Intensionele Logica: Kinderziekten met vlekjes

Mark Beumer (0477672)
mbeumer@science.uva.nl
11/04/2006

Inleiding

In het kader van het vak Intensionele Logica 05-06 dient een Bayesiaans netwerk te worden gebouwd voor het domein 'kinderziekten met vlekjes'. Dit domein behelst kinderziekten zoals de mazelen, waterpokken en rode hond: aandoeningen met als gemene deler het veroorzaken van diverse soorten vlekjes op de kinderhuid. Dit rapport is een verslag van het construeren van dit netwerk met het programma Hugin Lite[1].

De aanleiding voor het bouwen van zo'n netwerk wordt gevonden in een notoir probleem bij dergelijke ziekten: in de praktijk blijkt het lastig de werkelijke boosdoener vast te stellen. Zelfs als men de beschikking heeft over uitgebreide mogelijkheden voor laboratorisch onderzoek kan in slechts 65% van de gevallen een juiste diagnose worden gesteld [Win96]. Het huidige netwerk kan worden gezien als het voorzichtig aftasten van de mogelijkheden van Bayesiaanse netwerken op dit domein. De nadruk ligt hierbij op voorzichtig: de beschikbare informatie ([Win96, GGD03]) was beperkt en vooral kwalitatief van aard. Dientengevolge zijn er gaandeweg een aantal simplificerende aannames gedaan. Onder meer in het laatste hoofdstuk zullen deze als aanwijzingen voor verdere ontwikkeling worden behandeld.

In dit rapport wordt enige voorkennis van Bayesiaanse netwerken en de daarmee geassocieerde *conditional probability tables* (CPTs) bekend verondersteld. Een gedegen omschrijving kan gevonden worden in [RNo03].

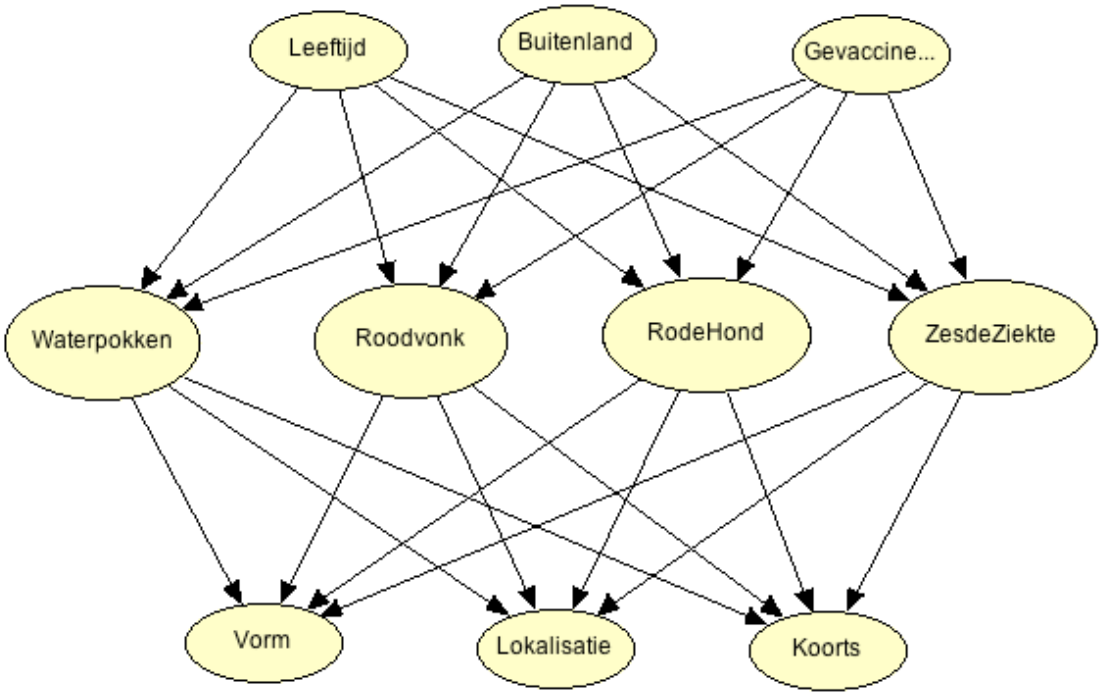
Het verslag is als volgt opgebouwd: in het eerste hoofdstuk wordt het geconstrueerde netwerk geïntroduceerd tezamen met de belangrijkste overwegingen die in dit proces gemaakt zijn. In het daaropvolgende hoofdstuk staat een uitgebreide beschrijving van de 'noisy-OR': een generalisatie van de meer bekende logische OR. Het gebruik van de noisy-OR vermindert de hoeveelheid te schatten waarschijnlijkheden aanzienlijk. Tenslotte volgt een hoofdstuk met enige aanwijzingen voor verbeteringen van het netwerk.

