elektronisch patiëntendossier uitrekenen of een patiënt een hoog risico heeft. Dit kan vervolgens eenvoudig worden doorgegeven aan de huisarts, waarna deze weet dat dat patiënt een hoog risico heeft. Dit kan bijvoorbeeld in een volgend consult ter sprake gebracht worden of de huisarts kan de patiënt proactief benaderen. In deze studie maken we gebruik van de machinelearning-technieken LASSO regressie en Random forest om op basis van eerder zorggebruik, medicatie voorschriften en beschikbare meetwaarden een verhoogd risico op HVZ te voorspellen. Daarmee onderzoeken we de potentie van machine learning en zetten we een eerste stap naar geautomatiseerde screening voor HVZ in de huisartsenpraktijk.

# 2 Methoden

Op basis van gegevens van Nivel Zorgregistraties Eerste Lijn over de periode 2009-2017 en met behulp van LASSO regressie en random forest zijn predictiemodellen ontwikkeld om huisartsen te helpen voorspellen of een patiënt grote kans heeft om een (risicofactor voor) HVZ te ontwikkelen.

## 2.1 Nivel Zorgregistraties Eerste Lijn

Nivel Zorgregistraties Eerste Lijn verzamelt routinematig gegevens uit het EPD van ongeveer 500 huisartsenpraktijken en beschikt daarmee over informatie over gezondheidsproblemen, contacten, prescripties, uitslagen van diagnostische tests van zo'n 1,7 miljoen Nederlanders [7]. Diagnoses (zowel symptomen als aandoeningen) worden door huisartsen gecodeerd volgens de International Classification of Primary Care (ICPC), versie 1 [8]. Voor voorschriften van geneesmiddelen wordt de Anatomisch Therapeutisch Chemisch (ATC) classificatie gebruikt [9]. Voor het ontwikkelen van het risico predictie model zijn alle beschikbare gegevens uit de huisartsenregistratie van Nivel Zorgregistraties Eerste Lijn gebruikt over de periode 2009 – 2017.

## 2.2 Patiënten en controles

Als eerste zijn alle volwassen patiënten (18 jaar en ouder) geselecteerd die in de periode 2013-2017 een eerste HVZ of een risicofactor voor HVZ ontwikkelden. Patiënten mochten dus voor 1 januari 2013 nog geen geregistreerde (risicofactor voor) HVZ hebben (zie Tabel 1) of geen aan HVZ gerelateerde medicatie voorgeschreven hebben gekregen (zie Tabel 2). De data over de periode 2009-2012 zijn gebruikt om dit uit te sluiten.

Basis voor de definitie van HVZ vormt de NHG richtlijn cardiovasculair risicomanagement [10]. Aanvullend zijn ook enkele aanverwante cardiometabole aandoeningen en risicofactoren meegenomen, zoals hypercholesterolemie, hypertensie, diabetes mellitus, reumatoïde artritis en psoriasis (zie Tabel 1).

*Tabel 1 Geselecteerde aandoeningen en bijbehorende ICPC codes*

| Diagnose | ICPC |
|---|---|
| Angina pectoris | K74 |
| Acuut myocardinfarct | K75 |
| Andere/chronische ischemische hartziekte | K76 |
| Hypertensie | K86/K87 |
| TIA | K89 |
| CVA | K90 |
| Claudicatio | K92 |
| Aneurysma aorta | K99 |
| Reumatoïde artritis | L88 |
| Diabetes mellitus | T90 |
| Psoriasis | T91 |
| Hypercholesterolemie | T93 |

De aandoeningen uit Tabel 1 zijn voor sommige analyses geclusterd in drie groepen. De groep chronische HVZ bevat angina pectoris, claudicatio, aneurysma aorta en andere/chronische ischemische hartziekten. De groep acute HVZ bestaat uit TIA, CVA en acuut myocardinfarct. En de groep risicofactoren bevat de diagnoses hypertensie en hypercholesterolemie.

Naast de geregistreerde diagnose is gekeken naar geneesmiddelen die gerelateerd zijn aan HVZ (zie Tabel 2).

*Tabel 2 Geselecteerde ATC-codes gerelateerd aan HVZ*

| Medicatie | ATC3 |
|---|---|
| Antihypertensiva | C02/C03/C07/C08/C09 |
| Antilipaemica | C10 |
| Antitrombotica | B01A |
| Vasodilatoren voor cardiale aandoeningen | C01D |
| Diabetesmedicatie | A10 |
| Antipsoriatica | D05 |

Bij patiënten die geneesmiddelen uit Tabel 2 voor 1 januari 2013 kregen voorgeschreven, zijn we er vanuit gegaan dat zij al een bestaande (risicofactor voor) HVZ hadden. Deze patiënten zijn uitgesloten van de analyse. Wanneer bij patiënten deze geneesmiddelen tussen 1 januari 2013 en de geregistreerde diagnosedatum werd voorgeschreven, hebben we de diagnosedatum (casedatum) en eventueel de diagnose aangepast volgens het schema in Tabel 3.

*Tabel 3 Overzicht aanpassingen casedatum op basis van medicatie voorschriften voor de casedatum*

| Medicatie | <3 maand voor geregistreerde diagnose |
|---|---|
| Antihypertensiva | • Casedatum aanpassen naar datum eerste voorschrift<br>• Diagnose hypertensie wanneer >3 mnd voor geregistreerde diagnose |
| Antilipaemica | • Casedatum aanpassen naar datum eerste voorschrift<br>• Diagnose hypercholesterolemie wanneer >3 mnd voor geregistreerde diagnose |
| Antitrombotica | • Casedatum aanpassen naar datum eerste voorschrift bij aandoeningen uit K-hoofdstuk, behalve hypertensie. Bij overige aandoeningen patiënt uitsluiten. |
| Vasodilatoren voor cardiale aandoeningen | • Casedatum aanpassen naar datum eerste voorschrift bij angina, myocardinfarct en claudicatio. Anders patiënt uitsluiten. |
| Diabetesmedicatie | • Casedatum aanpassen naar datum eerste voorschrift<br>• Diagnose diabetes |
| Antipsoriatica | • Casedatum aanpassen naar datum eerste voorschrift<br>• Diagnose psoriasis |

Naast de gedefinieerde patiëntengroep is er ook een groep controles geselecteerd die geen geregistreerde HVZ hadden in de periode 2009-2017. Zowel patiënten als controles moesten minstens 4 kwartalen in de huisartspraktijk ingeschreven staan. Per case zijn 2 controle patiënten uit dezelfde praktijk random gekozen. Een controle patiënt kan maar één keer meedoen als controle patiënt (trekking zonder terugleggen).

## 2.3 Voorbewerking data

### 2.3.1 Voorspellers

Als mogelijke voorspellers voor het hebben van hoog risico op HVZ, zijn de volgende type variabelen in de modellen meegenomen:

- Aandoeningen waarvoor de huisarts werd bezocht (o.b.v. ICPC-codes) en het aantal consulten per ICPC hoofdstuk.
- Geneesmiddelen die werden voorgeschreven (o.b.v. ATC-codes, niveau 3).
- Indien beschikbaar werden de volgende meetwaarden van lichamelijk of laboratorium onderzoek meegenomen: BMI, rookstatus, systolische- en diastolische bloeddruk.

