Validering av analysekvalitet, Del 1. Tester og riktighet

PÅL RUSTAD

Fürst Medisinsk Laboratorium, Søren Bulls vei 25, N-105l Oslo
(prustad@furst.no)

Denne arikkelen gir en kortfattet fremstilling av noen av de vesentligste elementene i en metodevalidering for et klinisk kjemisk laboratorium. Emnene er basert på en "Standard Prosedyre" for metodevalidering som gjelder for vårt laboratorium.

Det er her brukt betegnelsen «bruksstandard» (working standard [2]) for det mer vanlig brukte «kalibrator» .

I et senere nummer av Klinisk Kjemi i Norden vil presisjon, måleområde og måleusikkerhet bli omtalt i en egen artikkel (Del 2).

Takk til Heidi Steensland som har kommet med nyttige innspill til utformingen av artikkelen.

Validering/verifiksjon

Definisjon av validering [1]: *Bekreftelse ved undersøkelse og fremskaffelse av objektive bevis, på at de spesielle krav for en spesifisert, antatt bruk er tilfredsstilt. (Objektive bevis: Informasjon som kan bevises sanne, basert på fakta fremskaffet ved observasjoner, målinger, forsøk eller på andre måter).*

Hvis reagens/bruksstandard for påvisning av en komponent er innkjøpt, skal metodene være validert av produsent før metoden valideres i laboratoriet. Dokumentasjon skal være tilgjengelig. I laboratoriet kan det dermed være tilstrekkelig å gjennomføre en forenklet validering (verifikasjon) for å kontrollere de mange forhold, både på instrument- og reagenssiden som kan bidra til at målesystemet ikke virker optimalt. For validering som er vanskelig å utføre lokalt, f.eks. av interferenser, må man stole på de opplysninger produsenten gir. Hvis det ikke finnes opplysninger som anses nødvendige, må det kreves at produsentleverandør bidrar med informasjon i slike tilfeller.

T-test

Den eneste hensikten med å gjennomføre en statistisk test er i grunnen å forvisse seg om at man ikke trekker slutninger fra data som er utilstrekkelige. Det er viktig å skille mellom signifikant (tydelig) og viktig avvik. Man kan få en hvilken som helst reell forskjell til å bli signifikant med mange nok data, men hvor viktig forskjellen er kan bare vurderes fra kvalitetsmål og har ingenting med statistikk å gjøre.

Anta at det er gjort et forsøk hvor to middelverdier M_1 og M_2 er beregnet basert på. hhv. n_1 og n_2 målinger med standard avvik hhv. s_1 og s_2. Testen består i å finne ut om forskjellen mellom middelverdiene er tydelig (signifikant) forskjellig fra 0 (nullhypotesen, H_0) mot en alternativ hypotese (H_l), enten at det er signifikant forskjell ($M_l \neq M_2$), at $M_l > M_2$ eller $M_l < M_2$.

Hvor tydelig, uttrykkes med signifikansnivået (p) som forteller hvor stor sannsynligheten er for at nullhypotesen forkastes når den er riktig (type 1 feil). Denne sannsynligheten bør være liten (alltid ≤ 0.05). Hvis det skal testes om det er signifikant forskjell, må absoluttverdien av denne, $|M_l - M_2|$, sammenlignes med standard avviket for forskjellen, $\sqrt{(s_1^2/n_l + s_2^2/n_2)}$ ved å beregne forholdet mellom dem: $|Ml-M2|/\sqrt{(s_1^2/n_l + s_2^2/n_2)}$. Hvis forskjellen utgjør mange standard avvik, er det stor sannsynlighet for at nullhypotesen må forkastes, dvs den absolutte forskjellen er reelt >0. Hvis standard avvikene hadde vært beregnet med mange resultater (>60) kan man i stedet for t benytte z (f.eks. gir p=0.05 en z-verdi på 1.96 ($z_{0.975}$), dvs hvis forholdet over er >1.96, så er forskjellen signifikant på 5% nivå. Hvis forskjellen er beregnet med et mindre antall resultater, må den relevante t finnes i stedet. Med f=l0 (bruker f for frihetsgrader i stedet for n) må man i stedet for

å bruke $Z_{0.975}$=1.96 bruke $t_{l0,\ 0.025}$ = 2.23. Grunnen til at t>z, er at usikkerheten i det beregnede standard avvik øker når antall måleresultater minker.