[1] http://www.hugin.com/Products_Services/Products/Demo/Lite/

Belief Network & CPTs

Het 'Belief Network'

'Beginnen bij het begin' is vaak raadzaam, maar toch zal eerst het eindproduct van het proces worden geïntroduceerd: het gebouwde netwerk. Het toelichten van de belangrijkste overwegingen is namelijk duidelijker aan de hand van een concreet gegeven:



De geassocieerde CPTs zijn te vinden in de appendix.

Belangrijkste overwegingen

Beperkt

Ten eerste valt op dat het netwerk vrij bescheiden van omvang is: een bewuste beperking van het gegeven domein. Deze keuze is ingegeven door de snel toenemende gecompliceerdheid van de CPTs. Om een meer waarheidsgetrouw model te kunnen bouwen is het noodzakelijk meer ziektes op te nemen (e.g., de vijfde ziekte, mazelen, Ziekte van Kawasaki), meer symptomen te beschouwen (e.g., verspreiding van de vlekjes, opzwellen van lymfeklieren, aanwezigheid van een frambozentong), alsmede meer predisposities te overwegen (e.g., seizoensinvloed, contact met andere patiënten i.v.m. incubatietijd van de ziekte).

Voor het moment is het netwerk echter groot genoeg om de belangrijkste ideeën in deze context over te brengen. Daarnaast is het verzinnen van meer kansen weinig zinvol, te meer gezien de beperkte informatie die voor dit project ter beschikking stond. De betreffende informatie was bovendien nogal kwalitatief van aard: voor een gerede invulling van de CPTs zijn meer harde cijfers nodig.

Knopen

Een korte omschrijving van de knopen is op zijn plaats. Aangezien ze allen discreet zijn treden er een aantal problemen op.

In de bovenste 'laag' vinden we de knopen 'Leeftijd', 'Buitenland' en 'Gevaccineerd'. De eerste geeft de leeftijd van het kind weer en bestaat uit drie *states*: 0-2 jaar, 2-6 jaar en 6-8 jaar. Deze keuze is afdoende voor het huidige, beperkte domein. Er is namelijk alleen informatie beschikbaar in de trant van "de ziekte komt veel voor bij zuigelingen" (zesde ziekte) of "vooral bij oudere kinderen" (rode hond). In het model corresponderen deze uitspraken met respectievelijk de intervallen 0-2 en 6-8.

Deze keuze stipt direct een belangrijk punt aan: in werkelijkheid is een discretisatie van een dergelijke grove granulariteit niet betrouwbaar. Het gebruiken van een normale verdeling ligt meer voor de hand, of als men wél wil discretiseren dan toch in ieder geval met een kleinere stapgrootte. Het is bijvoorbeeld mogelijk dat een kind van 4 of 5 reeds rode hond krijgt, hoewel deze kans misschien klein is. Deze mogelijkheid is in het huidige model echter volledig genegeerd, i.e. krijgt een (impliciete) waarschijnlijkheid van 0.

In het algemeen kan gesteld worden dat het huidige domein een aantal continue variabelen bevat die zich slecht lenen voor discretisatie. Beschouw bijvoorbeeld 'Vorm' in de onderste laag, welke de vorm van de vlekjes beschrijft. Het blijkt dat deze niet zomaar gecategoriseerd kunnen worden, aangezien ze zich in vele overgangsvormen manifesteren. Dit verschijnsel kan beter gemodelleerd worden met een groot aantal staten (een staat voor elke vorm) of met een bepaalde kansverdeling (normaal, binomiaal). Voorlopig voldoet de grove discretisatie echter.

Om terug te komen op de beschrijving van knopen: de *random variable* 'Buitenland' geeft aan of een kind recentelijk een bezoek aan het buitenland heeft gebracht. Dit is van belang voor de ziekte rode hond, welke in Nederland niet meer voorkomt. Als een kind niet gevaccineerd is én kort geleden naar een land op vakantie is geweest waar de ziekte nog wél voorkomt, dan is er een kans dat het kind alsnog rode hond krijgt.

