DIAGNOSTISCHE TESTEN: VALIDATIE, INTERPRETATIE EN DE GEVOLGEN OP DE BESLUITVORMING

D. Verloo[1], J. Dewulf[2], D. Maes[2], K. Mintiens[1], H. Laevens[1], F. Boelaert[3]

[1] Coördinatiecentrum voor Diergeneeskundige Diagnostiek,
Centrum voor Onderzoek in de Diergeneeskunde en Agrochemie,
Groeselenberg 99, B-1180 Brussel
[2] Vakgroep Voortplanting, Verloskunde en Bedrijfsdiergeneeskunde,
Afdeling voor Veterinaire Epidemiologie,
Faculteit Diergeneeskunde,Universiteit Gent, Salisburylaan 133, B-9820 Merelbeke
[3] Europese Autoriteit voor Voedselveiligheid (European Food Safety Authority),
Palazzo Ducale, Parco Ducale 3, I-43100 Parma, Italië

SAMENVATTING

Het inschatten van en het rekening houden met de precisie en de accuraatheid van diagnostische testen zijn van groot belang voor de interpretatie van testresultaten en hebben een invloed op zeer verscheiden onderdelen van de diergeneeskunde. Praktijkdierenartsen dienen dit in acht te nemen bij het stellen van hun diagnosen; in bewakingsprogramma's moet de steekproefgrootte aangepast zijn, terwijl epidemiologen deze parameters nodig hebben om de resultaten van hun studie te corrigeren, en risicoanalysten deze gebruiken om hun kansmodellen op te stellen. In het eerste deel van dit artikel wordt een inleiding gegeven tot de schatting van de prevalentie en de interpretatie ervan en worden de precisieparameters herhaalbaarheid en reproduceerbaarheid toegelicht. Vervolgens worden de diagnostische accuraatheidsparameters, sensitiviteit en specificiteit, verklaard uitgaande van de aanwezigheid van een gouden standaard. Daaropvolgend beschrijven we de relatie tussen de diagnostische accuraatheid en de gebruikte afkapwaarden, en wordt dit gevisualiseerd door Receiver Operator Characteristics (ROC) curves. Finaal wordt het begrip voorspellende waarde besproken en bespreken we een methode om de populatieprevalentie te schatten bij gebruik van imperfecte testen.

INLEIDING

Elke handeling die de onzekerheid rond de "ziektestatus" van een individu of een groep individuen doet afnemen, kunnen we beschouwen als een diagnostische test. Ziektestatus dient echter niet te eng geïnterpreteerd te worden. Ook het bepalen of een dier al dan niet drachtig is of het meten van de bacteriële contaminatie van een karkas kan beschouwd worden als een diagnostische test. In dit artikel zullen we ons echter toespitsen op diagnostische testen die informatie geven of een individueel dier al dan niet ziek is. Grosso modo kan men deze testen opdelen in testen die het pathogeen agens zelf detecteren, testen die delen van of metabolieten van het agens detecteren, testen die immunologische reacties van de gastheer op het agens of metabolieten ervan detecteren en testen die de aanwezigheid van delen van het genetisch materiaal van het agens detecteren.

De prestatie van een diagnostische test wordt uitgedrukt door de twee onafhankelijke parameters precisie en accuraatheid. Precisie slaat op de overeenkomst tussen herhaalde metingen van een test op hetzelfde staal onder welbepaalde condities (Standard Operating Procedures of SOP's), terwijl accuraatheid de overeenkomst kwantificeert tussen het testresultaat en de echte waarde van hetgeen wordt gemeten. Daar we bij diagnostische testen vooral geïnteresseerd zijn in de ziektestatus van het dier, is het belangrijk informatie te hebben over de diagnostische accuraatheid. Het schatten van deze parameters hoort bij het validatieproces van een diagnostische test.

PREVALENTIE

De prevalentie π is de proportie zieke dieren in de populatie of de kans dat een dier dat at random (aselect, willekeurig: men laat de selectie volledig

aan het toeval over) uit de populatie geselecteerd wordt, de ziekte heeft (=Pr($D+$)).

$$\pi = Pr(D+) = \frac{\textit{aantal zieke dieren in de populatie}}{\textit{totaal aantal zieke dieren in de populatie}}$$

Het spreekt voor zich dat men hiervoor aanneemt dat er een perfecte test (gouden standaard) gebruikt wordt om het aantal zieke dieren in de populatie te bepalen. Later zullen we zien hoe men dit kan corrigeren voor imperfecte testen.

