

Ontwikkeling en validatie van een systematisch intern kwaliteitsbewaking-programma voor microscopische semenbeoordeling met behulp verse semenstalen

A.K. BOER

Afgelopen jaar is de vijfde editie van de WHO-handleiding voor semendiagnostiek (1) verschenen. Eén van de grote verschillen met de vierde editie uit 1999 betreft de positionering van een kwaliteitsbewaking-programma binnen de semendiagnostiek. Werd in de vierde editie slechts een subhoofdstuk van 15 bladzijden aan kwaliteit gewijd, de vijfde editie bestaat uit drie even belangrijke hoofdonderdelen: 1) semenanalyse, 2) semenopwerking en 3) kwaliteitsbewaking. Met 26 bladzijden is dit laatste deel ook qua tekstuele omvang fors toegenomen. Een kwaliteitsysteem, omvattende zowel interne als externe kwaliteitsbewaking, wordt als norm gesteld. De meeste laboratoria participeren wel aan een extern kwaliteitsbewaking-programma, maar een rondvraag bij diverse collegae leerde me dat er eigenlijk nog geen simpel algoritme bestaat dat op een systematische wijze op wekelijkse basis met minimale inspanning zowel individuele als longitudinale informatie genereert over precisie en juistheid, en dat bovendien geheel overeenkomt met de door de WHO voorgestelde statistiek. Bij de ontwikkeling van een dergelijk algoritme werden de volgende vier uitgangspunten geformuleerd. 1. Het algoritme moet in staat zijn om de juistheid en precisie tussen analisten in kaart te brengen voor zowel absolute metingen (bijvoorbeeld celtelling), als percentuele metingen (bijvoorbeeld motiliteit). 2. Het algoritme moet schaalbaar zijn zodat het niet uitmaakt of er nu twee analisten of bijvoorbeeld twaalf analisten participeren. 3. Het algoritme moet aansluiten bij de praktijk, waarbij verse semenstalen worden gebruikt. 4. Het algoritme moet met minimale tijd en kosten alle door de WHO vereiste informatie genereren. In dit manuscript wordt de ontwikkeling en validatie beschreven van dit nieuwe algoritme, dat gebruikt maakt van verse semenstalen in plaats van beads of gefixeerd semen. Om dit mogelijk te maken werd de statistiek achter de rekenvoorbeelden van de WHO herleid. Vervolgens werd deze op twee punten aangepast zodat enerzijds het algoritme geschikt is voor meer dan tien analisten en anderzijds dat verse semenstalen kunnen worden gebruikt. Om niet gebruik te hoeven maken van dure statistische software pakketten, is de door de WHO gebruikte statistiek in dit manuscript uitgeschreven. Bij de validatie werd naast het valideren van de rekenalgoritmes ook expliciet gekeken of vanuit

het analistenperspectief werd voldaan aan het vierde uitgangspunt zoals hierboven geformuleerd. Het algoritme is geprogrammeerd in Microsoft Excel en is op aanvraag beschikbaar voor geïnteresseerden.

Methode

Uitgangspunt bij het vergelijken van celtellingen is dat de betrouwbaarheidsinterval van een celtelling (N -cellen) afhankelijk is van diezelfde telling¹: $N \pm 1,96 \times \sqrt{N}$. Voor percentages geldt dat de betrouwbaarheidsinterval gelijk is aan 1,96 maal de geschatte standaardfout van percentage (p). De betrouwbaarheidsmarge² bij een gegeven percentage p , gemeten op basis van

N -spermatozoën is
$$p \pm 1,96 \times \sqrt{\frac{p \times (100 - p)}{N}}$$
 (voor een gegeven percentage p tussen 20% en 80%.

en daarbuiten
$$\left(\sin \left(\sin^{-1} \left(\sqrt{\frac{p}{100}} \right) \pm \frac{1,96}{2 \times \sqrt{N}} \right) \right)^2$$

Gebaseerd op deze betrouwbaarheidsmarges definieert de WHO alarmgrenzen waartussen de tellingen van de analisten moeten liggen. Voor percentages geldt dat dit gelijk is aan de betrouwbaarheidsinterval rondom het gevonden groepsgemiddelde volgens bovenstaande formules. Voor absolute celtellingen is dit iets gecompliceerder omdat de alarmgrenzen afhangen van het aantal analisten dat de celtelling heeft uitgevoerd. Onderstaande formules komen geheel overeen met tabel 7.1 (blz. 186 WHO-handleiding), met dien verstande dat deze nu ook toepasbaar zijn voor meer dan tien analisten.