De binaire knoop 'Gevaccineerd' geeft als laatste aan of een kind reeds gevaccineerd is of niet. Voor het gemak zijn verschillende vaccinaties op één hoop gegooid.

In de tweede laag bevinden zich de ziektes zelf. Dit zijn binaire knopen: óf een kind heeft de ziekte, óf een kind heeft de ziekte niet. In praktijk zijn er uiteraard mildere (tussen)vormen van elke ziekte, bijvoorbeeld als een kind de ziekte al eens heeft gehad. In het huidige netwerk wordt aangenomen dat dit niet het geval is, wederom voor simpliciteit.

In de onderste laag staan tenslotte 'Vorm', 'Lokalisatie' en 'Koorts'. De eerste slaat op de vorm van de vlekjes die optreden bij het kind. Zoals reeds opgemerkt is zijn er vele overgangsvormen. Hier heeft de knoop echter maar vier staten:

1) vesiculeus / hemorragisch
2) maculopapuleus
3) maculopapuleus / hemorragisch
4) maculeus / maculopapuleus

Voor het huidige domein is dit afdoende. 'Lokalisatie' is de plaats waar de vlekjes optreden. In de praktijk zal dit niet zo eenduidig zijn als de volgende verdeling doet vermoeden:

1) gehele lichaam
2) oksels / liezen
3) gezicht / nek
4) extremiteiten (i.e., armen en benen)
5) romp

Koorts is tenslotte onderverdeeld in 'geen', 'laag' en 'hoog'.

Causaal

Het netwerk is in drie lagen verdeeld, waarbij de causale relaties tussen de lagen is gehandhaafd. Van boven naar beneden:

1. Predisposities
2. Ziekten
3. Symptomen

In de eerste laag vinden we de zogeheten *predisposities*: deze factoren kunnen de kans op een bepaalde ziekte verhogen of verlagen. Zo komt de zesde ziekte bijna uitsluitend voor bij zuigelingen. Een jonge leeftijd predisponeert (vergroot de kans) dus voor het krijgen van deze ziekte.

In de tweede laag staan de ziektes, welke op hun beurt de oorzaak zijn van vele symptomen in de derde laag. Er zijn meer relevante symptomen dan hier genoemd worden: wederom is voor de eenvoud een groot deel weggelaten.

De causale ordening van het netwerk biedt als grote voordeel een intuïtief duidelijke structuur. Daarbij zorgt de ordening er voor dat het netwerk in één van zijn meest simpele verschijningsvormen staat.

Verbindingen

Twee dingen zijn interessant om op te merken wat betreft de verbindingen tússen de knopen. Ten eerste zijn er geen links tussen de knopen binnen een laag. Daarnaast zijn de lagen onderling *fully connected*: elke knoop is met elke andere knoop verbonden.

Wat betreft de laatste waarneming: het is bijvoorbeeld opvallend dat een ziekte als rode hond verbonden is met de knoop Koorts. Doorgaans heeft rode hond geen koorts tot gevolg. Merk echter op dat juist de aanwezigheid van koorts dan een belangrijke aanwijzing is omtrent de waarschijnlijkheid van rode hond en een link dus essentieel is. Overigens wordt deze invloed ook indirect duidelijk door het waaschijnlijker worden van ziektes die wél koorts veroorzaken.

Verder zijn er geen links tussen de knopen binnen een laag. Gezien de omschrijvingen van de huidige knopen is het aannemelijk dat deze factoren elkaar niet onderling beïnvloeden. Eén extra aanname die tijdens het construeren van de CPTs is gedaan mag echter niet onvermeld blijven: er wordt van uitgegaan dat slechts één ziekte tegelijkertijd optreedt. Het waarnemen van een ziekte zou op deze manier direct gevolg moeten hebben op de waarschijnlijkheid van andere ziektes. Het verwerken van deze aanname zou echter onvermijdelijk *cycles* tot gevolg hebben: dit is per definitie niet toegestaan in Bayesiaanse netwerken. Daarnaast kan het mogelijk zijn dat een kind met waterpokken door een verlaagde weerstand ook vatbaarder is voor andere ziektes. Hier is echter aanvullende informatie nodig.

CPTs en noisy-OR

De met het netwerk geassocieerde CPTs zijn te vinden in de Appendix. Bij het invullen van de tabellen zijn een aantal zaken van belang geweest.