Onder reële omstandigheden is het meestal niet mogelijk een volledige populatie te testen en wordt een aselecte steekproef met steekproefgrootte n genomen om de prevalentie te schatten. Op basis van de proportie zieke dieren in de steekproef kan men een schatting maken van de proportie zieke dieren in de populatie. Door een 'hoedje' te plaatsen boven de parameter $\hat{\pi}$ geven we aan dat het om een schatting gaat van de populatieparameter π.

$$\hat{\pi} = \frac{\textit{aantal zieke dieren in de steekproef}}{n}$$

Aangezien een steekproef slechts bestaat uit een fractie van de populatie zal een nieuwe steekproef (i.e. opnieuw een aantal dieren aselect uit de populatie halen en de ziektestatus bepalen) meestal een ander resultaat opleveren. De onzekerheid die men heeft door slechts een deel van de populatie te onderzoeken en niet de gehele populatie, wordt weergegeven door het betrouwbaarheidsinterval rond de schatting (een maat voor de precisie van de schatting).

Een prevalentieschatting van 10% met een 95% betrouwbaarheidsinterval van 5-15% geeft weer dat, mocht men 100 maal een steekproef nemen, en voor elk van die steekproeven een betrouwbaarheidsinterval voor de prevalentie berekenen, de echte populatieprevalentie gemiddeld in 95 gevallen van de 100 binnen het berekende betrouwbaarheidsinterval zou liggen. Men kan dus stellen dat de kans gelijk is aan 0,95 dat een 95% betrouwbaarheidsinterval de populatieprevalentie bevat.

Er zijn verschillende manieren om betrouwbaarheidsintervallen rond de prevalentie te berekenen. De eenvoudigste manier (en de enige manier die we hier bespreken) is de klassieke manier. Men neemt hier aan dat de selectie van een bepaald dier volledig onafhankelijk is van de selectie van een ander dier en dat de steekproefgrootte niet meer bedraagt dan 10% van de populatiegrootte (indien het aantal dieren in de steekproef groter is dan 10% van de populatie kan men niet meer aannemen dat de kans om een ziek (of niet-ziek) dier te selecteren gelijk blijft voor elk geselecteerd dier).

Het 95% betrouwbaarheidsinterval (BI) wordt gegeven door:

$$95\%\ BI = \hat{\pi} \pm 1{,}96\sqrt{\hat{\pi}\left(1-\hat{\pi}\right)/n}$$

Het is duidelijk dat een betrouwbaarheidsinterval het best zo nauw mogelijk is, omdat dit de onzekerheid rond de prevalentieschatter $\hat{\pi}$ verkleint. Dit kan men bekomen door de steekproefgrootte te laten toenemen (maar zoals hoger vermeld, onder deze assumpties, mag de steekproefgrootte niet groter worden dan 10% van de populatiegrootte).

PRECISIE (HERHAALBAARHEID EN REPRODUCEERBAARHEID) EN ANALYTISCHE ACCURAATHEID

Indien een diagnostische test herhaald wordt op hetzelfde staal dienen de resultaten zoveel mogelijk gelijk te blijven. Hoe groter de eigenschap van een test is om steeds hetzelfde resultaat te reproduceren, hoe hoger de precisie van een test is. Precisie is een maat van overeenkomst tussen verschillende testuitslagen op hetzelfde staal zonder echter rekening te houden met wat het resultaat eigenlijk moet zijn gegeven de informatie die we hebben over de geteste stalen.

De accuraatheid handelt over het verschil tussen de waarde bekomen in de test en de echte waarde. Voor diagnostische testen kunnen we onderscheid maken tussen analytische en diagnostische accuraatheid.

Indien men de eigenschap van een test wil nagaan om correct een staal van een ziek (niet-ziek) dier als positief (negatief) te classificeren, spreken we van diagnostische accuraatheid (zie verder).

Precisie kan, afhankelijk van de omstandigheden waarin men het hertesten uitvoert, onderverdeeld worden in herhaalbaarheid en reproduceerbaarheid.

Herhaalbaarheid is een maat voor de overeenkomst van herhaalde metingen op een aantal stalen door dezelfde persoon (of hetzelfde laboratorium), terwijl reproduceerbaarheid een maat van overeenkomst is van metingen door verschillende personen (of laboratoria) op dezelfde stalen. Herhaalbaarheid kan gemeten worden tijdens één uitvoering van een test door bijvoorbeeld een ELISA-("Enzyme Linked Immuno Sorbent Assay") plaat te testen waarvan alle reactievaatjes gevuld zijn met hetzelfde staal. Bij een dergelijke opstelling spreekt men van de "intratestprecisie"

(herhaalbaarheid). Indien men de herhaalbaarheid tussen twee verschillende uitvoeringen van de test wil bepalen (meerdere malen na elkaar dezelfde test uitvoeren op dezelfde stalen) spreekt men van "uitvoering-tot-uitvoering-precisie". Als men dagelijks eenmalig de test uitvoert dan spreken we van "dag-tot-dag-precisie" (herhaalbaarheid). Indien dezelfde stalen door twee of meer laboratoria worden getest en men de overeenkomst tussen de uitslagen bepaalt, dan spreken we niet meer over de herhaalbaarheid maar wel over de reproduceerbaarheid van de testen, of "laboratorium-tot-laboratorium-precisie". Zoals reeds hoger besproken dient de test, om vergelijking mogelijk te maken, steeds onder gecontroleerde en vaststaande omstandigheden uitgevoerd te worden.

