

Samenvatting

Klinische predictiemodellen combineren patiëntgegevens om de kans te voorspellen dat een ziekte aanwezig is (diagnose) of dat een bepaalde ziekte status zich zal voordoen (prognose). De voorspelde kans op de diagnostische of prognostische uitkomst kan een arts helpen bij het maken van een therapiekeuze voor een bepaalde patiënt. Voordat een predictiemodel gebruikt kan worden in de klinische praktijk, dient de prestatie van het model te worden bestudeerd in nieuwe patiënten (externe validiteit). Dit proefschrift beschrijft verschillende theoretische en praktische aspecten van externe validatiestudies voor klinische predictiemodellen. Hiervoor zijn vier doelstellingen gedefinieerd: i) het beschrijven van aspecten die de prestatie van het model bepalen en het beschrijven van relevante maten om de prestatieaspecten te bestuderen; ii) het schatten van de power van de prestatiematen; iii) het extern valideren van een bestaand predictiemodel en iv) het actualiseren van het bestaande predictiemodel.

Drie prestatieaspecten van een klinisch predictiemodel zijn besproken in hoofdstuk 2: calibratie, discriminatie en klinisch nut. Calibratie heeft betrekking op de overeenkomst tussen de geobserveerde uitkomst frequenties en de kansen op de uitkomst zoals voorspeld met het model. Voor een model dat de histologie van restweefsel in gemetastaseerde kiemceltumoren voorspelt betekent dit, dat de frequentie van goedaardig restweefsel overeen dient te komen met de voorspelde kans op goedaardig restweefsel. Een groep van 100 patiënten die allen een voorspelde kans op goedaardig restweefsel van 70% hebben dient te bestaan uit ongeveer 70 patiënten met goedaardig restweefsel en ongeveer 30 patiënten met tumor. Discriminatie heeft betrekking op het vermogen van het model om een patiënt met goedaardig restweefsel te onderscheiden van een patiënt met tumor. In het ideale geval liggen de voorspelde kansen voor goedaardig restweefsel voor patiënten die daadwerkelijk goedaardig restweefsel hebben dicht bij de 100% en voor patiënten met tumor dicht bij 0%. Klinisch nut geeft aan in welke mate het model de arts werkelijk van dienst kan zijn bij het maken van een therapiekeuze zoals het aanbieden van een operatie om restweefsel te verwijderen.

Tevens zijn in hoofdstuk 2 voor ieder prestatieaspect enkele maten beschreven. Calibratie wordt vaak bestudeerd met de Hosmer-Lemeshow statistiek. Deze vergelijkt de geobserveerde uitkomst frequentie met de gemiddelde voorspelde kans per deciel voorspelde kansen. Een figuur die de relatie tussen de geobserveerde frequenties met de voorspelde kansen weergeeft (validatiefiguur), spreekt echter meer tot de verbeelding. De lijn in deze validatiefiguur kan bij perfecte calibratie beschreven worden met een intercept gelijk aan 0 en een helling gelijk aan 1. De waarden van deze parameters kunnen statistisch getoetst

worden (apart of samen). Discriminatie wordt over het algemeen bestudeerd met de concordantie (*c*)-statistic. Voor binaire uitkomsten, zoals ziekte aanwezig ja/nee, komt deze statistic overeen met de oppervlakte onder de ‘receiver-operating-characteristic’ curve die gebruikt wordt voor het evalueren van diagnostische testen. De waarde van de *c*-statistic kan geïnterpreteerd worden als de waarschijnlijkheid dat een aselekt getrokken patiënt met de uitkomst (bijv. goedaardig restweefsel) een hogere voorspelde kans op de uitkomst heeft dan een aselekt getrokken patiënt zonder de uitkomst (bijv. tumor). Relevante maten voor klinisch nut zijn moeilijker te definiëren. Om aan de hand van een voorspelde kans een therapiekeuze te maken is een drempelwaarde nodig. Voor patiënten met een voorspelde kans boven de drempelwaarde zal de therapiekeuze anders zijn dan voor patiënten met een voorspelde kans onder de drempelwaarde. Gegeven de drempelwaarde kan bijvoorbeeld de proportie juiste keuzes (‘accuracy’), de sensitiviteit en de specificiteit berekend worden. Een betere weergave van klinisch nut wordt gegeven door maten die tevens de consequenties van therapiekeuzes meewegen zoals de gewogen proportie foute keuzes (weighed error rate).