Om te beginnen is er gebruik gemaakt van de zogeheten 'noisy-OR'. De assumpties die dit model maakt verminderen het aantal *a priori* benodigde kansen aanzienlijk. Een uitgebreide beschrijving van de toepassing in het netwerk zal in het volgende hoofdstuk worden gegeven.

Ook is er de vraag of uitgegaan moet worden van een *Closed World Assumption* (CWA). Als geen van de ziektes in ons netwerk het geval is, kan er dan bijvoorbeeld sprake van koorts zijn? Uitgaande van een CWA is dit niet mogelijk; dit lijkt echter een wel zeer vergaande simplificatie te zijn.

Bij het vinden van een oplossing is het nu gewenst dat niet direct alle mogelijke oorzaken van koorts beschouwd hoeven worden. Een idee is dan het toevoegen van een speciale staat die dergelijke onvoorziene omstandigheden op één hoop gooit, de zogeheten *leak node*. De waarschijnlijkheid van deze oorzaak kan willekeurig klein worden gemaakt, min of meer naar gelang het gewenste gedrag van het netwerk.

In het huidige model is deze knoop niet expliciet toegevoegd, hoewel er geen sprake is van een CWA. In de hierboven geschetste situatie zou de waarschijnlijkheid van een koorts een kleine, niet-0 waarschijnlijkheid krijgen.

‘Noisy-OR’: Consistent schatten

Generalisatie

Een potentieel probleem bij het construeren van een Bayesiaans netwerk is het invullen van de CPTs. Als elke knoop in het netwerk maximaal *k* binaire ouders heeft, kan het voorkomen dat voor elke CPT 2^k verschillende getallen geschat moeten worden: het conditioneren van het netwerk vereist aldus exponentieel veel data.

In werkelijkheid is dit *worst-case scenario* vaak te vermijden, aangezien de relaties binnen een netwerk vaak te beschrijven zijn met een standaard canonische distributie [RNo03]. In een dergelijk geval kan een CPT worden gespecificeerd met het standaardpatroon en enkele additionele parameters.

Relaties met onzekerheid kunnen aldus worden beschreven met de ‘noisy-OR’, een generalisatie van de standaard logische OR. Dit model staat onzekerheid toe in het vermogen van elke ouder om het kind waar te laten zijn: in plaats van ‘true’ of ‘false’ is er sprake van ‘inhiberen’ (minder waarschijnlijk doen zijn) en ‘exciteren’ (waarschijnlijker laten worden).

De noisy-OR gaat uit van twee aannames:

1) Alle mogelijke oorzaken van een knoop zijn bekend; eventueel wordt een ‘leak node’ gespecificeerd.
2) De inhibitie van elke ouder is onafhankelijk van de inhibitie van andere ouders (e.g., als waterpokken koorts inhibeert dan is deze oorzaak onafhankelijk van datgene waarom rode hond koorts inhibeert)

Met name de tweede aanname van onafhankelijkheid is interessant. Gegeven deze aanname is koorts alleen ‘niet waar’ als al zijn ouders niet waar zijn. De kans op geen koorts is dan ook het product van de inhiberende waarschijnlijkheden van de ouders. Dit verkleint het aantal te schatten kansen.

De uitwerking van het model kan het beste worden uitgelegd aan de hand van een voorbeeld.

Voorbeeld: Rode hond

In het netwerk zien we dat drie ouders van belang zijn:

1) Gevaccineerd (wel / niet)
2) Buitenland (recent_bezoek / geen_bezoek)
3) Leeftijd (0-2 / 2-6 / 6-8)

De literatuur vermeldt dat rode hond in Nederland vrijwel uitgebannen is[2]: het wordt slechts bij oudere kinderen zónder BMR-vaccinatie waargenomen. Daarbij moet het kind

[2] http://www.blutner.de/Intension/kinderziekten_met_vlekjes/kinderziekten_met_vlekjes.htm.

recent in een land zijn geweest waar de ziekte nog veelvuldig voorkomt. Er is dus sprake van drie inhiberende waardes voor de variabelen:

1) Gevaccineerd = wel
2) Buitenland = geen_bezoek
3) Leeftijd = 0-6

Merk op dat van 'Leeftijd' een binaire knoop is gemaakt: óf je bent een ouder kind in de leeftijd 6 tot 8, óf je bent het niet en zit in de categorie 0 tot 6. Deze opdeling in twee categoriën is noodzakelijk voor het gebruik van de noisy-OR.