Precisie en analytische accuraatheid worden gemeten en gecontroleerd door het opmaken van controlekaarten (Statistical Process Control) en de organisatie van interlabotesten (ringtesten, proficiency testing).

Het opmaken van controlekaarten is een continu proces en dient dus te gebeuren bij elke uitvoering van de test. Het principe van controlekaarten bestaat erin dat tijdens elke nieuwe uitvoering van de test steeds dezelfde controlestalen meegetest worden. De resultaten van deze controlestalen worden dan in een grafiek uitgezet en de statistische mogelijkheid wordt nagegaan of het resultaat van de controlestalen bij de laatste uitvoering van de test gelijkaardig is met de resultaten van het controlestaal bij een reeks vorige uitvoeringen van de test (zie htttp://www.westgard.com). Het voorleggen van controlekaarten wordt tegenwoordig geëist indien een laboratorium een bepaalde diagnostische test wil accrediteren.

Interlaboratoriumtesten worden meestal georganiseerd door een referentielaboratorium of een al dan niet officiële instantie die zich daarmee bezighoudt. Resultaten van interlabotesten dienen steeds beschikbaar te zijn voor de deelnemende laboratoria zodat er op elk moment kan ingegrepen worden indien er anomalieën vastgesteld worden.

DIAGNOSTISCHE ACCURAATHEID (SENSITIVITEIT EN SPECIFICITEIT)

In tegenstelling tot de precisie en de analytische accuraatheid van een test, die eigenlijk meer de labotechnische kant van een testvalidatie omvatten, zijn de sensitiviteit en de specificiteit echte populatieparameters die omschrijven hoe een test zich in een bepaalde populatie gedraagt.

De sensitiviteit (*Se*) van een test is de kans dat een test een positief resultaat geeft wanneer men aselect één dier uit de populatie zieke dieren selecteert, terwijl de specificiteit (*Sp*) de kans is op een negatief testresultaat wanneer men aselect één dier uit de populatie niet-zieke dieren selecteert. Daar hier de zieke dieren van de niet-zieke dieren moeten onderscheiden worden, moet men een perfecte referentietest (gouden standaard) ter beschikking hebben. Weergegeven in een kruistabel (Tabel 1) bekomt men vier cellen waarbij *EP* en *VP* respectievelijk het aantal echtpositieve en valspositieve zijn, *EN* en *VN* de echtnegatieve en de valsnegatieve in de populatie.

Tabel 1. Schematische voorstelling van echtpositieve, valspositieve, valsnegatieve en echtnegatieve testresultaten in een kruistabel.

	Ziek	**Niet-ziek**
Testpositief	*EP*	*VP*
Testnegatief	*VN*	*EN*

De *Se* en *Sp* kunnen als volgt uit tabel 1 berekend worden.

$$Se = \Pr(T+|D+) = \frac{\text{aantal testpositieve dieren in de zieke populatie}}{\text{totaal aantal dieren in de zieke populatie}} = \frac{EP}{EP+VN}$$

$$Sp = \Pr(T-|D-) = \frac{\text{aantal testnegatieve dieren in de niet-zieke populatie}}{\text{totaal aantal dieren in de niet-zieke populatie}} = \frac{EN}{EN+VP}$$

met $\Pr(T+|D+)$ de kans dat een ziek dier positief test
$\Pr(T-|D-)$ de kans dat een niet-ziek dier negatief test

Voorbeeld

Alle varkens in een bedrijf van 1000 dieren worden getest voor een bepaalde ziekte met een gouden standaardtest (ziek of niet-ziek) en een nieuwe test (testpositief of testnegatief). Men wil de prevalentie van de ziekte kennen in het bedrijf en de diagnostische accuraatheid van de nieuwe test op het bedrijf (het bedrijf is hier de populatie). De bekomen resultaten worden weergegeven in de volgende kruistabel (Tabel 2).

Uit de bovenstaande formules voor sensitiviteit en specificiteit is het duidelijk dat de sensitiviteit van de test *Se* =90/(90+10)=0,9 en *Sp*=800/(800+100)= 0,89. Merk op dat in dit geval de prevalentie Pr(*D*+)= (90+10)/1000=0,1.