### 2.3.2 Bewerken van de data

Diagnoses zijn los meegenomen, maar ook geclusterd naar ICPC-hoofdstuk (17 verschillende hoofdstukken), welke uiteenlopen van algemene klachten en huidaandoeningen tot psychologische of sociale problemen. Per hoofdstuk is weer een onderscheid gemaakt in de codes die klachten weergeven (codes 1 tot en met 29) en de codes die aandoeningen weergeven (codes 70 tot en met 99). Zo is bijvoorbeeld het hoofdstuk D ingedeeld in de codes D1 tot en met D29 huidklachten, en D70 tot en met D99 als huidaandoeningen.

ATC-codes zijn meegenomen op ATC-3 niveau, dat wil zeggen de therapeutische/farmacologische subgroep. Een voorbeeld is middelen bij ulcus pepticum en gastro-oesofageale refluxzijn met ATC3 code A02B.

De geselecteerde meetwaarden roken, systolische- en diastolische bloeddruk en BMI worden onregelmatig en beperkt gemeten en geregistreerd in de huisartsenpraktijk en waren dus niet voor alle patiënten beschikbaar. De huisarts zal namelijk deze metingen alleen uitvoeren (en registreren) als daar een indicatie voor is. De meest recente meetwaarde in de periode 3 tot 12 maanden voor het event is meegenomen. Dit betekent dat de bloeddruk of bijvoorbeeld rookstatus van maximaal 1 jaar voor het event gebruikt kan zijn om te voorspellen. De vier meetwaarden zijn gehercodeerd naar twee categorieën om de interpretatie van de analyses te verbeteren. Roken en BMI zijn op twee verschillende manieren gehercodeerd en elk afzonderlijk in de analyses meegenomen. De meetwaarden zijn gehercodeerd tot dichotome variabelen (zie Tabel 4).

*Tabel 4 Overzicht hercoderingen meetwaarden tot dichotome variabelen*

| Meetwaarde | Hercodering |
|---|---|
| Roken | • Nee/Voorheen versus Ja<br>• Nee versus Voorheen/Ja |
| Systolische bloeddruk | • Normaal (60-139 mm Hg) versus Afwijkend (<60 of >=140 mmHg) |
| Diastolische bloeddruk | • Normaal (40-90 mmHg) versus Afwijkend (<40 of >=90 mm Hg) |
| BMI | • Ondergewicht/Normaal (<25 kg/m$^2$) versus Overgewicht/Obesitas (>=25 kg/m$^2$)<br>• Ondergewicht/Normaal/Overgewicht (<30 kg/m$^2$) versus Obesitas (>=30 kg/m$^2$) |

### 2.3.3 Aggregatie van de data

Alle voorspellende variabele zijn geaggregeerd naar kwartalen om de densiteit van de data te vergroten en data van individuele patiënten beter vergelijkbaar te maken. Voor de individuele ATC en ICPC codes is gekeken of zij per kwartaal tenminste één keer gerapporteerd waren. Daarnaast is per hoofdstuk (uitgesplitst naar de groep klachten en aandoeningen), berekend hoeveel contacten iemand in een kwartaal had. Voor leeftijd is de waarde meegenomen ten tijde van het event. Voor de meetwaarden is de laatst bekende waarde binnen een kwartaal meegenomen.

Informatie over het kwartaal voorafgaand aan het event is niet meegenomen als voorspeller (een time-lag van drie maanden) om rekening te houden met mogelijk verlate registratie van de diagnose, bijvoorbeeld wanneer deze afkomstig is uit ziekenhuisbrieven. We gebruikten de informatie van 3-12 maanden voor het event als voorspeller. Hier is specifiek voor gekozen omdat het waarschijnlijk is dat er al meerdere kwartalen voor een diagnose tekenen in de data te vinden zijn dat de aandoening zich aan het ontwikkelen is. Door het volledige jaar voor de diagnose, behalve het laatste kwartaal voor het event, mee te nemen in de voorspelling houden we hier rekening mee. Voor controles namen we dezelfde kwartalen mee als de case waar ze aan gematcht zijn.

Om data van de drie kwartalen, die samen als voorspellende periode werden gebruikt, te aggregeren, hebben we het gemiddelde aantal consulten per ICPC hoofdstuk in die periode berekend, en net als bij de eerdere aggregatie gekeken of ATC en ICPC codes in die periode tenminste één keer gerapporteerd waren. Voor geslacht, leeftijd en meetwaarden is wederom de laatste waarde binnen deze periode meegenomen. Ook is er een extra variabele toegevoegd, namelijk het gemiddelde totale aantal consulten per kwartaal in deze periode.

### 2.3.4 Variabele selectie en preparatie

Om het totale aantal voorspellers minder omvangrijk te maken, is bekeken of er ICPC en ATC codes en hoofdstukken zijn die nagenoeg niet geobserveerd zijn bij de patiënten. Voorspellers zijn verwijderd wanneer zij bij minder dan 0,1% van de patiënten geobserveerd waren.

Alle numerieke variabelen zijn gecentreerd rond hun gemiddelde en gestandaardiseerd om te voorkomen dat voorspellers met een hoge variantie alleen al om deze reden een hogere mate van belangrijkheid (importance) krijgen toegeschreven in de machinelearning-analyses.

Tenslotte is elk van de 12 databestanden (voor iedere (risicofactor voor) HVZ een eigen databestand), in een training en een test set uitgesplitst. Op de training set wordt het meest geschikte predictiemodel ontwikkeld. Voor de patiënten in de test set wordt vervolgens hun kans op het ontstaan van de aandoening berekend op basis van het in de training set gemaakte predictiemodel, en kan de accuraatheid van deze voorspelling worden getest. Bij het construeren van de training en test set is gekozen voor een verhouding van 4 op 1; 80 procent van de patiënten uit het databestand zijn toegewezen aan de training set en 20 procent aan de test set.

Leeftijd, geslacht en de vier meetwaarden zijn hieraan toegevoegd om tot een uiteindelijke selectie van variabelen te komen voor gebruik in de machinelearning-analyses. Leeftijd en geslacht zijn in alle modellen meegenomen; vervolgens zijn deze modellen nog een keer uitgevoerd met toevoeging van meetwaarden. Dit alleen voor die patiënten waarvoor de desbetreffende meetwaarde bekend is.

### 2.3.5 Voorbereidende analyses

De volgende stap om het potentieel aantal predictoren enigszins te bepreken is met behulp van univariabele logistische regressie gedaan. Per potentiele predictor is gekeken of deze significant verschilde tussen de patiënten en controles. Voor iedere van de 12 uitkomstmaten (HVZ of risicofactoren voor HVZ), zijn de 50 aandoeningen en 100 geneesmiddelen met de hoogste odds ratios geselecteerd.

Zoals eerder beschreven zijn de vier meetwaarden, leeftijd en geslacht buiten deze voorselectie gelaten. Deze variabelen gaan sowieso mee in de machinelearning-modellen. Leeftijd en geslacht omdat dat bekende voorspellers zijn; meetwaarden omdat we dit alleen van een subgroep hebben waar deze waarden gemeten en geregistreerd zijn in het EPD van de huisarts.