Det generelle uttrykket for signifikant forskjell på p nivå i forsøket skissert over, blir:

$|M_1-M_2|/\sqrt{(s_1^2/n_l+s_2^2/n_2)}>t_{n1+n2-2,p/2}$ som blir $|M_1-M_2|\sqrt{[n/(s_1^2+s_2^2)]}$ når $n_l = n_2$ (her forutsettes at $E(s_1)=E(s_2)=E(s)$ dvs. forventningen av s_1 og s_2 er den samme).

Hvis den alternative hypotesen (Hl) enten er at $M_1>M_2$ eller $M_l<M_2$ (ensidig test), vil uttrykket for signifikant forskjell bli
$|M_1-M_2|/\sqrt{(s_1^2/n_l+s_2^2/n_2)}>t_{2n-2,p}$ (p/2 over er erstattet med p).

Et l-p konfidensintervall er et intervall rundt den målte middelverdi som (i gjennomsnitt ved mange gjentagelser av testen) med sannsynligheten l-p inneholder den sanne verdi.

For eksempelet over er dette intervallet:
$M_1-M_2 \pm t_{n1+n2-2,p/2}\sqrt{(s_1^2/n_l+s_2^2/n_2)}$.

Hvis forskjellen IKKE er signifikant på p nivå (0 er inneholdt i l-p konfidensintervallet), men viktig (se «Analytiske mål» under), har man brukt for få data i forsøket.

Hvordan man i et forsøk kan sørge for at en viktig feil blir signifikant?

Hvis G er en viktig forskjell, må man, i et forsøk for å finne ut om forskjellen er så stor, passe på at man har et tilstrekkelig stort datamateriale (n) slik at en så stor feil blir signifikant på et ønsket nivå (p). Flg. relasjon gjelder i dette tilfellet:
$G>t_{n1+n2-2,p/2}\sqrt{(s_1^2/n_l+s_2^2/n_2)}$. Hvis $n_l=n_2=n$ og $s_1=s_2=s$, fås $n>2x[t_{2n-2,p/2}xs/G]^2$.

Eksempel

Man ønsker å teste om det er forskjell mellom to lot'er av bruksstandard ved å måle begge annen hver en i en serie med det samme antall av hver (n). Anta at innen serie variasjonen er ca. 2% (s), det tillates en forskjell på 3% (G), signifikansnivå velges til p=0.05. Da fås $n>2x[t_{n-2,p/2}x0.02/0.03]^2$ som gir n=5 dvs. 5 stk av hver bruksstandard vil være tilstrekkelig til at en feil på 3 % skal bli signifikant på 5% nivå.

Analytisk kvalitet

Riktighet

Definisjon: Grad *av overensstemmelse mellom gjennomsnittlig verdi fremskaffet fra en stor serie måleresultater* og *en sann verdi [2].*

Kvantitativt uttrykkes riktighet ved bias eller systematisk feil. Som et mål for systematisk feil (B) når den analytiske variasjonskoeffisient CVa er ubetydelig og total biologisk variasjon er CVt ($=\sqrt{(CVw^2+CVb^2)}$ der CVw og CVb er hhv. intra- og interindividuell biologisk variasjon), kan denne inndelingen være nyttig [3]:

Optimalt: B < 0,125 x CVt, ønskelig: B < 0,250 x CVt, minimum: B < 0,375 x CVt .

Hvis samme komponent måles på ulike instrumenter innen et laboratorium, foreslås flg. kriterium for akseptabel forskjell [3]: B<CVw/3.