De berekening van de alarmgrenzen rondom een gemiddelde semenconcentratie $X_{bar,i}$ bij i QC-ronden gaat als volgt³: $X_{bar,i} \pm A_{2,n} \times S_{bar,i}$, waarbij $A_{2,n}$ bij een even aantal analisten ($n=2k$) gelijk is aan

$$\sqrt{\frac{2}{\pi(2k-1)}} \times \frac{2(k-1) \times (2^{k-2} \times (k-2)!)^2}{(2k-3)!}$$

en bij een oneven aantal analisten ($n=2k+1$) gelijk is aan:

$$\sqrt{\frac{\pi}{k}} \times \frac{(2k-1)!}{2 \times (2^{k-1} \times (k-1)!)^2}$$

Algemeen Klinisch Laboratorium, Catharina-Ziekenhuis Eindhoven

E-mail: arjen-kars.boer@cze.nl

1 Zie WHO-handleiding box 2.5 en paragraaf 2.5

2 Zie WHO-handleiding box 2.7 en paragraaf 7.8

3 Zie WHO-handleiding box 7.2 en tabel 7.1

Bij het gebruik van een geconserveerd QC-materiaal is $S_{bar,i}$ gelijk aan de gemiddelde standaarddeviatie van elke QC-ronde. Bij het gebruik van verse materialen met wisselende concentraties is de rekenwijze van de WHO voor het berekenen van $S_{bar,i}$ niet bruikbaar. Door de concentratieafhankelijkheid van de standaarddeviatie wordt $S_{bar,i}$ derhalve beter berekend met een lineaire regressie model tussen de standaarddeviatie per ronde (y_i) tegen de concentratie (x_i). $S_{bar,i}$ is dan gelijk aan $a \times x_i + b$, waarbij geldt dat

$$b = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n} (\sum x_i)^2} \quad \text{en} \quad a = \bar{y} - b \times \bar{x}$$

Deze benadering geeft een realistischer beeld van de alarmgrenzen in het hoge gebied.

Tot slot stelt de WHO dat systematische verschillen tussen de analisten moet worden opgespoord met een F-toets. In box 7.6 (blz. 195 WHO-handleiding) wordt een rekenvoorbeeld gegeven, maar de onderliggende formules zijn niet expliciet uitgeschreven. Hieronder vindt u de formules achter de F-toets en de omrekening naar een p-waarde.

$$F = \frac{\sum_{j=1}^t n_j (\bar{x}_j - \bar{x})^2 / (t-1)}{\sqrt{\frac{1}{t-1} \times \sum_{j=1}^t \left(\frac{1}{n_j - 1} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_i)^2 \right)}}$$

waarbij t het aantal analisten is, n_j het aantal meetwaarden van de j^{de} analist, \bar{x}_j de gemiddelde meetwaarde van de j^{de} analist, \bar{x}_i de gemiddelde meetwaarde van de i^{de} QC-ronde en x_{ij} de meetwaarde van de j^{de} analist in de i^{de} QC ronde. Deze F-waarde kan als volgt worden uitgedrukt in een p-waarde;

$$p(F, q, r) = \int_{m=F}^{m=\infty} \frac{1}{\int_{k=0}^{k=1} k^{\frac{q-1}{2}} \times (1-k)^{r-1} dk} \times \left(\frac{q}{r}\right)^{\frac{q}{2}} \times m^{\left(\frac{q-1}{2}\right)} \times \left(1 + \frac{q}{r} m\right)^{\left(\frac{q+r}{-2}\right)} dm$$

waarbij geldt dat q het aantal QC ronden minus 1 is en r gelijk is aan $q \times$ (het gemiddelde aantal deelnemers per QC ronde minus 1). Deze F-toets geeft alleen aan of er een significant verschil bestaat tussen de analisten, maar geeft niet aan door wie dit verschil wordt veroorzaakt.