## 2.4 Machine learning

### 2.4.1 LASSO regressie

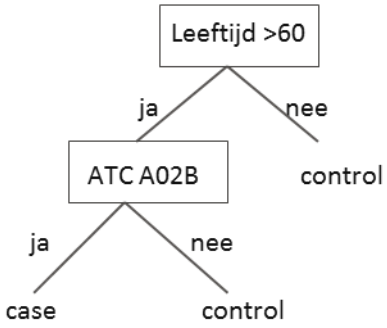
LASSO regressie is een extensie van de meer traditionele regressie, zoals die veel gebruikt wordt in de sociale wetenschappen. LASSO is een afkorting van Least Absolute Shrinkage and Selection Operator. De methode combineert regressiemethoden met variabelen selectie. Vergeleken met standaard regressie kan LASSO regressie beter omgaan met data met veel variabelen, en met variabelen die niet direct een verband hebben met de te voorspellen variabele. Middels een zogenaamde ‘tuning parameter’ krimpt het LASSO algoritme de regressie coëfficiënten, en zet sommige coëfficiënten die te weinig voorspellende waarde hebben, op nul. LASSO is een populaire en veel toegepaste machinelearning-techniek, die bij grote datasets tot betere voorspellingen dan standaard regressie leidt [11].

### 2.4.2 Rondom forest

Random forest is een populaire machinelearning-methode met als doel observaties zo goed mogelijk te classificeren. Hierbij wordt gebruik gemaakt van beslisbomen om tot de best mogelijke voorspelling voor elke observatie te komen. Per beslisboom wordt de voorspeller, die het sterkste effect heeft op de uitkomst, boven in de beslisboom opgenomen, gevolgd door voorspellers die daarna volgen in voorspellingskracht. Het is een complexer machinelearning-algoritme vergeleken met LASSO regressie, omdat het naast de lineaire hoofdeffecten tussen variabele ook kijkt naar alle niet lineaire interacties tussen variabelen.

Een voorbeeld van een deel van een typische beslissingsboom voor deze studie is weergegeven in Figuur 1. In deze beslisboom is de meest belangrijke voorspeller leeftijd, en specifiek of iemand ouder is dan 60. Daarom wordt op basis daarvan de dataset gesplitst in mensen die ouder zijn dan 60 en degenen die jonger zijn. Vervolgens wordt er gekeken of er binnen deze subgroepen nog variatie is in het voorkomen van cases en controles, en of er variabelen beschikbaar zijn die deze groepen nog verder van elkaar kunnen onderscheiden. In dit voorbeeld is het voor mensen van 60 jaar en jonger niet mogelijk om een verdere nuttige splitsing te maken, en wordt voor iedereen binnen deze groep voorspeld dat ze control zijn. Voor mensen ouder dan 60 jaar is dit wel het geval: cases en controles kunnen verder van elkaar worden onderscheiden door te kijken of de ATC code B01a (anti-trombose medicatie) voorgeschreven was in de periode van 3-12 maanden voor diagnose.

*Figuur 1 Een voorbeeld van een beslisboom zoals gebruikt in random forest*




Een kenmerk van random forest classificatie is dat niet één, maar een groot aantal beslisbomen worden gemaakt. Deze beslisbomen worden telkens met een aselecte steekproef van alle voorspellers gemaakt op een subset van alle observaties. Elke beslisboom heeft hierdoor andere beslispunten waardoor ze deels onafhankelijk van elkaar zijn. Uiteindelijk worden voor elke observatie de schattingen van alle beslisbomen gemiddeld om tot één modelschatting te komen [11].

### 2.4.3 Datasets

De analyses zijn uitgevoerd voor verschillende combinaties aan voorspellers. Ten eerste is een selectie gemaakt van alle voorspellers behalve de meetwaarden. Dit omdat veel meetwaarden slechts voor een klein aantal individuen beschikbaar waren, waardoor de dataset veel kleiner zou worden als alleen individuen met beschikbare data voor de meetwaarden zouden worden meegenomen. Om alsnog het effect van de meetwaarden te kunnen evalueren is per meetwaarde een subset van de data gemaakt voor de individuen waarbij de meetwaarde gemeten was. Voor elk van deze subsets is de random forest analyse uitgevoerd met de meetwaarde als voorspeller en zonder de meetwaarde mee te nemen. Op deze manier kon de toegevoegde waarde van de meetwaarde voor het voorspellen van de diagnose goed worden gemeten. Met een kleinere dataset wordt namelijk sowieso een verschil in prestatie van de random forest verwacht. De subsets met en zonder meetwaarden hadden allebei even veel observaties, waardoor het enige verschil tussen de datasets het toevoegen van de meetwaarde was.

### 2.4.4 Model evaluatie

Om de prestaties van de random forest modellen per dataset en diagnose te evalueren is de voorspelde waarde vergeleken met de geobserveerde waarde. Deze vergelijking wordt door middel van verschillende uitkomstmaten weergegeven, omdat iedere uitkomstmaat een andere nadruk heeft. Het doel van het model was niet de perfecte voorspeller, maar een betere voorspeller dan iedereen zonder (risicofactoren voor) HVZ tussen de 45 en 70 jaar. Dus stel dat in die hele groep de kans om iemand te vinden met een hoog risico op (risicofactoren voor) HVZ 5% is (1 op de 20), en door gebruik te maken van dit model kan die kans al verhoogd worden tot 20% (1 op de 5), dan is dat al winst in efficiëntie van opsporing. De discussie of dit genoeg winst is, en het daarmee een bruikbaar model is, moet gevoerd worden met de praktijk zelf en haar stakeholders. Het doel van deze studie was om te onderzoeken of en in welke mate de kans op het vinden van patiënten met een verhoogd risico op (risicofactoren voor) HVZ verbeterd kan worden door middel van het gebruik van machine learning op gegevens uit het EPD van de huisarts. De vergelijking tussen de geobserveerde en voorspelde uitkomst wordt weergegeven in verschillende uitkomstmaten. Een belangrijke uitkomstenmaat bij ongebalanceerde datasets is de kappa. Een kappa geeft aan hoeveel

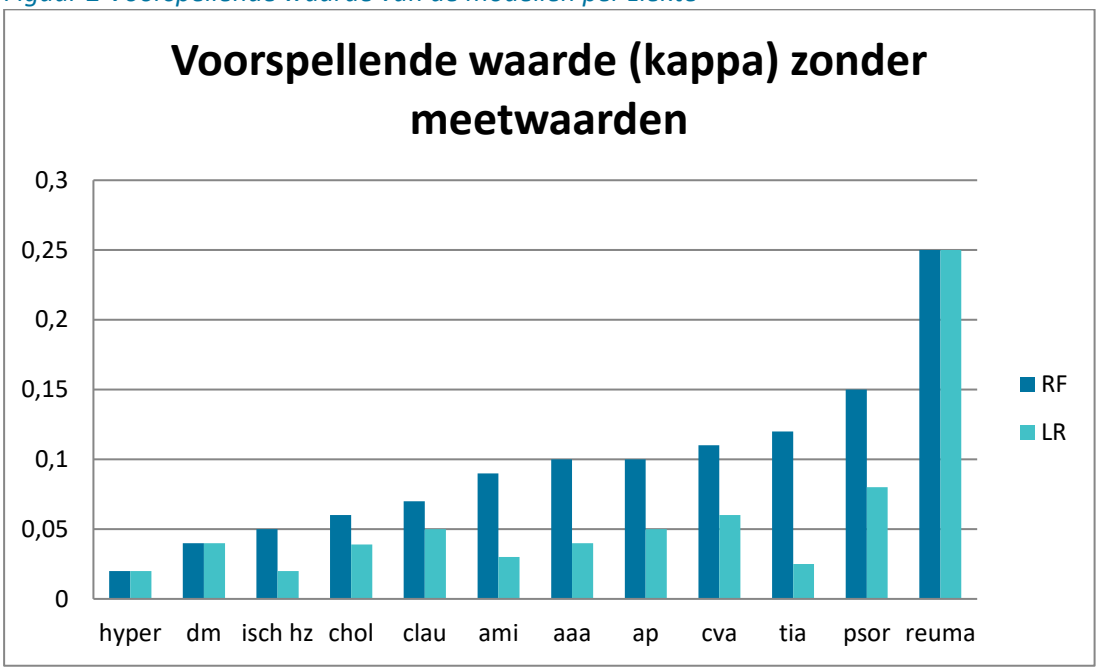
beter een model het doet ten opzicht van alleen toeval. Een kappa van 0 betekent dat je net zo goed een muntje op kan gooien, en een kappa van 1 geeft aan dat het model iedereen goed classificeert. Een ondergrens in de literatuur is een kappa van 0.2.