De power van een aantal maten om verandering in prestatie van het model aan te tonen in externe validatiestudies is bestudeerd met powerberekeningen en Monte Carlo simulaties in hoofdstuk 3. Een statistische toets die de waarden van het intercept en de helling van de calibratielijn test op afwijking van respectievelijk 0 en 1 had meer power om veranderingen in calibratie op te sporen dan de Hosmer-Lemeshow statistic. De power van de *c*-statistic om veranderingen in discriminatie aan te tonen hing sterk af van de prevalentie van de uitkomst. Dit resultaat laat zien dat het aantal patiënten met de uitkomst (effectief patiëntenaantal) belangrijker is in powerberekeningen dan het totale aantal patiënten. Als vuistregel stellen we voor minimaal 100 patiënten met de uitkomst en 100 patiënten zonder de uitkomst te gebruiken in externe validatiestudies.

De externe validiteit van een model wordt niet alleen bepaald door de nieuwe patiëntengroep, maar ook door het ontwikkelingsproces van het model zoals de selectie van voorspellers, het schatten van de vormen van de associaties en het corrigeren voor te optimistische schattingen. Daarom zijn in dit proefschrift ook enkele illustraties van aspecten van modelontwikkeling opgenomen. In hoofdstuk 4 is een overzicht gegeven van de in de literatuur bekende voorspellers van de aanwezigheid van metastase in patiënten met klinisch stadium I nonseminomateuze kiemceltumoren. Daarnaast is de sterkte van de voorspellers geschat in een meta-analyse. De resultaten kunnen een leidraad vormen voor de selectie van belangrijke voorspellers voor een predictiemodel. Een dergelijk predictiemodel dat het risico op metastase beter schat dan huidige modellen kan de therapiekeuzes voor dit type patiënten verbeteren. De hoog risico patiënten kan direct aanvullende therapie geboden worden, terwijl mogelijk overbodige behandeling van laag risico patiënten kan worden vermeden.

Het ontwikkelen van een predictiemodel voor de histologie van restweefsel in nonseminomateuze kiemceltumoren is beschreven in hoofdstuk 5. Het model voorspelt de

kans dat retroperitoneaal restweefsel geheel goedaardig is na chemotherapie met cis-platina voor gemetastaseerde tumoren. Het ontwikkelen van het model omvatte de selectie van variabelen met behulp van de literatuur (in totaal 996 patiënten), een gedetailleerde schatting van de vormen van de associaties met behulp van individuele patiëntengegevens (n=544, ontwikkelingsset) en correctie voor het optimisme van de geschatte regressiecoëfficiënten met behulp van een *zgn.* ‘shrinkage factor’. Het definitieve model heeft zes voorspellers. De calibratie van het model was, zoals verwacht, goed in de ontwikkelingsset. Het discriminerend vermogen was in de ontwikkelingsset ook goed (c -statistic=0.83, na correctie voor overoptimisme). De externe validiteit van het predictiemodel voor restweefsel was goed in een eerste externe validatieset van 172 patiënten. Het model toonde in deze patiënten adequate calibratie en goede discriminatie (c -statistic=0.82).