Grafische weergave

De door ons ontwikkelde spreadsheet genereert de volgende grafieken voor een parameter als de semenconcentratie: een Levey-Jennings grafiek met daarin de X_{bar} alarmgrenzen (figuur 1A), een cumulatieve afwijking per analist in de tijd (figuur 1C), de absolute resultaten per QC ronde (niet getoond) en de absolute resultaten uitgezet tegen de gemiddelde concentratie

(vergelijkbaar met figuur 1B). Voor percentuele parameters, waaronder de motiliteit, worden ook vier grafieken gegenereerd die vergelijkbaar zijn met die van de semenconcentratie. De X_{bar} berekening is echter niet toepasbaar op percentages, vandaar dat deze niet worden weergegeven.

Individuele scores

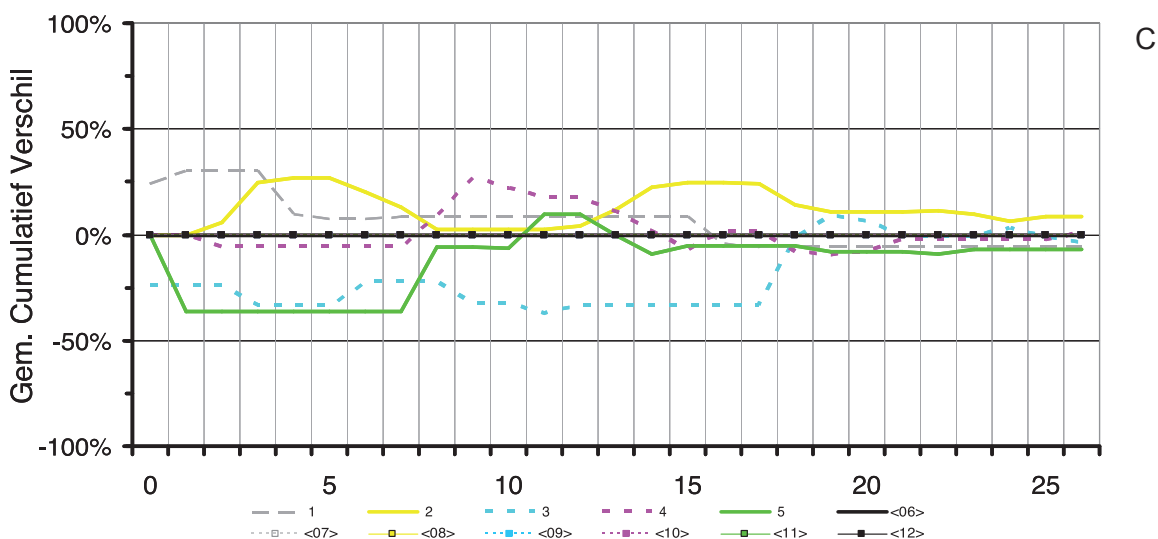