# 3 Resultaten

## 3.1 Uitkomsten voorspellingen

In Figuur 2 vatten we de resultaten voor alle aandoeningen samen. Zoals eerder beschreven, geeft de kappa aan in hoeverre een voorspelling het beter doet dan het gooien met een muntje, met 0.2 als minimum. De enige uitkomst die boven de grenswaarde scoort is reumatoïde artritis. Psoriasis wordt daarna het beste voorspelt. Bij de meeste gevallen zijn de voorspellingen van de random forest beter dan de voorspellingen van de LASSO regressie. Bij diabetes mellitus en reumatoïde artritis doen beide algoritmes het even goed.

*Figuur 2 Voorspellende waarde van de modellen per ziekte*



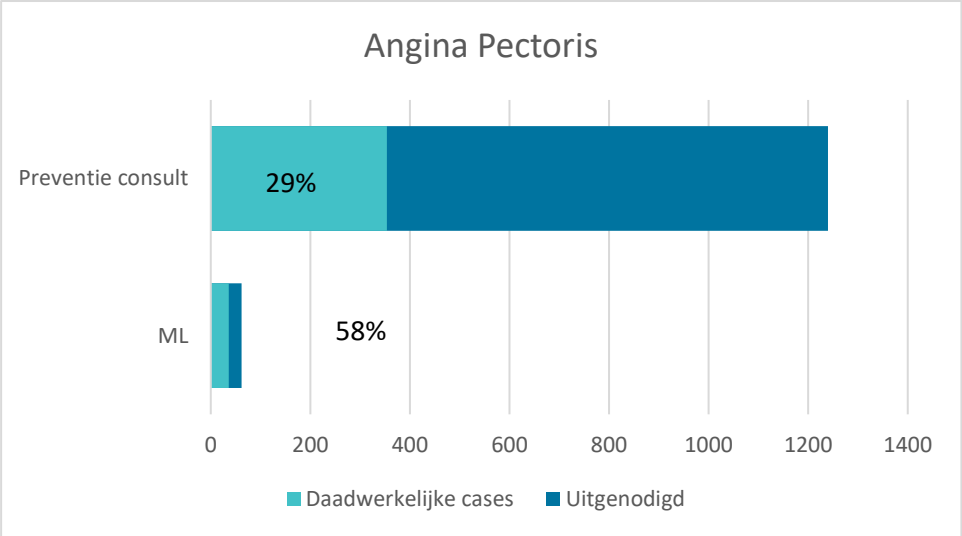

hyper = hypertensie, dm = diabetes mellitus, isch hz = Andere/chronische ischemische hartziekte, clau = Claudicatio intermittens, ami = acuut myocardinfarct, aaa = Aneurysma Aortae Abdominalis, ap = agina pectoris cva = Cerebrovasculair accident, tia = transient ischemic attack, psor = psoreasis, reuma = reumatoide artritis, RF = random forest, LR = LASSO Regression. Kappa = hoeveel beter het algoritme voorspelt dan het gooien met een muntje. Als kappa 0 is dan is de voorspelling even goed als kans, hoe hoger, hoe beter de voorspelling van het algoritme is.

Maar wat betekenen deze uitkomsten nu concreet? In onze dataset hebben we bij elke case twee controles gezocht. Daarna hebben we nog enige databewerkingen uitgevoerd, zodat de daadwerkelijk verhouding iets afwijkt van 33%. Deze afwijking was verschillende voor iedere uitkomstmaat. Om alle daadwerkelijke cases te vinden moet de huisarts in de huidige situatie elke patiënt uitnodigen en screenen. Door te werken met de machinelearning-algoritmes hopen we de pakkans te verhogen, de te onderzoeken groep patiënten te verkleinen, of allebei. Hieronder werken we dit uit voor een aandoening, namelijk angina pectoris (zie Figuur 3).

Onze dataset voor angina pectoris bevat 354 patiënten (cases) en 886 controles; van de 1240 patiënten is 29% dus een case. Om alle cases te vinden moet een huisarts in de huidige situatie alle 1240 patiënten uitnodigen en screenen.

Wanneer we het random forest algoritme zouden toepassen, en daarna alleen degene die positief zijn uitnodigen en screenen, dan zouden slechts 62 patiënten worden uitgenodigd en gescreend. Hiervan zijn er 36 'echte' cases (58%). Dus na het toepassen van het algoritme worden aanzienlijk minder patiënten uitgenodigd, 54 i.p.v. 1240 terwijl de opbrengst van het tijdig identificeren van cases hoger is (58% in plaats van 29%). Echter, door het toepassen van het algoritme zijn 330 cases, patiënten met een hoog-risico op het ontwikkelen van angina pectoris, onterecht niet uitgenodigd en worden daardoor minder tijdig herkend. De afweging die de praktijk hierbij moet maken is of de 'winst' van minder patiënten screenen met een hogere opbrengst, opweegt tegen het missen van een groep potentieel hoogrisico-patiënten. Hierbij moet de belasting voor de huisarts om al deze patiënten te screenen, en de haalbaarheid daarvan, meegenomen worden als ook de ernst van het 'missen' van een behoorlijk aantal hoog-risico patiënten. De uitkomsten voor de andere aandoeningen staan in de appendix.

*Figuur 3: Verhoogde kans om een patiënt vroegtijdig te detecteren na gebruik van selectie algoritmes*




## 3.2 Toevoegen van meetwaarden

Niet voor alle patiënten bleken er meetwaarden in het EPD aanwezig te zijn. Daarom is een directe vergelijking met de analyses met en zonder meetwaarden lastig te maken. Daar komt nog bij dat het hebben van meetwaarden niet random is; vermoedelijk was er voor het uitvoeren van een diagnostische test namelijk een medische aanleiding. Wanneer we de beschikbare gegevens van de verschillende meetwaarden (roken, BMI en bloeddruk) toevoegden aan de gehele dataset, voor de cases en controles waarvoor deze informatie beschikbaar was, zagen we in de meeste gevallen zeer weinig verbetering of soms zelf een verslechtering van de voorspellende waarde. De voorspellende waarde werd wel iets beter bij bijvoorbeeld angina pectoris wanneer we informatie hadden of patiënten rookten of niet. Soms werden de voorspellingen slechter, zoals het geval was bij TIA.

## 3.3 Belangrijkste voorspellers

Per aandoening hebben we een lijst gemaakt van de tien belangrijkste voorspellers. Omdat LASSO regressie en random forest op een andere manier tot de beste voorspelling komen, verschillen ook de geselecteerde tien belangrijkste voorspellers. De verwachting was dat leeftijd en geslacht vaak een belangrijke rol bij de voorspellingen zou spelen.