Twee andere externe validatiestudies zijn beschreven in de hoofdstukken 6 en 7. De studies maakten gebruik van patiëntengegevens van Indiana University Medical Center (IUMC, n=276) en van een EORTC/MRC trial (n=105). De prevalentie van goedaardig weefsel was lager in deze datasets (28% en 26% respectievelijk) dan in de ontwikkelingsset (45%). Voor de IUMC patiënten is een alternatief model ontwikkeld, omdat van één voorspeller veel waarden ontbraken en omdat voor twee andere voorspellers slechts categorische waarden bekend waren, terwijl het model continue waarden gebruikt. De lage prevalentie voor goedaardig weefsel kon voor een deel verklaard worden door de voorspellers in het model. De patiënten van IUMC hadden bijvoorbeeld vaker dan de patiënten van de ontwikkelingsset een grote restmassa dat kan duiden op residuele tumor. Na correctie voor de voorspellers in het model was de prevalentie goedaardig restweefsel nog steeds lager dan in de ontwikkelingsset, hoewel niet meer statistisch significant. Dit verschil moet worden toegeschreven aan patiënteneigenschappen die niet in het model zijn opgenomen, gerelateerd bijvoorbeeld aan een ander verwijspatroon voor de patiënten van IUMC. Het alternatieve model kon vrij goed onderscheid maken tussen patiënten met goedaardig weefsel en patiënten met tumor (c -statistic=0.79).

De lage prevalentie goedaardig weefsel voor de patiënten van de EORTC/MRC trial (26%) was opmerkelijk, omdat deze patiënten tot de categorie ‘goede prognose patiënten’ behoorden. De patiënten die geselecteerd waren voor operatie waren echter een klein deel van alle patiënten met goede prognose. Veel van de patiënten hadden kleine, vaak goedaardige, restmassa’s die niet verwijderd zijn. De patiënten die wel zijn geopereerd hadden daarom relatief vaak tumor. De lage prevalentie voor goedaardig restweefsel kon ook geheel verklaard worden door de voorspellers in het model. Dit betekent dat de gemiddelde geobserveerde frequentie van goedaardig restweefsel overeen kwam met de gemiddelde voorspelde kans op goedaardig restweefsel. Lage voorspelde kansen waren echter niet in overeenstemming met de geobserveerde frequenties. Het model zou toch behulpzaam kunnen zijn bij de therapiekeuze, omdat de drempelwaarde voor niet opereren versus wel opereren rond de 70% ligt en in dit gebied de voorspelde kansen wel correct waren. De verdeling van de voorspelde kansen liet echter zien dat weinig kansen hoger dan 70% waren. Veel therapiekeuzes gebaseerd op het predictiemodel zouden dus hetzelfde zijn

geweest als de gemaakte keuzes, namelijk opereren. Het model had daarom voor deze patiënten weinig toegevoegde waarde.

Alle beschikbare individuele patiëntengegevens (de ontwikkelingsset en de 3 validatiesets) zijn in hoofdstuk 8 samengevoegd (n=1094) om het model voor restweefsel te actualiseren. De regressiecoëfficiënten van dit actuele model verschilden slechts in geringe mate van de originele coëfficiënten. De coëfficiënten waren wel preciezer door het grote aantal patiënten. De prestatie van het actuele model was goed in vier van zes patiëntengroepen. De lage voorspelde kansen voor de patiënten van de EORTC/MRC trial waren nog steeds slecht gecalibreerd, terwijl de voorspelde kansen voor patiënten van een gespecialiseerd Amerikaans centrum systematisch te hoog waren. Voor deze patiënten kan overwogen worden om het intercept van het model aan te passen.

Dit proefschrift eindigt met een algemene discussie (hoofdstuk 9). Hierin zijn een aantal praktische aanbevelingen voor externe validatiestudies gedaan. In een externe validatiestudie dienen adequate patiëntengegevens te worden verzameld, retrospectief of prospectief. De definities van de voorspellers dienen overeen te komen met de definities van de ontwikkelingsset en missende waarden dienen zoveel mogelijk beperkt te worden. Om voldoende power te behalen in een validatiestudie dienen veel patiënten met de uitkomst en veel patiënten zonder de uitkomst verzameld te worden. Een essentieel element in de data analyse is een validatiefiguur die de associatie tussen de geobserveerde frequenties en de voorspelde kansen weergeeft. De figuur kan snel inzicht geven in alle relevante prestatieaspecten van het model.