Definisjon av sporbarhet: Egenskap for et måleresultat eller verdi for en standard som gjør at den kan relateres til angitte referanser, vanligvis nasjonale eller internasjonale standarder gjennom en ubrutt kjede av sammenligninger, alle med angitt usikkerhet [1].

Legg merke til at sporbarhet er etablert hvis «måleresultatet kan relateres til angitte referanser», det er altså ikke nødvendig å ha korrigert en evt. systematisk feil for at måleresultatet skal være sporbart. Ved usikkerhetsberegninger for måleresultatet forutsettes imidlertid at alle kjente systematiske feil er korrigert (se «Standard/ekspandert usikkerhet» i neste artikkel).

Test ved bruk av sertifisert referansemateriale

Definisjon: Referansemateriale: Et materiale eller stoff hvor en eller flere egenskaper er tilstrekkelig godt kjent til å kunne brukes for kalibrering av utstyr, vurdering av en målemetode eller for å bestemme verdier på *materialer [1].*

I korthet betyr «sertifisert» at verdiene for referansematerialet skal være funnet ved teknisk gyldig målemetode med angitt sporbarhet og usikkerhet.

Eksempel: Anskaff et sertifisert referansemateriale med oppgitt verdi (OV) og mål konsentrasjonen på en slik måte at usikkerheten i resultatet er tilstrekkelig lav. Mål prøven i N (f.eks. 10) replikater i ulike serier. Beregn middelverdi (M) og standard avvik (s) for de N målingene. For-

skjellen mellom oppgitt verdi og den målte verdi, D=M-OV, er et uttrykk for den systematiske feil for metoden ved denne konsentrasjonen med standard usikkerhet $uD=\sqrt{(uOV^2+s^2/N)}$ der uOV er angitt/beregnet standard usikkerhet for OV (som skal være oppgitt for et sertifisert referansemateriale). Forskjellen er signifikant hvis $|M-OV|>\sqrt{(uOV^2+s^2/n)}xt_{n-1,p/2}$.

Test mot referansemetode eller annen metode.
En meget vanlig og god metode for å teste overensstemmelse med annen metode, er å velge ut et antall prøver jevnt fordelt i hele måleområdet, måle dem med begge metoder, plotte dataene i et xy-diagram med verdier for referansemetode på x-akse og den andre metodens verdier på y-aksen (eller differansen mellom dem, da kalles det et differanseplott).

Hvis antall prøver som måles på begge metoder er større enn ca 5, kan man bruke regresjonsmetoder for å beregne parametre for systematiske (helning og skjæringspunkt) og tilfeldige forskjeller (variasjon rundt regresjonslinjen i form av for eksempel $s_{y/x}$). Det finnes mange typer regresjonsmetoder, alle med sine fordeler og ulemper. De viktigste er:

Standard minste kvadraters metode: Beregner regresjonslinjen slik at summen av de kvadrerte avvikene (i y-retning) fra regresjonslinjen er minimert. Antar normalfordelte og konsentrasjonsuavhengige avvik og usikkerhet bare i y-verdier. Finnes tilgjengelig i Excel regneark.

Veiet minste kvadraters metode: Som over, men usikkerhet varierer med konsentrasjon. I prinsippet kan man selv angi usikkerheten i hele måleområdet, men det vanlige er å anta konstant variasjonskoeffisient.

Demings metode: Antar usikkerhet både i x- og y-verdier (man kan også bruke veiet Demings metode).

Passing og Bablocks metode: Prinsippet er at man finner regresjonslinjens helning ved å beregne median av alle mulige kombinasjoner av linjer mellom to og to punkter. Skjæringspunktet med y-aksen finnes ved å legge linjer gjennom alle punkter med helning som funnet foran og beregne medianen av alle skjæringspunktene med y-aksen. Fordelen med denne metoden er at punkter som avviker mye fra resten får liten innvirkning på resultatet. Metoden er tilgjengelig i for eksempel statistikkprogrammet Analyse-It (http://www.analyse-it.com/info/genstat.htm).