Dit zien we terug bij random forest, maar niet bij LASSO. Bij LASSO komen deze variabelen wel voor, maar niet als eerste en vaak ook niet in de top tien. Wat opviel was dat de voorgeschreven geneesmiddelen, naast aandoeningen, belangrijke voorspellers waren bij LASSO, maar een groep symptomen of diagnoses binnen een ICPC hoofdstuk niet of nauwelijks. Bij random forest zien we geneesmiddelen en juist symptomen of aandoeningen binnen een ICPC hoofdstuk vaker als belangrijk voorspeller en individuele ICPC-codes veel minder vaak. Bij random forest was het aantal contacten ook vaak een belangrijke voorspeller; bij LASSO kwam deze variabele niet voor in de top tien. Ter illustratie presenteren we de top tien belangrijkste voorspellers voor angina pectoris op basis van LASSO en random forest technieken in Tabel 5.

*Tabel 5 Top 10 beste voorspellers van Angina pectoris**

| LASSO regressie | Random forest |
|---|---|
| T81 (Struma/moduli) | leeftijd |
| M09A (overige middelen voor problemen van het skeletspierstelsel) | aantal contacten |
| X75 (Maligniteit cervix/uteri) | geslacht |
| S01F (Mydriatica en cycloplegica) | A02B (Middelen bij ulcus pepticum en gastro-oesofageale reflux) |
| K02 (Druk beklemming hart) | J01C (Betalactam-antibiotica penicillines) |
| T83 (Overgewicht) | M01A (Niet-steroide anti-inflammatoire en antireumatische midd.) |
| C01B (Anti-aritmische middelen klasse i en iii) | Ds (symptomen maag-darm kanaal) |
| D11A (Overige dermatologische preparaten) | Rs (symptomen luchtwegen) |
| A11D (Vitamine b1 enkelvoudig en met vitamine b6 en/of b12) | Rd (diagnoses luchtwegen) |
| R03D (Overige middelen bij astma/copd voor systemisch gebruik) | Ld (diagnoses bewegingsapparaat) |

* Geneesmiddelen zijn in blauw gepresenteerd voor de herkenbaarheid.

De resultaten van random forest voor alle aandoeningen staan in bijlage B.

# 4 Discussie

In deze studie onderzochten we de potentie van machine learning om hoogrisico-patiënten voor (risicofactoren voor) HVZ op te sporen. We maakten daarbij gebruik van routinematig verzamelde gegevens uit de EPDs van huisartsenpraktijken, die deelnemen aan Nivel Zorgregistraties, met diagnose informatie tijdens zorgconsulten en voorschriften van geneesmiddelen . Waar mogelijk verrijkten we de data met meetwaarden (BMI, roken en bloeddruk). We maakten gebruik van twee veel gebruikte machinelearning-algoritmen: LASSO regressie en random forest.

De resultaten van het onderzoek lieten zien dat het op basis van de beschikbare data in het EPD van de huisarts en de door ons gebruikte technieken nagenoeg niet mogelijk was om tot een goed voorspellend model te komen. Alleen de voorspelling voor één risicofactor, namelijk reumatoïde artritis, was acceptabel. Het algoritme detecteerde 1 op de 4 patiënten dat binnen zes maanden reumatoïde artritis zou ontwikkelen correct. Alle andere aandoeningen, zoals hypertensie of een herseninfarct, waren niet goed te voorspellen op basis van de beschikbare data. Wel zorgde de voorselectie door de algoritmes ervoor dat de huisarts minder patiënten zou moeten screenen, met een grotere pakkans binnen deze kleine groep patiënten. Aangezien de huisarts nooit alle patiënten met een eventueel risico op HVZ kan uitnodigen voor consult, valt deze uitkomst als winst te zien. Daarmee wordt de screening in potentie dus aanzienlijk efficiënter. Het grote probleem is dat het algoritme teveel cases miste. Blijkbaar bevat de huisartsendata niet genoeg onderscheidende informatie waarop een automatische selectie gemaakt kan worden. Vergelijkbare resultaten werden gevonden in een Amerikaanse studie, waarin de machinelearning-methoden op basis van routinematig verzamelde data geen betere voorspellingen opleverde om hartfalen te voorspellen vergeleken met standaard logistische regressie [12]. Net als in die studie zagen we dat het toevoegen van de beschikbare informatie over de meetwaarden de voorspellingen niet verbeterde.

Een belangrijke reden waarom reumatoïde artritis het beste te voorspellen was op basis van eerder zorggebruik, is dat reumatoïde artritis, vergeleken met de andere aandoeningen, een duidelijker preklinisch beeld heeft. In het voorstadium van reuma zal de patiënt al op consult geweest zijn voor klachten aan het bewegingsapparaat. Voorspellende factoren op basis van random forest waren dan ook symptomen in het L-hoofdstuk (bewegingsapparaat), de losse ICPC-code L20 (Symptomen meerdere/niet-gespecificeerde gewrichten), en geneesmiddelen als NSAID's. Bij de overige aandoeningen zoals bijvoorbeeld angina pectoris of hypertensie is het preklinische beeld veel minder eenduidig.

De voorspellingen met het complexere machinelearning-algoritme random forest waren in de meeste gevallen beter dan de voorspellingen met de machinelearning-techniek LASSO regressie. Random forest kan beter met de complexe niet-lineaire relatie tussen variabelen omgaan dan LASSO regressie, wat de voorspellende waarde kan verhogen. De coëfficiënten van de LASSO regressie zijn wel makkelijker te interpreteren, namelijk vergelijkbaar met een standaard logistische regressie. De random forest geeft geen interpreteerbare coëfficiënten, maar alleen een rangorde van de belangrijkste voorspellers. Wanneer het doel van het algoritme is om zo goed mogelijk te voorspellen hebben complexere algoritmen zoals random forest of neural networks de voorkeur. Wanneer interpretatie van belang is heeft LASSO regressie de voorkeur.

Wanneer we kijken naar de belangrijkste voorspellende variabelen bij de random forest modellen valt op dat bij bijna alle aandoeningen leeftijd, geslacht en het aantal contacten met de huisartsenpraktijk in de top 10 staan. Deze drie variabelen als sterke voorspellers zijn in lijn met de verwachtingen en ook al eerder gevonden. Minder voor de hand liggend was dat maagzuurremmers bijna altijd in de top 10 staan. Echter de voorspellende waarde van maagzuurremmers op zich is weer heel laag. Indien we een simpele logistische regressie analyse draaide met alleen de variabelen leeftijd, geslacht, aantal contacten en maagzuurremmers kwam daar een model uit met een zeer lager voorspellende waarde.

De variabelen die de meeste bijdrage hadden aan de voorspellende modellen verschilden enorm tussen LASSO en random forest; bij een aantal aandoeningen was er geen enkele variabele in de top tien hetzelfde. Allereerst zijn LASSO en random forest twee verschillende algoritmen; LASSO is een lineair model en random forest niet. Random forest neemt complexere relaties mee, namelijk het houdt rekening met niet-lineaire relaties tussen de voorspellers onderling en de uitkomst. LASSO neemt deze niet vanzelf mee, alleen als die er actief aan toegevoegd worden. De moeilijkheid daarbij is dat het vooraf onduidelijk is welke niet-lineaire relaties toegevoegd moeten worden. Een ander verschil is dat de variabelen in LASSO die in de top tien staan gebaseerd zijn op de magnitude van de regressiecoëfficiënten, terwijl random forest een criterium op basis van 'variable importance' hanteert. Dit wil zeggen de volgorde is op basis van welke variabelen voor de beste verbetering in de predictie zorgen. De bevinding dat LASSO en random forest totaal verschillende modellen opleveren is niet nieuw. Dit is al eerder beschreven in de literatuur en diverse fora op internet. (zie o.a. https://stats.stackexchange.com/questions/155192/why-discrepancy-between-lasso-and-randomforest).