Net zoals bij de populatieprevalentie is het hier eveneens zo dat, aangezien de gehele populatie getest

Tabel 2. Schematische voorstelling van de testresultaten van de 1000 varkens in een kruistabel.

	Ziek	Niet-ziek
Testpositief	90	100
Testnegatief	10	800

werd, de berekende *Se* en *Sp* de exacte waarden voor de diagnostische accuraatheid van de test in deze bepaalde populatie zijn. Dit houdt onder andere in dat de bekomen waarden strikt genomen niet gelden voor een andere populatie (een ander varkensbedrijf).

Indien men de prevalentie van de ziekte en de diagnostische accuraatheid van de test wil kennen bij bijvoorbeeld alle Belgische varkens, is het duidelijk dat zowel om praktische als financiële redenen nooit alle Belgische varkens zullen onderzocht worden, maar slechts een deel van de populatie. Net zoals bij de prevalentieschattingen door middel van een steekproef zal dit bij de sensitiviteit- en specificiteitschattingen ook aanleiding geven tot een betrouwbaarheidsinterval rond de schattingen. Men is pas zeker van de *Se* (*Sp*) van een test als men ALLE zieke (niet-zieke) dieren uit de populatie heeft getest. Indien dit niet zo is, en men slechts een aselecte steekproef van de populatie test, geeft dit onvermijdelijk aanleiding tot een onzekerheid rond de schatting, wat zich uitdrukt in betrouwbaarheidsintervallen. Net zoals voor de populatiewaarden geldt ook voor de schattingen dat de bekomen waarden en betrouwbaarheidsintervallen slechts gelden voor de populatie waaruit de steekproef is geselecteerd.

Dit houdt dus in, zoals ook beschreven door de Office International des Epizoöties (OIE, 2000) dat het onontbeerlijk is de diagnostische accuraatheid (sensitiviteit en specificiteit) van een test te bepalen voor de populatie waarvoor de test uiteindelijk zal gebruikt worden. De diagnostische accuraatheidsparameters zijn dus karakteristieken van een test die gelden voor de populatie waarin ze bepaald werden. Verdere extrapolatie naar andere populaties is niet mogelijk want daar gelden ze niet (i.e. men zou geheel andere resultaten kunnen bekomen als men de test in een andere populatie valideert).

RELATIE TUSSEN DE DIAGNOSTISCHE ACCURAATHEID EN DE GEBRUIKTE AFKAPWAARDE

De meeste serologische testen produceren uitslagen die gemeten worden op de ordinale schaal (eindtiters of scores) of de continue schaal (een kwantitatieve aflezing van één enkele dilutie, bijvoorbeeld ELISA OD's (Optische Densiteiten)). Daar een score of een cijfer niet rechtstreeks weergeeft of een dier positief of negatief test, bepaalt men een punt op de originele schaal vanaf waar men de test positief beschouwt. Testuitslagen die onder dit punt liggen, worden dan als negatief beschouwd. Dit punt is de afkapwaarde (cut-off). Het is het punt waarop men de beslissing neemt: "vanaf hier is de test positief". Een vergelijking van deze "gedichotomiseerde" resultaten met de echte ziektestatus van het dier leidt tot schattingen van de sensitiviteit en de specificiteit zoals hoger beschreven. Afkapwaarden en diagnostische accuraatheidsparameters zijn dus onlosmakelijk met elkaar verbonden. Een verandering van de afkapwaarde leidt automatisch tot een verandering van de sensitiviteit en de specificiteit van een test. Dit wordt gevisualiseerd voor een hypothetische ELISA in Figuur 1.

Op Figuur 1 zien we links de verdeling van de ELISA-resultaten (OD's) voor de niet-zieke dieren en rechts de verdeling voor de zieke dieren (de onderverdeling niet-ziek en ziek wordt door een gouden standaard bepaald). De stippellijnen a, b en c stellen drie verschillende afkapwaarden voor (a=0,5, b=0,8, c=1,17). Het gebied links van een stippellijn wordt als testnegatief beschouwd terwijl de stippellijn en het gebied rechts ervan testpositief is. De relatie tussen de afkapwaarde en de diagnostische accuraatheid is gemakkelijk te zien op de figuur. Een verlaging van de afkapwaarde (van c naar b naar a) leidt tot een stijging van het aantal testpositieve dieren uit de zieke populatie (verhoging van de sensitiviteit) maar ook tot een stijging van het aantal testpositieve dieren uit de

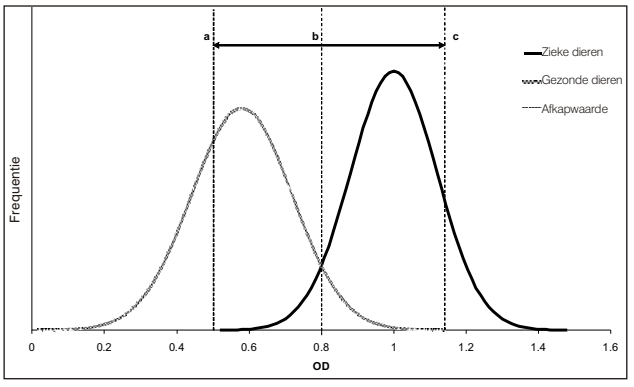

Fig. 1. Distributie van de ELISA-resultaten voor de zieke en niet zieke dieren.