Voor het invullen van de volledige CPT van 'RodeHond' is het nu voldoende om met een beetje *Fingerspitzegefühl* de volgende drie kansen te schatten:

μ(RodeHond = nee | Gevaccineerd = niet, Buitenland = geen_bezoek, Leeftijd = 0-6) = 0.1
μ(RodeHond = nee | Gevaccineerd = wel, Buitenland = recent_bezoek, Leeftijd = 0-6) = 0.01
μ(RodeHond = nee | Gevaccineerd = wel, Buitenland = geen_bezoek, Leeftijd = 6-8) = 0.1

Hierin bevat elke waarschijnlijkheid steeds twee inhiberende ouders. Dit levert via de noisy-OR de volgende CPT op:

Gevaccineerd	Buitenland	Leeftijd	RodeHond = ja	RodeHond = nee
niet	recent_bezoek	6-8	0.1981	$0.9 \times 0.99 \times 0.9 = 0.8019$
niet	recent_bezoek	0-6	0.109	$0.9 \times 0.99 = 0.891$
niet	geen_bezoek	6-8	0.19	$0.9 \times 0.9 = 0.81$
niet	**geen_bezoek**	**0-6**	**0.1**	**0.9**
wel	recent_bezoek	6-8	0.109	$0.9 \times 0.99 = 0.891$
wel	**recent_bezoek**	**0-6**	**0.01**	**0.99**
wel	**geen_bezoek**	**6-8**	**0.1**	**.9**
wel	geen_bezoek	0-6	0.001	0.999

Vetgedrukt staan de aangenomen kansen. De bovenste tabel-regel kan nu als volgt worden berekend:

μ(RodeHond = ja | Gevaccineerd = niet, Buitenland = recent_bezoek, Leeftijd = 6-8)

= 1 - μ(RodeHond = nee | Gevaccineerd = niet, Buitenland = recent_bezoek, Leeftijd = 6-8)

= 1 - μ(RodeHond = nee | Gevaccineerd = niet, Buitenland = geen_bezoek, Leeftijd = 0-6) ×
μ(RodeHond = nee | Gevaccineerd = wel, Buitenland = recent_bezoek, Leeftijd = 0-6) ×
μ(RodeHond = nee | Gevaccineerd = wel, Buitenland = geen_bezoek, Leeftijd = 6-8)

= 1 - $0.9 \times 0.99 \times 0.9$
= 1 - 0.8019
= 0.1981

De overige kansen uit de CPT volgen een soortgelijke redenatie. Merk tenslotte op dat er geen Closed World Assumption is in de laatste regel, maar een impliciete 'leak node':

μ(RodeHond = ja | Gevaccineerd = wel, Buitenland = geen_bezoek, Leeftijd = 0-6) = $0.001 \neq 0$

Discussie & Conclusie

Uit simpele experimenten met het netwerk blijkt dat het geven van bewijs inderdaad de ziekte oplevert die men verwacht aan de hand van de ingevulde CPTs. Hoe veel waarde hier aan gehecht kan worden is onzeker: het schatten van de waarschijnlijkheden is op weinig meer dan gebakken lucht gebaseerd.

Er zijn vele andere aanmerkingen te maken op de betrouwbaarheid van het netwerk met betrekking tot het modelleren van de werkelijkheid. De belangrijkste aannames zijn reeds genoemd en worden hier nog eens op een rijtje gezet:

1) Het netwerk is zeer beperkt. Een volgend netwerk zou uitgebreid moeten worden met meer predisposities, ziekten en symptomen. Hiervoor is meer informatie nodig, zowel kwalititatief (causale verbanden tussen variabelen) als kwantitatief (harde cijfers).
2) Een belangrijk punt is het discretiseren van continue variabelen zoals de leeftijd van de patiënt en de vorm van de vlekjes. De stapgrootte waarmee nu een opdeling is gemaakt zal in werkelijkheid niet acceptabel zijn. Een oplossing kan gevonden worden in het toevoegen van meer discrete staten of het definiëren van een standaard kansverdeling over het bereik van de variabele. In beide gevallen is het van essentieel belang de bruikbaarheid van de noisy-OR in de gaten te houden.
3) De in dit netwerk niet expliciet opgenomen 'leak node' verdient nadere studie. In principe kan zo'n knoop in elke laag worden opgenomen. Het effectief implementeren vereist ook hier meer medische kennis.
4) Tenslotte is de aanname of er slechts één ziekte tegelijk het geval kan zijn twijfelachtig. De CPTs kunnen op basis van nadere informatie echter gemakkelijk worden aangepast om een dergelijk geval te accomoderen.