Het doel van dit onderzoek was om te onderzoeken of machinelearning-technieken een bijdrage kunnen leveren aan het identificeren van patiënten met een verhoogd risico op HVZ of risicofactoren voor HVZ. De resultaten waren minder goed dan verwacht, en zijn nog lang niet geschikt voor toepassing in de praktijk. Toch wil dit niet zeggen dat machine learning niet kan zorgen voor een verbeterde voorspelling. Wanneer gegevens uit het EPD van de huisarts worden uitgebreid met extra variabelen, die nu niet standaard in het EPD worden geregistreerd, zou de voorspelling aanzienlijk kunnen worden verbeterd. Zo vonden we nu dat het toevoegen van rookstatus, BMI en bloeddruk niet tot betere voorspelling leidden, terwijl bekend is dat dit belangrijke risicofactoren zijn voor het krijgen van HVZ. Een enorme beperking was dat er van de meeste patiënten geen meetwaarden bekend waren. Daar komt bij dat bij de personen waar het wel bekend van was, er vermoedelijk vanwege een medische reden is gemeten, en er dus geen sprake is van een random selectie. Een verbetering van het model zou mogelijk kunnen zijn als we deze meetwaarden voor alle patiënten beschikbaar zouden hebben. Mogelijk kan de Persoonlijk Gezondheids Omgeving (PGO) in de toekomst uitkomst bieden als patiënten zelf via hun PGO meer data met hun huisarts kunnen delen.

Een ander mogelijk verbeterpunt zit in de gekozen tijdsperiode. We hebben gebruik gemaakt van de laatste 3 tot 12 maanden voor het event, en hebben de meest recente waarde genomen. We hebben echter meer informatie per patiënt. Zo zouden eerdere metingen en veranderingen over de tijd binnen patiënten nog meegenomen kunnen worden in de analyses, en kan er ook langer terug gekeken worden in de tijd; mogelijk tot twee of drie jaar voor het event. Dat biedt dan tevens meer tijd voor preventie. In onze analyse hebben we tijd platgeslagen en verdeeld in blokken van drie maanden. Wanneer je per patiënt wilt kijken naar de afstand in tijd tussen consulten zal je dit soort technieken moeten gebruiken. Deep learning algoritmes, d.w.z. algoritmes die gebaseerd zijn op neurale netwerken, worden momenteel steeds meer toegepast om tijd dynamisch te modelleren [13].

De lage opbrengst kan ook samenhangen met de gekozen klinische uitkomst, namelijk het voorspellen van (risicofactoren voor) HVZ op basis van eerder zorggebruik. Mogelijk zijn dit aandoeningen en risicofactoren die zich te generiek presenteren, dus zonder een uniek preklinisch beeld. Een verwachting is dan ook dat de meerwaarde van machine learning vooral te verwachten is bij aandoeningen met een sterk onderscheidende preklinische fase, zoals bijvoorbeeld bij sommige vormen van kanker. Een andere manier om de voorspelling te verbeteren is om alleen te kijken naar een subset van patiënten met al een iets verhoogd risico. Zo zochten we in een eerdere studie naar suïcide(pogingen) naar verschil in zorggebruiker tussen depressieve patiënten met en zonder suïcide(pogingen). Dit werkte veel beter dan wanneer we alle patiënten includeerden om suïcide(pogingen) te voorspellen (resultaten nog niet gepubliceerd). Mogelijk dat het voorspellen van het krijgen van nieuwe HVZ voor de patiënten die al in CVRM zitten wel tot bruikbare modellen leidt.

Tot nu toe hebben alle bij ons bekende studies analyses uitgevoerd op historische data. Op basis van deze historische data voorspellen we het krijgen van een aandoening, waarvan we al weten dat die patiënt deze daadwerkelijk gekregen heeft. Dit is nog niet hetzelfde als wanneer we het zouden implementeren in de praktijk. Dan zou een patiënt die binnenkomt op basis van zijn eigen gegevens een risicoscore ten opzichte van de andere patiënten in een grotere database moeten krijgen. Deze risicoscore moet real time worden geüpdatet op basis van het consult dat de patiënt op dat moment heeft. Hoe goed machinelearning-modellen deze overstap van historische data naar real time maken moet ook nog uitgezocht worden.

## Conclusie

De resultaten van machinelearning-technieken, LASSO of random forest, op beschikbare EPD gegevens van huisartsen om patiënten met een verhoogd risico op het ontwikkelen van (risicofactoren voor) HVZ te voorspellen, zijn nog van onvoldoende kwaliteit om tot een voorspellend model te komen met voldoende betrouwbaarheid. Er werd een kleine verbetering gevonden in de patiëntselectie met daarbij een grote groep potentieel hoogrisico-patiënten die niet geïdentificeerd werden als zodanig. Er moeten nog de nodige stappen gezet worden voordat deze technieken een meerwaarde kunnen hebben in de praktijk. Allereerst is het van belang de beschikbare data uit te breiden, bijvoorbeeld met meer persoonlijke gegevens van de patiënt. Daarnaast zijn er nieuwe machinelearning-technieken beschikbaar die de voorspelling zouden kunnen verbeteren. Ook de randvoorwaarde waaraan de algoritmes moet voldoen voor daadwerkelijke toepassing in de praktijk moeten nog beter uitgezocht worden. Pas als hier stappen in gezet zijn kan het mogelijk worden dat machinelearning-technieken in de dagelijkse huisartsen praktijk tot geautomatiseerde voorspellingen kunnen leiden.

# Literatuur


1. Van der Meer V, Nielen MM, Drenthen AJ, Van Vliet M, Assendelft WJ, Schellevis FG. Cardiometabolic prevention consultation in the Netherlands: screening uptake and detection of cardiometabolic risk factors and diseases--a pilot study. BMC Fam Pract. 2013;14:29.

2. Stol DM, Hollander M, Badenbroek IF, Nielen MMJ, Schellevis FG, de Wit NJ. Uptake and detection rate of a stepwise cardiometabolic disease detection program in primary care-a cohort study. Eur J Public Health. 2019. pii: ckz201.

3. Nielen M, Davids R, de Bakker D. Het PreventieConsult Cardiometabool risico. Huisarts Wet. 2011;3:121.

4. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One. 2017;12(4):e0174944.

5. Barak-Corren Y, Castro VM, Javitt S, Hoffnagle AG, Dai Y, Perlis RH, Nock MK, Smoller JW, Reis BY. Predicting Suicidal Behavior From Longitudinal Electronic Health Records. Am J Psychiatry. 2017;174(2):154-162.

6. Reis BY, Kohane IS, Mandl KD. Longitudinal histories as predictors of future diagnoses of domestic abuse: modelling study. BMJ. 2009;339:b3677.

7. Website Nivel Zorgregistraties Eerste Lijn. 2018: www.nivel.nl/nl/nzr/zorgregistraties-eerstelijn.

8. Lamberts H, Wood M. ICPC, International Classification of Primary Care. Oxford: Oxford University Press; 1987.

9. World Health Organization collaborating center for drug statistics methodology. Guidelines for ATC cliassification and DDD assignment 2010. Oslo: WHO; 2009.

10. Banga J, Van Dijk J, Van Dis I, Giepmans L, Goudswaard A, Grobbee D. NHG-standaard cardiovasculair risicomanagement (eerste herziening). Huisarts Wet 2012;55:14-28.