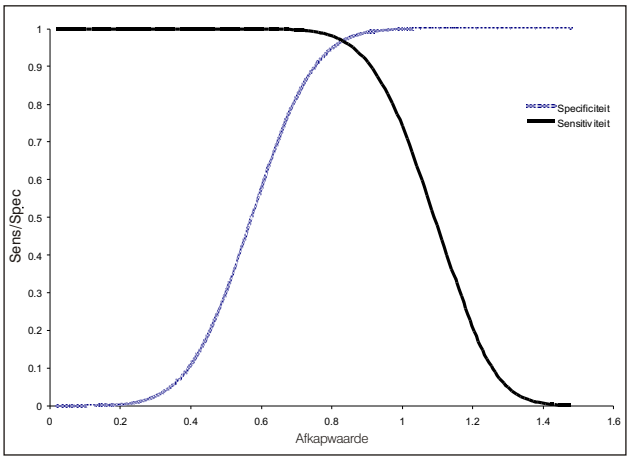

Fig. 2. Diagnostische sensitiviteit en specificiteit in functie van de afkapwaarde.

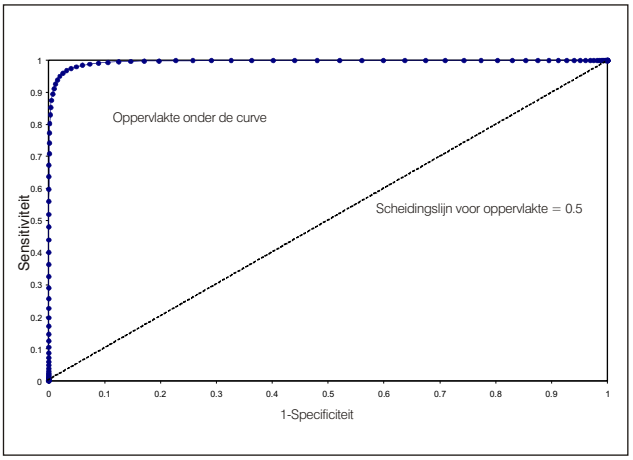

Fig. 3. ROC curve.

niet-zieke populatie (verlaging van de specificiteit). De uiterste situatie is dat de afkapwaarde zodanig laag is dat alle, zowel zieke als niet-zieke dieren, testpositief worden. In dit geval kan men absoluut zeker zijn dat men alle zieke dieren gedetecteerd heeft (sensitiviteit = 1) maar ook dat men alle niet-zieke dieren als valspositief heeft beschouwd (specificiteit = 0). Een verlaging van de afkapwaarde leidt dus tot een hogere sensitiviteit maar ook tot een lagere specificiteit. Dezelfde redenering gaat op voor een verhoging van de afkapwaarde (van a naar b naar c). In dit geval zal de specificiteit van de test stijgen maar de sensitiviteit dalen. Indien we voor de resultaten van Figuur 1 de sensitiviteit en de specificiteit van de test uitzetten in functie van de afkapwaarde, bekomen we Figuur 2 waar duidelijk te zien is hoe de sensitiviteit en de specificiteit van de test variëren in functie van de gebruikte afkapwaarde.

In Figuur 3 wordt de Receiver Operating Characteristic (ROC) voorgesteld waarin voor alle mogelijke afkapwaarden de sensitiviteit en (1-specificiteit) ten opzichtte van elkaar worden uitgezet.

De oppervlakte onder de curve (OOC) is dan een maat voor de diagnostische accuraatheid van de test, samengevat voor alle afkapwaarden. Een test die voor alle afkapwaarden een perfecte sensitiviteit en specificiteit vertoont, zal een OOC=1 hebben.

Volgens arbitraire richtlijnen (Swets *et al*., 1988) kunnen we een onderscheid maken tussen niet-informatieve (OOC=0,5), minder accurate (0,5<OOC ≤0,7), gemiddeld accurate (0,7<OOC≤0,9), hoogaccurate (0,9<OOC<1) en perfecte testen (OOC=1).