Av disse 4 metodene er de tre første såkalte parametriske metoder fordi teorien forutsetter normalfordelte avvik fra regresjonslinje, mens den siste (Passing/Bablock) ikke forutsetter dette og derfor kalles en uparametrisk metode. I de fleste sammenhenger anbefales enten (veiet) Deming eller Passing og Bablocks metode.

Mye viktig informasjon kan hentes fra et slikt plott f.eks.: Særlig avvikende punkter kan tyde på interferenser eller molekylære isoformer, ikke lineær sammenheng tyder på problemer med standardkurven for minst en av metodene, skjæringspunkt med y-aksen forskjellig fra 0-prøve kan tyde på matriksproblem for blind eller en konstant interferens, avvik for helning kan tyde på gal verdi på minst en av bruksstandardene, stor spredning rundt regresjonslinje (men god presisjon for begge metodene), kan bety at komponentene som måles i metodene kan være noe forskjellige osv.

Man bør vurdere om sammenhengen er lineær, evt. om helning og skjæringspunkt er signifikant forskjellige fra ideelle verdier (1 og 0). Se for øvrig [4] for nødvendig antall prøver for at spesifiserte krav (forskjeller) skal bli signifikante ved bruk av ulike regresjonsmetoder.

Testing ved tilsetting av ren komponent til prøve («spiking»)

Hvis ren komponent er tilgjengelig, kan denne tilsettes aktuell matriks, måle prøve med og uten tilsatt komponent, og finne hvor stor mengde av det tilsatte som «gjenfinnes» (recovery) ved analyse. Finner man igjen 100% av tilsatt mengde, er dette en god indikator på at man ikke har systematiske feil for analysen, men konstante feil (som ikke varierer med konsentrasjonen) kan ikke påvises med et slikt forsøk.
Eksempel: Del en prøve (helst lav konsentrasjon) i to og tilsett kjent mengde (T) komponent til en

av dem. Prøve med og uten tilsatt komponent analyseres 5 ganger innen serie, og test utføres på differansen mellom middelverdiene. T-test for forskjell mellom middelverdier benyttes. Man tester om forskjellen mellom middelverdiene med (Ma) og uten (Mo) tilsatt komponent er signifikant (p=0.05) forskjellig fra tilsatt mengde:

$|Ma-Mo-T|/s\sqrt{(2/n)} > t_{2n-2,0.025}$ der n=5 og $t_{8,0.0.025}=2.306$. I dette tilfellet er forskjellen signifikant når forskjellen er > 1.46s. Man kan beregne % gjenfinning (recovery) etter formelen 100x(Ma-Mo)/T. 100% er ideelt. For at gjenfinning skal være signifikant forskjellig fra 100% på p nivå, må gjenfinning ligge i intervallet:

$100\% \pm t_{2n-2,0.025} x s\sqrt{(2/n)}/T$ (bare en omforming av formelen for konfidensintervall) eller i tilfellet over: 100 ± 146s/T. Hvis innen serie %CV er 1.0%, vil en gjenfinning mellom 98.5 og 101.5 % ikke være forskjellig fra 100 % på 5 % (signifikans) nivå.

Linearitet

Det kan være et problem å få testet hele måleområdet ved å sammenligne reelle prøver med referansemetode. Ved å velge en høy og en lav prøve, lage fortynninger ved å blande disse og måle dem, vil man på en enkel måte kunne få et klart bilde av linearitetet for analysen.

Eksempel: Velg en lav og en høy prøve, fortynn disse med hverandre til f.eks. 5 konsentrasjoner jevnt fordelt i måleområdet og mål disse i en serie. Plott resultatene og vurder linearitetet (f.eks. ved regresjonsanalyse - her testes ikke om dataene er lineære, dette gjøres ved visuell inspeksjon).