11. Kuhn M, Johnson K. Applied predictive modeling. Springer. New York. 2013.

12. Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes. JAMA Netw Open. 2020;3(1):e1918962.

13. Kop R, Hoogendoorn M, Teije AT, Büchner FL, Slottje P, Moons LM, Numans ME. Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records. Comput Biol Med. 2016;76:30-8.

# Bijlage A Resultaten van de random forest

| Aandoening | Meetwaarden | True_pos | False_pos | True_neg | False_neg | Kappa |
|---|---|---|---|---|---|---|
| ap | geen | 12 | 26 | 8 | 318 | 0,1 |
| ap | BMI | 3 | 8 | 58 | 27 | -0,03 |
| ap | diastolische bloeddruk | 23 | 16 | 132 | 54 | 0,22 |
| ap | roken | 7 | 0 | 53 | 16 | 0,38 |
| ap | systolische bloeddruk | 22 | 13 | 134 | 55 | 0,23 |
| | | | | | | |
| aaa | geen | 32 | 38 | 903 | 296 | 0,08 |
| aaa | BMI | 9 | 6 | 59 | 19 | 0,26 |
| aaa | diastolische bloeddruk | 7 | 9 | 137 | 56 | 0,06 |
| aaa | roken | 2 | 2 | 40 | 16 | 0,08 |
| aaa | systolische bloeddruk | 6 | 6 | 142 | 58 | 0,07 |
| | | | | | | |
| ami | geen | 26 | 26 | 1099 | 287 | 0,09 |
| ami | BMI | 2 | 2 | 65 | 26 | 0,06 |
| ami | diastolische bloeddruk | 3 | 9 | 177 | 55 | 0 |
| ami | roken | 4 | 1 | 57 | 15 | 0,26 |
| ami | systolische bloeddruk | 5 | 5 | 182 | 53 | 0,08 |
| | | | | | | |
| ap | geen | 36 | 26 | 860 | 318 | 0,1 |
| ap | BMI | 3 | 8 | 58 | 27 | -0,03 |
| ap | diastolische bloeddruk | 23 | 16 | 132 | 54 | 0,22 |
| ap | roken | 7 | 0 | 53 | 16 | 0,38 |
| ap | systolische bloeddruk | 22 | 13 | 134 | 55 | 0,23 |
| | | | | | | |
| chol | geen | 122 | 116 | 5527 | 1785 | 0,06 |
| chol | BMI | 46 | 46 | 307 | 178 | 0,08 |
| chol | diastolische bloeddruk | 60 | 69 | 827 | 464 | 0,04 |
| chol | roken | 9 | 22 | 242 | 130 | -0,02 |
| chol | systolische bloeddruk | 62 | 61 | 839 | 463 | 0,06 |
| | | | | | | |
| clau | geen | 28 | 35 | 1276 | 401 | 0,05 |
| clau | BMI | 8 | 5 | 72 | 29 | 0,18 |
| clau | diastolische bloeddruk | 4 | 8 | 203 | 63 | 0,03 |
| clau | roken | 7 | 5 | 57 | 26 | 0,15 |
| clau | systolische bloeddruk | 3 | 6 | 206 | 64 | 0,02 |
| | | | | | | |
| cva | geen | 35 | 32 | 889 | 266 | 0,11 |
| cva | BMI | 6 | 14 | 64 | 20 | 0,06 |
| cva | diastolische bloeddruk | 13 | 10 | 145 | 56 | 0,15 |
| cva | roken | 3 | 6 | 54 | 14 | 0,09 |
| cva | systolische bloeddruk | 11 | 9 | 146 | 58 | 0,13 |

| Aandoening | Meetwaarden | True_pos | False_pos | True_neg | False_neg | Kappa |
|---|---|---|---|---|---|---|
| isch hz | geen | 9 | 18 | 288 | 82 | 0,05 |
| isch hz | BMI | 2 | 0 | 17 | 12 | 0,15 |
| isch hz | diastolische bloeddruk | 3 | 3 | 48 | 18 | 0,11 |
| isch hz | roken | 0 | 0 | 18 | 7 | 0 |
| isch hz | systolische bloeddruk | 4 | 3 | 48 | 17 | 0,16 |
| | | | | | | |
| dm | geen | 30 | 34 | 2486 | 712 | 0,04 |
| dm | BMI | 19 | 17 | 143 | 47 | 0,21 |
| dm | diastolische bloeddruk | 11 | 10 | 406 | 104 | 0,1 |
| dm | roken | 7 | 4 | 124 | 36 | 0,17 |
| dm | systolische bloeddruk | 11 | 13 | 406 | 104 | 0,09 |
| | | | | | | |
| hyper | geen | 71 | 101 | 9020 | 2854 | 0,02 |
| hyper | BMI | 15 | 14 | 604 | 241 | 0,05 |
| hyper | diastolische bloeddruk | 103 | 80 | 1433 | 442 | 0,17 |
| hyper | roken | 6 | 6 | 473 | 149 | 0,04 |
| hyper | systolische bloeddruk | 125 | 123 | 1393 | 420 | 0,18 |
| | | | | | | |
| psor | geen | 66 | 51 | 1343 | 382 | 0,15 |
| psor | BMI | 5 | 5 | 91 | 31 | 0,11 |
| psor | diastolische bloeddruk | 15 | 7 | 233 | 78 | 0,17 |
| psor | roken | 3 | 5 | 68 | 23 | 0,06 |
| psor | systolische bloeddruk | 11 | 8 | 232 | 82 | 0,11 |
| | | | | | | |
| reuma | geen | 55 | 32 | 646 | 177 | 0,24 |
| reuma | BMI | 3 | 1 | 43 | 16 | 0,17 |
| reuma | diastolische bloeddruk | 10 | 10 | 100 | 24 | 0,24 |
| reuma | roken | 3 | 0 | 37 | 8 | 0,37 |
| reuma | systolische bloeddruk | 8 | 10 | 101 | 26 | 0,17 |
| | | | | | | |
| tia | geen | 34 | 32 | 810 | 235 | 0,12 |
| tia | BMI | 4 | 2 | 63 | 19 | 0,19 |
| tia | diastolische bloeddruk | 9 | 13 | 121 | 48 | 0,07 |
| tia | roken | 3 | 8 | 44 | 14 | 0,03 |
| tia | systolische bloeddruk | 9 | 12 | 122 | 48 | 0,08 |

hyper = hypertensie, dm = diabetes mellitus, isch hz = Andere/chronische ischemische hartziekte, clau = Claudicatio intermittens, ami = acuut myocardinfarct, aaa = Aneurysma Aortae Abdominalis, ap = agina pectoris cva = Cerebrovasculair accident, tia = transient ischemic attack, psor = psoreasis, reuma = reumatoide artritis

# Bijlage B Voorspellende variabelen van de random forest

Per aandoening zijn de 10 beste voorspellers per uitkomstmaat weergegeven.

*Top 10 beste voorspellers van hypertensie**

| Random forest |
|---|
| leeftijd |
| aantal contacten |
| A02B (Middelen bij ulcus pepticum en gastro-oesofageale reflux) |
| G03A (Hormonale anticonceptiva voor systemisch gebruik) |
| Rd (diagnoses luchtwegen) |
| *A06A (Middelen bij obstipatie)* |
| H02A (Corticosteroiden voor systemisch gebruik) |
| M01A (Niet-steroide anti-inflammatoire en antireumatische midd.) |
| Ds (symptomen maagdarmkanaal) |
| Ls (symptomen bewegingsapparaat) |

* Geneesmiddelen zijn in blauw gepresenteerd voor de herkenbaarheid.