Verschillende manieren kunnen aangewend worden om de OOC te berekenen en we verwijzen naar Greiner (2000c) voor een overzicht.

Zoals beschreven door Hanley en McNeil (1982) is de OOC gelijk aan de kans dat één aselect gekozen staal uit de zieke populatie een hoger testresultaat zal vertonen dan één aselect gekozen staal uit de niet-zieke populatie. Indien de OOC=0,5 hebben we dus een gelijke kans op een hoger resultaat in de zieke populatie en de niet-zieke populatie en is de test dus waardeloos. Door het berekenen van het betrouwbaarheidsinterval rond de OOC kunnen we nagaan in hoeverre 0,5 al dan niet vervat zit in dit betrouwbaarheidsinterval. Voor testen waar dit zo is, kunnen we zeggen dat er statistisch geen bewijs is dat de test werkt. Dit kan gekwantificeerd worden met een p-waarde door een hypothesetest uit te voeren. Analoog kan men voor het vergelijken van twee testen het verschil in oppervlakte tussen beide OOC's berekenen en nagaan of dit statistisch van nul verschilt.

VOORSPELLENDE WAARDEN

Een belangrijke vraag is in hoeverre men iets bijgeleerd heeft over de echte ziektestatus van het dier nadat men het getest heeft. Dit wordt weergegeven door de voorspellende waarde van een testresultaat. De positieve voorspellende waarde $PVW=\Pr(D+|T+)$ geeft weer wat de kans is op ziekte van een bepaald dier gegeven dat men een positief testresultaat voor dit dier bekwam, en de negatief voorspellende waarde $NVW=\Pr(D-|T-)$ drukt de kans uit dat het dier niet ziek is indien men voor dit dier een negatief testresultaat bekwam. Beide parameters zijn dus verschillend van de sensitiviteit en de specificiteit.

De voorspellende waarden geven dus eigenlijk antwoord op de vraag: "wat heeft de test ons bijgeleerd?".

Indien men over gouden standaard informatie beschikt, kan de voorspellende waarde van een nieuwe test rechtstreeks uit Tabel 2 berekend worden als volgt

$$PVW=\frac{EP}{EP+VP} \text{ en } NVW=\frac{EN}{EN+VN}$$

De voorspellende waarden PVW en NVW kunnen ook gegeven worden in termen van de prevalentie, de sensitiviteit en de specificiteit. Met name geldt dat

$$\begin{cases} EP &= \pi*Se \\ VP &= (1-\pi)*(1-Sp) \\ EN &= (1-\pi)*Sp \\ VN &= \pi*(1-Se) \end{cases}$$

Bijgevolg kunnen voorgaande uitdrukkingen voor PVW en NVW herschreven worden als

$$PVW = \frac{\pi*Se}{\pi*Se + (1-\pi)*(1-Sp)}$$

$$NVW = \frac{(1-\pi)*Sp}{(1-\pi)*Sp + \pi*(1-Se)}$$

De prevalentie is dus de kans op een ziek dier alvorens men test (vandaar ook wel pre-test probabiliteit genoemd), en de voorspellende waarde de kans op een ziek of niet-ziek dier na een gegeven testresultaat (post-test probabiliteit). Beide zijn gelinkt door de sensitiviteit en de specificiteit van de gebruikte test.

Indien Se+Sp=1 zal PVW=π en NVW=(1-π). De test heeft dus met andere woorden geen meerwaarde gegeven. In het geval dat Se+Sp<1 is er zelfs een verlies van informatie daar een testpositief (-negatief) dier minder kans heeft om ziek (niet ziek) te zijn dan een lukraak uit de populatie gekozen dier. Testen met Se+Sp<=1 reduceren dus niet de onzekerheid rond de ziektestatus en zijn dus, volgens de definitie gegeven in de inleiding, geen diagnostische testen.

SCHATTING VAN DE POPULATIEPREVALENTIE MET IMPERFECTE TESTEN

In dit deel bespreken we de accuraatheidsfout die men bekomt als men de populatieprevalentie wil schatten met imperfecte testen en hoe die te corrigeren.

Als een imperfecte test gebruikt wordt, dan kan een schatting van de ware populatieprevalentie gebaseerd op het aantal testpositieve n_{T+} van een aselecte steekproef van n dieren bekomen worden via de Rogan-Gladen schatter (Rogan en Gladen, 1978) gegeven door

$$\hat{\pi} = \frac{n_{T+}/n+Sp-1}{Se+Sp-1}$$

De Rogan-Gladen schatter kan als volgt worden afgeleid.