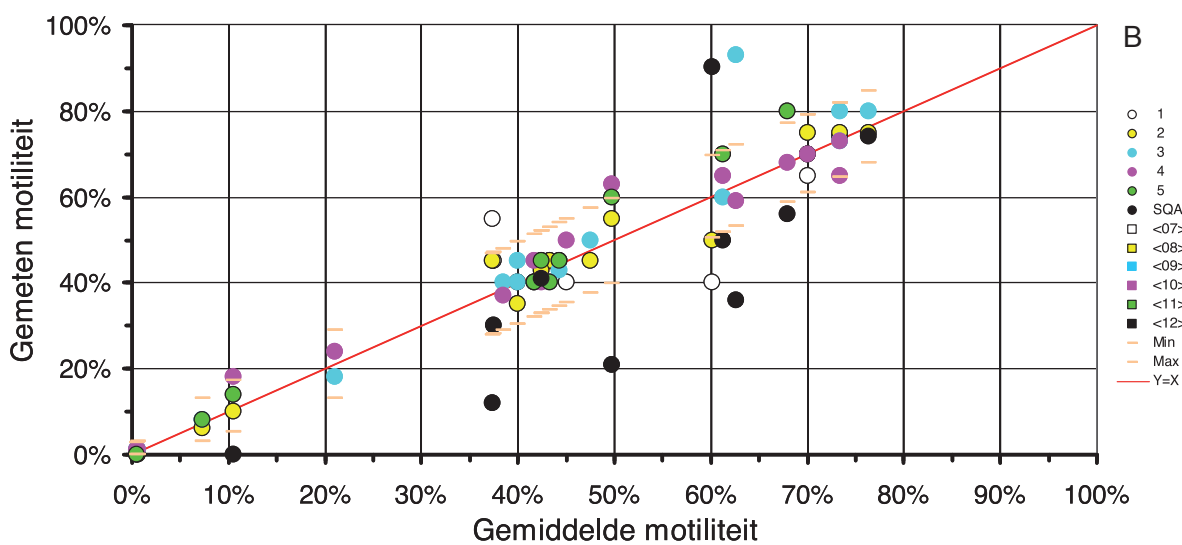
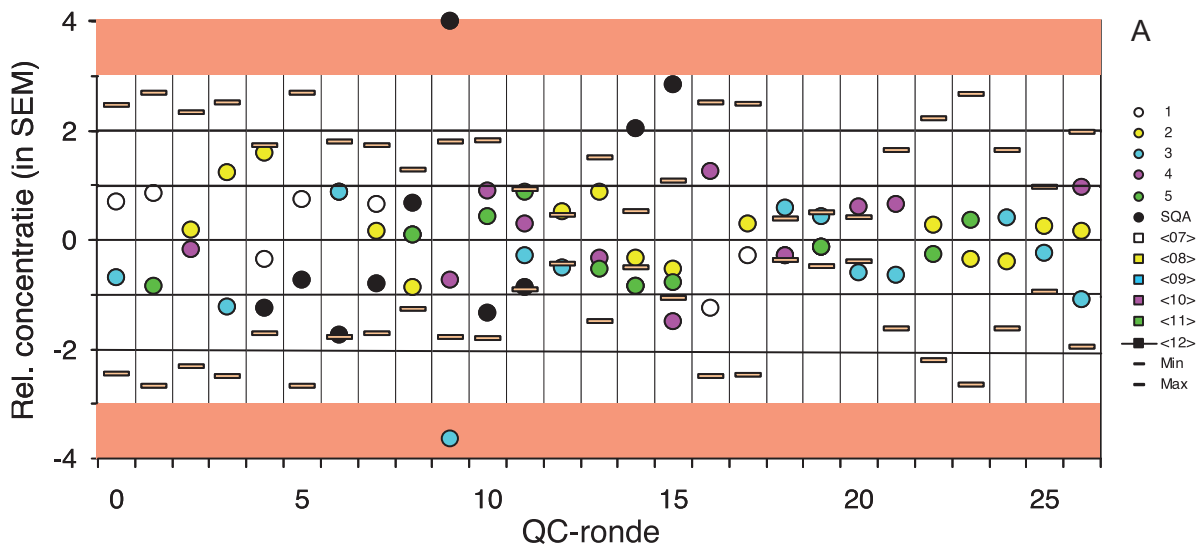
Van alle deelnemers worden de volgende statistieken bijgehouden: het aantal en percentage QC-ronden waaraan is deelgenomen (n QC's en %QC's), het percentage van de QC-resultaten die binnen de norm vielen (% goed), de algehele relatieve en absolute afwijking (rel.bias en abs.bias), en de SE-score (tabel 1). De SE-score wordt berekend door abs.bias te delen door de gemiddelde standaarddeviatie over alle QC-ronden en van alle analisten. Wanneer de SE-score groter is dan 3 wordt een significant verschil verondersteld en wordt dit visueel in de tabel weergegeven met een rode achtergrond. Daarnaast wordt bijgehouden met welke analisten (in percentages) de QC-materialen zijn mede-beoordeeld.