Concluderend valt te stellen dat, hoewel er een aantal simplificerende aannames zijn gedaan, het Bayesiaanse netwerk een veelbelovend formalisme is voor het beschrijven van kinderziekten met vlekjes. De structuur van het domein kan intuïtief worden gevat in een 3-lagenstructuur met een causale ordening tussen de knopen. Het gebruik van de noisy-OR vergemakkelijkt het leven, maar zal aangepast moeten worden voor een uitgebreidere omschrijving van het domein. Ondanks de vele noodzakelijke assumpties zit er zeker toekomst in deze aanpak.

Literatuur

RNo03 Russel, S. & Norvig P., 2003. Artificial Intelligence: A Modern Approach. Prentice Hall, 2nd edition

Hal03 Halpern, J. Y., 2003. Reasoning about Uncertainty. MIT Press.

Win96 Winterberg, D.H., 1996. Exanthemen bij kinderen. Ned. Tijdschrift Geneeskunde, 20 juli, 140(29).

GGD03 Div. auteurs, 2003. Richtlijn voor GGD bij melding exantheem. Landelijke Coördinatiestructuur Infectieziektebestrijding.

Appendix: alle CPTs

	Leeftijd
0-2	0.25
2-6	0.5
6-8	0.25

	Buitenland
recent_bezoek	0.2

	Gevaccineerd
wel	0.95

Gevaccineerd	**Waterpokken**
wel	0.01
niet	0.1

- De variabele Buitenland is alleen relevant voor rode hond
- waterpokken komen voor bij kinderen van alle leeftijden

Gevaccineerd	**Roodvonk**
wel	0.01
niet	0.1

- De variabele Buitenland is alleen relevant voor rode hond
- roodvonk komt voor bij kinderen van alle leeftijden

Gevaccineerd	**Buitenland**	**Leeftijd**	**RodeHond = ja**
niet	recent_bezoek	6-8	0.1981
niet	recent_bezoek	0-6	0.109
niet	geen_bezoek	6-8	0.19
niet	geen_bezoek	0-6	0.1
wel	recent_bezoek	6-8	0.109
wel	recent_bezoek	0-6	0.01
wel	geen_bezoek	6-8	0.1
wel	geen_bezoek	0-6	0.001

- Rode hond komt bijna uitsluitend voor bij niet-gevaccineerde kinderen in de leeftijd 6-8 die een recent bezoek aan het buitenland hebben gebracht.

Gevaccineerd	**Leeftijd**	**ZesdeZiekte**
Niet	0-2	0.19
Niet	2-8	0.1
Wel	0-2	0.1
Wel	2-8	0.001

- De zesde ziekte komt met name voor bij ongevaccineerde zuigelingen.

Water-Pokken	Roodvonk	RodeHond	Zesde ziekte	Koorts = none	Koorts = low	Koorts = high
yes	yes	yes	yes	1	1	1
yes	yes	yes	no	1	1	1
yes	yes	no	yes	1	1	1
yes	yes	no	no	1	1	1
yes	no	yes	yes	1	1	1
yes	no	yes	no	1	1	1
yes	no	no	yes	1	1	1
yes	no	no	no	0.1	0.8	0.1
no	yes	yes	yes	1	1	1
no	yes	yes	no	1	1	1
no	yes	no	yes	1	1	1
no	yes	no	no	0.1	0.8	0.1
no	no	yes	yes	1	1	1
no	no	yes	no	0.8	0.1	0.1
no	no	no	yes	0.1	0.1	0.8
no	no	no	no	1	1	1

- De CPTs voor 'Vorm' en 'Lokalisatie' zijn volgens dezelfde principes opgebouwd en vanwege omvang hier weggelaten.
- Let op de aanname dat er slechts één ziekte tegelijk kan optreden: regels met meer dan één 'yes' bevatten alleen 1'en. Het geven van een 0-waarschijnlijkheid aan zo'n regel is door Hugin niet toegestaan. Het definiëren van een uniforme distributie sorteert echter ook het gewenste effect.
- De impliciete 'leak node' is weggelaten, aangezien het niet duidelijk is uit de literatuur welke mogelijkheid (geen, lage of hoge koorts) de kleine niet-0 waarschijnlijkheid zou moeten krijgen.