Aangezien er geen gouden standaard informatie aanwezig is, kent men enkel het aantal dieren dat positief zal zijn voor de test. De kans op een positief testresultaat is gelijk aan de som van de kans op een EP− resultaat en de kans op een VP− resultaat

$$Pr(T+)=\pi*Se+(1-\pi)*(1-Sp)$$

Pr(T+) wordt ook wel de schijnbare prevalentie genoemd. Voorgaande uitdrukking kan herschreven worden als:

$$\pi=\frac{Pr(T+)+Sp-1}{Se+Sp-1}$$

De schijnbare prevalentie kan geschat worden door het aantal testpositieve dieren in de steekproef te delen door de steekproefgrootte, n_{T+}/n , en als we deze schatter introduceren in de formule geeft ons dat een schatter voor de prevalentie, $\hat{\pi}$.

De Rogan Gladen schatter is enkel geldig voor testen die de onzekerheid rond de ziektestatus doen afnemen (zie de definitie van een diagnostische test). Dit houdt in dat, zoals beschreven in het hoofdstukje van de predictieve waarden Se + Sp groter dan 1 dient te zijn.

DISCUSSIE

Testvalidatie is een onderzoeksgebied dat slechts de laatste jaren in de diergeneeskunde (en de humane geneeskunde) de nodige aandacht gekregen heeft. Men komt langzamerhand tot het inzicht dat er geen absolute waarden bestaan voor de sensitiviteit en de specificiteit van een test, maar dat deze afhankelijk zijn van de populatie waarin ze gebruikt worden. Voor bepaalde aandoeningen kan men ook opperen dat voor serologische testen de sensitiviteit zal variëren naargelang het stadium van de infectie of de immuunstatus van de gastheer (Greiner en Gardner, 2000a, Greiner en Gardner, 2000b). Het is bijvoorbeeld bekend dat de sensitiviteit van de tuberculinetesten zal variëren naargelang het stadium van de infectie. Voor paratuberculose (*Mycobacterium paratuberculosis*) is beschreven dat de sensitiviteit van een antistof-ELISA opmerkelijk veel verschilde naargelang het dier zich in het eerste, tweede of derde stadium van de infectie bevond (Ridge *et al*., 1991; Socket *et al*., 1992). De specificiteit van antistofdetectietesten kan beïnvloed worden door de aanwezigheid van maternale antistoffen in het jonge dier, vaccinatie, behandeling of spontane genezing. Daarnaast zijn er ook

rapporteringen van diagnostische testen die verschillend presteren in bepaalde subgroepen van de populatie (geslacht, ras,…). Het is dus belangrijk daar zo veel mogelijk rekening mee te houden bij het opstellen van een validatiestudie. De gouden stelregel, die we al meerdere malen aangehaald hebben, blijft echter dat de test moet gevalideerd worden in de populatie waarin hij uiteindelijk gebruikt zal worden. Dit impliceert ook dat er bij een plotselinge insleep van ziekte in een gebied dat voordien vrij was er een hernieuwde validatie moet komen van de gebruikte diagnostische testen.

Het gevoeligste punt bij testvalidatie is natuurlijk de aanname van een gouden standaard als referentietest. Indien deze aanname niet klopt, is het duidelijk dat de bekomen diagnostische accuraatheidsschattingen van de te valideren test niet correct zullen zijn.

Sensitiviteit en specificiteit van de te valideren test kunnen zowel overschat als onderschat worden afhankelijk van de fouten die de referentietest maakt en de correctie daarvoor. Alhoewel ze mathematisch mogelijk is, moet ze met de nodige omzichtigheid benaderd worden. We verwijzen naar Pepe (2003) voor een gedetailleerde discussie.

Daar de gouden standaard test veelal een utopie is en de ziektestatus van een individu nooit met zekerheid kan vastgesteld worden, kan men stellen dat de ziektestatus een niet-observeerbare (latente) variabele is die moet onderscheiden worden van de observeerbare testresultaten. Vanuit dit gedachtegoed werden recent statistische technieken ontwikkeld die de assumptie van een gouden standaard ontwijken en toch toelaten schattingen te bekomen van prevalentie, sensitiviteit en specificiteit (en predictieve waarden). Deze technieken die samengevat kunnen worden onder de noemer “latent class analyse” worden meer en meer gebruikt zowel in de humane geneeskunde als in de diergeneeskunde (Boelaert *et al.*, 1999; Goetgebeur *et al.*, 2000; Enoe *et al.*, 2001). De bespreking van deze methoden vergt echter een zwaardere wiskundige en statistische benadering en valt dus buiten het doel van dit artikel.