Resultaat

Dit algoritme werd de tweede helft van 2010 door vijf analisten gebruikt (figuur 1). Daarnaast hebben we de SQA-V tijdens het validatietraject als analist laten meelopen. Na introductie van dit algoritme op ons laboratorium kwam naar boven dat één analist een systematische afwijking in de semenconcentratie had van gemiddeld 3%. De afwijking verdween nadat de gehele procedure van semenanalyse nogmaals was doorgesproken. Daarnaast nam het maximale cumulatieve verschil in semenconcentratie tussen twee analisten af van 50% (gedurende de eerste acht weken), naar minder dan 10% over een periode van 26 weken.

Discussie

Het algoritme werd door de analisten als zeer waardevol gezien, en de compactheid van registratie (slechts twee getallen invoeren per analist) werd gewaardeerd. Daarnaast werd tijdsinvestering van het tellen door de analisten als minimaal ervaren, omdat alleen semenstalen werden geteld die toch al geteld moesten worden voor de kliniek. In het verleden werden geregeld semenstalen door meerdere analisten geteld, maar een systematische analyse van longitudinale trends en tussen-analist-variatie vond niet plaats. Een belangrijke beperking van dit algoritme is de tijdsspanne waarbinnen de verse semen moeten worden beoordeeld. Binnen ons laboratorium hanteren wij een half uur als maximum tussen eerste en laatste beoordeling.



Figuur 1. Grafische weergave van statistische algoritmen voor de interne kwaliteitsbewaking van microscopische (semen) beoordelingen conform de vijfde WHO-handleiding. Voor zowel absolute parameters (bijvoorbeeld concentratie) als relatieve parameters (bijvoorbeeld motiliteit) kunnen maximaal 12 analisten hun bevindingen van verse semenstalen blind invoeren. De spreadsheet berekent Levey-Jennings plots [A]; *semenconcentratie* en *semenmotiliteit*, en geeft de bevindingen weer per QC-ronde (niet getoond) en B) als concentratie/percentageafhankelijke grafiek. C) Tot slot wordt ook nog de *cumulatieve afwijking in concentratie* en percentage van de analisten in de tijd weergegeven. Per analist wordt één curve weergegeven. De door de WHO gestelde maximale en minimale alarmeringswaarden (Tbar-score) staan als horizontale strepen weergegeven (min & max in figuur 1A en 1B). De cursieve parameters zijn hierboven als voorbeeld weergegeven.

Tabel 1. Individuele scores van het intern kwaliteitsbewakingprogramma voor de semenmotiliteit. Een vergelijkbare tabel wordt berekend voor de semenconcentratie. Van elke analist wordt weergegeven aan hoeveel QC-ronden hij/zij heeft deelgenomen (absoluut: n QC's en relatief: % QC's). Daarnaast wordt per analist het percentage weergegeven van het aantal QC-ronden dat binnen de WHO-norm valt (% goed), de cumulatieve absolute bias (Abs.Bias) en de bijbehorende SE-score (zie methode). In het rechter gedeelte is terug te vinden wie de QC-preparaten van de desbetreffende analist heeft medebeoordeeld.

	N QC's	% QC's	% goed	Abs Bias	SE-score	1	2	3	4	5	SQA
Analist 1	7	26%	71%	0%	-0,25		45%	0%	18%	9%	27%
Analist 2	18	67%	94%	0%	0,05	13%		26%	26%	21%	15%
Analist 3	14	52%	93%	3%	1,74	0%	40%		32%	16%	12%
Analist 4	16	59%	88%	1%	0,63	6%	30%	24%		21%	18%
Analist 5	9	33%	78%	3%	1,76	4%	32%	16%	28%		20%
SQA-V	9	33%	33%	-10%	-5,14	13%	26%	13%	26%	22%	

Door deze tijdslimiet en door roostertechnische redenen kan soms niet iedereen deelnemen aan de QC-ronde. Door de deelnemersparticipatie goed te monitoren (tabel 1), wordt dit naar onze mening goed opgevangen. In ons geval is, bij voorbeeld, een aandachtspunt dat analist één en analist drie nooit samen een semenstalen hebben beoordeeld. Dat de SQA-V niet geschikt is om de motiliteit te beoordelen, was al gebleken uit het validatietraject. Vandaar dat we na QC-ronde 16 gestopt zijn om de SQA-V mee te nemen als 'analist' voor de motiliteit. Strikt genomen zou de SQA-V data verwijderd moeten worden om het analisten gemiddelde juist vast te stellen. Zonder de SQA-V data vallen alle QC-metingen binnen de alarmgrenzen. Alleen analist 3 en analist 4 scoren bij de QC-ronde 9 inaccuraat. Door de SQA-V data toch mee te nemen, wordt goed geïllustreerd hoe een systematische afwijking van bijvoorbeeld 10% kan worden opgespoord.

Conclusie

Ons algoritme voor het monitoren van de tussen-analist-variantie bij microscopische semenbeoordeling voldoet aan de door de WHO gestelde criteria, zonder omslachtige of dure procedures en draagt bij aan een actieve verbetering van de tussen-analist-variantie. Bovendien wordt de werkelijke werkwijze getoetst en niet een artificiële afgeleide hiervan.

Referentie

1. WHO laboratory manual for the examination and processing of human semen – 5^e editie, 2010. ISBN 9789241547789.