Interferenser

Matriks

Definisjon: Alle komponentene i en *prøve, unntatt komponenten.*

Der skal angis hvilke matrikser (prøvetyper) som er aktuelle for metoden. Det bør enten sannsynliggjøres at de aktuelle matrikser er av liten betydning for analyseresultatet, eller utføres forsøk for å finne det ut.

Interferens (spesifisitet)

Definisjon: Systematisk feil som følge av at komponenter i prøven andre enn komponenten påvirker analyseresultatet.

Data om interferens hentes hovedsakelig fra reagensprodusentens valideringsdata, men kan være aktuelt å utføre ved laboratoriet i spesielle tilfeller.

Eksempel: Tilsett kjent mengde interferent til en prøve. Prøve med og uten interferent analyseres 5 ganger innen serie, og test utføres på differansen mellom middelverdiene. T-test for forskjell mellom middelverdier benyttes. Man tester om forskjellen mellom middelverdiene med (Mi) og uten (Mo) interferent er signifikant (p=0.05) forskjellig fra 0*:* $|Mi-Mo|/s\sqrt{(2/n)} > t_{2n-2,0.025}$ der n=5 og $t_{8,0.025}=2.306$. I dette tilfellet er forskjellen signifikant når forskjellen er >1.46s. Man kan beregne gjenfinning etter formelen 100xMi/Mo. 100% er ideelt.

Intervallet i % for IKKE signifikant forskjell fra 100 er: 100±146s/Mo.

Man bør angi hva som anses som viktige forskjeller.

Man bør angi eller henvise til dokumentasjon som omhandler interferenser og viser deres innflytelse på analyseresultatet.

Holdbarhet

Holdbarhet er ikke direkte relatert til måleprosessen, men tas med her.

Holdbarhetstest 1 (enkel): Hvis prøvene kan oppbevares under betingelser hvor man vet de er holdbare (f.eks. frosset*),* kan n prøver (f.eks. 5) hver deles i to, en del oppevares slik at den er holdbar (0-prøven) og en del oppbevares i det tidsrom og under de betingelser som skal testes ut (test-prøven). På analysedagen måles 0-prøvene og test-prøvene sammen i en serie. Beregn gjennomsnittlig forskjell mellom testprøvene og 0-prøvene: M-Mo. Hvis innen serie variasjon er s (hentes fra annen kilde enn fra selve forsøket), vil forskjellen være signifikant på p nivå hvis

$|M-Mo|/s\sqrt{(2/n)} > z_{1-p/2}$ (bruker z i stedet for t fordi s er hentet fra annen kilde og har antatt sikker verdi, dvs n er stor).

Holdbarhetstest 2: Man kan benytte buksemetoden slik den er beskrevet i [5].

Man bør ha gjort seg opp en mening om hvilken endring som er viktig (se «Analytiske mål») og beregne nødvendig antall prøver i forsøket.

Holdbarhetsdata kan hentes fra litteraturen. I tilfeller der slike data er ufullstendige, eller ikke relevante, skal holdbarhet testes.

Litteratur

1. Kvalitetsledelse *og* kvalitetssikring. Terminologi. NS ISO 8402 2. utgave oktober 1994.
2. VIM: «International vocabulary of basic and general terms in metrology», 1993, (BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML).
3. General strategies to set quality specifications for reliability performance characteristics. C.G. Fraser. Scand J Clin Lab Invest, Vol. 59, No. 7, Nov. 99.
4. Kristian Linnet. Necessary Sample Size for Method Comparison Studies Based on Regression Analysis. Clin Chem 45:6, 882-894 (1999).
5. "Buksemetode". Sample stability: A suggested definition and Method of determination. Thiers et. al. Clin. Chem. 22/2, 176-83 (1976).
6. Statistiske vurderinger ved endring av analysemetode. Lars Mørkrid. Klinisk Kemi i Norden, nr 2, vol 10, 1998