In de realiteit worden ook veelal meerdere testen gebruikt en worden er beslissingen genomen omtrent de behandeling of het afslachten gebaseerd op de resultaten van deze meerdere testen. Het bepalen van sensitiviteit, specificiteit en predictieve waarden van deze testreeksen is zeer belangrijk en kan een uitsluitsel geven over hoe men de verschillende testuitslagen het beste interpreteert. Hierbij is het vooral van belang dat men de correlatie tussen de testuitslagen voor een welbepaalde ziektestatus in acht neemt (conditionele afhankelijkheid). Testen die bijvoorbeeld consequent dezelfde dieren als valspositief en valsnegatief gaan beschouwen, leveren geen extra informatie op over de ziektetoestand van het dier indien ze gebruikt worden als gecombineerd diagnostisch middel. Het effect van de conditionele afhankelijkheid op diagnose en surveillance wordt besproken door Gardner *et al*. (2000).

Diagnostische testen kunnen ook aangewend worden om een beslag als positief of negatief te classificeren (Christensen en Gardner, 2000). Meestal neemt men hiervoor aan dat het beslag positief is als er één testpositief dier in het beslag wordt gevonden, maar deze grens kan ook op twee of meer testpositieve dieren gelegd worden. Maatstaven, zoals beslagprevalentie, -sensitiviteit en -specificiteit kunnen dan berekend worden, maar passen niet binnen de opzet van dit artikel. Ook het testen van bijeengevoegde of gepoolde stalen werd in dit artikel niet besproken (Vansteelandt *et al*. 2000).

Het kwantificeren van kansen en de interpretatie ervan om beslissingen te ondersteunen zijn samen te brengen onder de noemer kwantitatieve risico analyse (Vose, 2000). De validatie van diagnostische testen en de correcte interpretatie ervan zijn dus een onontbeerlijk onderdeel indien men de risico’s op uitbraak, ziekte-import of -export en contaminatie van de voedselketen wil kwantificeren en om tot de juiste maatregelen te komen om deze risico’s te beperken.

REFERENTIES

Boelaert M., Aoun K., Liinev J., Goetghebeur E., Van der Stuyft P. (1999). The potential of Latent Class Analysis in diagnostic test validation for canine *Leishmania infantum* infection. *Epidemiology and Infection 123*, 499-506.

Christensen J, Gardner I.A. (2000). Herd-level interpretation of test results for epidemiologic studies of animal diseases. *Preventive Veterinary Medicine 45*, 83-106.

Enoe C., Andersen S., Sorensen V., Willeberg P. (2001). Estimation of sensitivity, specificity and predictive values of two serologic tests for the detection of antibodies against Actinobacillus pleuropneumoniae serotype 2 in the absence of a reference test (gold standard). *Preventive Veterinary Medicine 51,* 227-43.

Gardner I.A., Stryhn H., Lind P., Collins M.T. (2000). Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. *Preventive Veterinary Medicine 45*, 107-122.

Goetghebeur E., Liinev J., Boelaert M., Van der Stuyft P. (2000). Diagnostic test analyses in search of their gold standard: latent class analyses with random effects. *Stat Methods Med Res 9*, 231 -248.

Greiner M. and Gardner I.A. (2000a). Application of diagnostic tests in veterinary epidemiologic studies. *Preventive Veterinary Medicine 45*, 43-59.

Greiner M., Gardner I.A. (2000b). Epidemiologic issues in the validation of veterinary diagnostic tests. *Preventive Veterinary Medicine 45*, 3-22.

Greiner M., Pfeiffer D. Smith R.D. (2000c). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine 45*, 23-41.

Hanley J.A. and McNeil B.J. (1982). The meaning and use of the area under a receiver operating characteristic curve. *Radiology 143,* 29-36.

OIE (2000). Principles of validation of diagnostic assays for infectious diseases. *Manual of Standards for Diagnostic Tests and Vaccines*, In: Office International des Epizooties (OIE).

Pepe M.S. (2003) *The statistical evaluation of mediacal tests for classification and prediction*. Oxford statistical science series 28, Oxford university press, England.

Ridge S.E., Morgan I.R., Socket D.C., Collins M.T., Condron R.J., Skilbeck N.W., Webber J.J. (1991). Comparison of the Johne's Absorbed EIA and the complement fixation test for the diagnosis of Johne's disease in cattle. *Australian Veterinarian Journal 68*, 253-257.

Rogan W.J., Gladen B. 1978. Estimating prevalence from the results of a screening test. *American journal of epidemiology 107*, 71-76.

Socket D.C., Conrad T.A., Thomas C.B., Collins M.T. (1992) Evaluation of four serological tests for bovine paratuberculosis. *Journal of Clinical Microbiology 30*, 1134-1139.

Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. *Science 240,* 1285-1293.

Vansteelandt S., Goetghebeur E., Verstraeten T. (2000). Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics 56*, 1126-1133.

Vose, D. (2000). *Risk Analysis: a Quantitative Guide*, 2nd Ed John Wiley and sons Chichester, England.