*Top 10 beste voorspellers van diabetes mellitus**

| Random forest |
|---|
| aantal contacten |
| geslacht |
| leeftijd |
| Rd (diagnoses ademhalingsorganen) |
| A02B (Middelen bij ulcus pepticum en gastro-oesofageale reflux) |
| A91 (afwijkende uitslag onderzoek) |
| Rs (symptomen ademhalingsorganen) |
| Td (diagnoses endocriene klieren, metabolisme, voeding) |
| Sd (diagnoses huid/onderhuisweefsel)) |
| Ld (diagnoses bewegingsapparaat) |

* Geneesmiddelen zijn in blauw gepresenteerd voor de herkenbaarheid.

*Top 10 beste voorspellers van Andere/chronische ischemische hartziekte**

| Random forest |
|---|
| geslacht |
| leeftijd |
| aantal contacten |
| Rd (diagnoses ademhalingsorganen) |
| Ds (symptomen maagdarm kanaal) |
| A02B (Middelen bij ulcus pepticum en gastro-oesofageale reflux) |
| Kd (diagnoses hart- vaatstelsel) |
| As (symptomen algemeen) |
| N05C (Hypnotica en sedativa) |
| Dd (diagnoses maagdarmkanaal) |

* Geneesmiddelen zijn in blauw gepresenteerd voor de herkenbaarheid.

*Top 10 beste voorspellers van hypercholesterolemie**

| Random forest |
|---|
| leeftijd |
| aantal contacten |
| A02B (Middelen bij ulcus pepticum en gastro-oesofageale reflux) |
| geslacht |
| *N06A (Antidepressiva)* |
| Td (diagnoses endocriene klieren, metabolisme, voeding) |
| A06A (Middelen bij obstipatie) |
| M01A (Niet-steroide anti-inflammatoire en antireumatische midd.) |
| R06A (Antihistaminica voor systemisch gebruik) |
| N05C (Hypnotica en sedativa) |

* Geneesmiddelen zijn in *blauw* gepresenteerd voor de herkenbaarheid.

*Top 10 beste voorspellers van Claudicatio intermittens**

| Random forest |
|---|
| leeftijd |
| aantal contacten |
| Rd (diagnoses ademhalingsorganen) |
| R95 (Emfyseem/COPD) |
| Ls (symptomen bewegingsapparaat) |
| R03A (Sympathicomimetica voor inhalatie) |
| N05B (Anxiolytica) |
| J01X (Overige antibacteriele middelen) |
| M01A (Niet-steroide anti-inflammatoire en antireumatische midd.) |
| Ps (symptomen psyche) |

* Geneesmiddelen zijn in blauw gepresenteerd voor de herkenbaarheid.

*Top 10 beste voorspellers van acuut myocard infarct**

| Random forest |
|---|
| geslacht |
| leeftijd |
| aantal contacten |
| M01A (Niet-steroide anti-inflammatoire en antireumatische midd.) |
| A02B (Middelen bij ulcus pepticum en gastro-oesofageale reflux) |
| Ld (diagnoses bewegingsapparaat) |
| Rd (diagnoses ademhalingsorganen) |
| J01C (Betalactam-antibiotica penicillines) |
| Sd (diagnoses huid/onderhuidsweefsel) |
| As (symptomen algemeen) |

* Geneesmiddelen zijn in blauw gepresenteerd voor de herkenbaarheid.

*Top 10 beste voorspellers van Aneurysma Aortae Abdominalis**

| Random forest |
|---|
| leeftijd |
| aantal contacten |
| geslacht |
| Rd (diagnoses ademhalingsorganen) |
| As (symptomen algemeen) |
| Ss (symptomen huid/onderhuidsweefsel) |
| A02B (Middelen bij ulcus pepticum en gastro-oesofageale reflux) |
| R03B (Overige middelen bij astma/copd voor inhalatie) |
| Ds (symptomen maagdarmkanaal) |
| Rs (symptomen ademhalingsorganen) |

* Geneesmiddelen zijn in blauw gepresenteerd voor de herkenbaarheid.

*Top 10 beste voorspellers van Angina pectoris**

| Random forest |
|---|
| leeftijd |
| aantal contacten |
| geslacht |
| A02B (Middelen bij ulcus pepticum en gastro-oesofageale reflux) |
| J01C (Betalactam-antibiotica penicillines) |
| M01A (Niet-steroide anti-inflammatoire en antireumatische midd.) |
| Ds (symptomen maagdarmkanaal) |
| Rs (symptomen ademhalingsorganen) |
| Rd (diagnoses ademhalingsorganen) |
| Ld (diagnoses bewegingsapparaat) |

* Geneesmiddelen zijn in blauw gepresenteerd voor de herkenbaarheid.

*Top 10 beste voorspellers van Cerebrale vasculair accident CVA**

| Random forest |
|---|
| leeftijd |
| aantal contacten |
| A02B (Middelen bij ulcus pepticum en gastro-oesofageale reflux) |
| M01A (Niet-steroide anti-inflammatoire en antireumatische midd.) |
| Rd (diagnoses ademhalingsorganen) |
| Ls (symptomen bewegingsapparaat) |
| Sd (diagnoses huid/onderhuidsweefsel) |
| Geslacht |
| *N03A (Anti-epileptica)* |
| Nd (diagnoses zenustelsel) |

* Geneesmiddelen zijn in blauw gepresenteerd voor de herkenbaarheid.

*Top 10 beste voorspellers van TIA**

| Random forest |
|---|
| leeftijd |
| aantal contacten |
| Rd (diagnoses ademhalingsorganen) |
| A02B (Middelen bij ulcus pepticum en gastro-oesofageale reflux) |
| geslacht |
| Ds (symptomen maagdarmkanaal) |
| M01A (Niet-steroide anti-inflammatoire en antireumatische midd.) |
| H03A (Thyreomimetica) |
| N06A (Antidepressiva) |
| Nd (diagnoses zenuwstelsel) |

* Geneesmiddelen zijn in blauw gepresenteerd voor de herkenbaarheid.

*Top 10 beste voorspellers van Psoriasis**

| Random forest |
|---|
| leeftijd |
| aantal contacten |
| D07X (Corticosteroiden met overige middelen) |
| A02B (Middelen bij ulcus pepticum en gastro-oesofageale reflux) |
| Sd (diagnoses huid/onderhuidsweefsel) |
| D01A (Antimycotica lokale) |
| S86 (Seborroïsch eczeem/roos) |
| Rd (diagnoses ademhalingsorganen) |
| As (symptomen algemeen) |

* Geneesmiddelen zijn in blauw gepresenteerd voor de herkenbaarheid.

*Top 10 beste voorspellers van reumatoide artritis**

| Random forest |
|---|
| M01A (Niet-steroide anti-inflammatoire en antireumatische midd.) |
| Ls (symptomen bewegingsapparaat) |
| *L04A (Immunosuppressiva)* |
| L20 (Symptomen meerdere/niet-gespecificeerde gewrichten) |
| Aantal contacten |
| B03B (Vitamine b12 en foliumzuur) |
| A02B (Middelen bij ulcus pepticum en gastro-oesofageale reflux) |
| H02A (Corticosteroiden voor systemisch gebruik) |
| Rd (diagnoses ademhalingsorganen) |
| J01C (Betalactam-antibiotica penicillines) |

* Geneesmiddelen zijn in blauw gepresenteerd voor de herkenbaarheid.