Metodesammenligning

Ulrik Gerdes, Klinisk Biokemisk Laboratorium,
Center for Psykiatrisk Grundforskning, Risskov
E-post: ulrik.gerdes@dadlnet.dk

Baggrund

Vi bruger mange resurser indenfor klinisk biokemi på at sammenligne forskellige målemetoder og procedurer. Undersøgelserne anvendes fx hvis man overvejer at indføre en ny målemetode som erstatning for en eksisterende metode, eller overvejer at ændre en procedure for prøvebehandling.

Der findes forskellige matematisk-statistiske modeller (værktøjer), der kan bruges til dataanalyser i sådanne undersøgelser. Nogle modeller er gode, fordi de giver en tydelig karakterisering af forskellene i resultaterne, og dermed giver støtte til besvarelsen af de centrale spørgsmål: Hvor meget adskiller resultaterne fra to metoder sig fra hinanden? Er forskellene så store, at rekvirenterne vil få problemer med fortolkningen af resultaterne, hvis vi udskifter den ene metode med den anden? Er der behov for at omregne måleværdierne?

Andre modeller er mindre gode, eller ligefremt dårlige, fordi de ikke kan afsløre betydningsfulde forskelle mellem to metoder og/eller fordi resultaterne af metodesammenligningerne udtrykkes i aggregerede talstørrelser (parameterestimater), der ikke er meningsfulde eller brugbare i forhold til undersøgelsens formål. Det gælder især den alt for hyppigt anvendte lineære regression af resultater fra en metode Y mod resultaterne fra en metode X, med estimering af et intercept, en hældningskoefficient og en korrelationskoefficient.

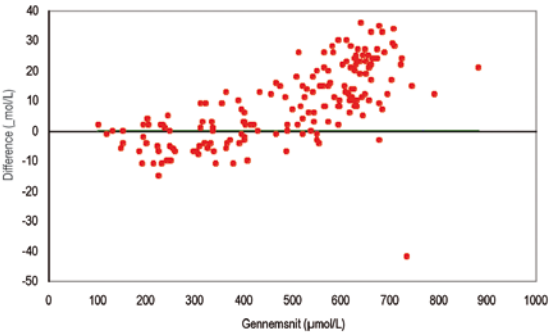
Vi vil gerne kvalitetsvurdere alt muligt indenfor klinisk biokemi, og vurderingerne burde også omfatte brugen af matematisk-statistiske modeller til metodesammenligninger. Det er vigtige undersøgelser, der ofte danner grundlag for væsentlige beslutninger i laboratoriet.

Jeg vil i denne artikel gennemgå brugen af Altman og Blands model, som baserer sig på grafiske præsentationer og enkle matematisk-statistiske analyser af differencer og middelværdier af måleresultater fra to forskellige metoder[1;2]. Udover at være en god model i den ovennævnte forstand, så er den også robust og let at implementere til brug i et regneark.

Gennemsnit og differencer for de individuelle måleværdier

Figur 1 viser resultaterne fra en sammenligning af målinger af P-Albumin på hhv. en Hitachi 917 (Roche) og en Aeroset (Abbott) autoanalysator (n=173). Jeg bruger 'A' og 'B' som henvisninger til de to metoder i det følgende, og bruger subskriptet 'i' for individuelle måleværdier.

Figur 1



Værdierne på abscissen (X-aksen) er gennemsnittet af de to målinger, dvs. værdierne $gns_i = (A_i+B_i)/2$. Man skal så godt som altid bruge gennemsnittene, også selvom den ene målemetode er en referencemetode, fordi der ellers introduceres en systematisk fejl i dataanalyserne[3].

Værdierne på ordinaten (Y-aksen) er differencerne mellem de to målinger, dvs. værdierne $d_i = (B_i$

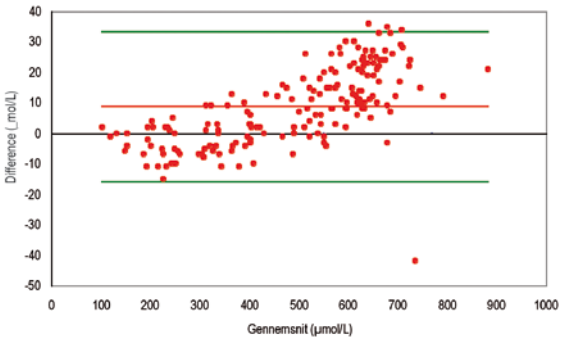
- A_i). Man kunne lige så godt bruge værdierne (A_i – B_i), men hvis B er den nye metode og A den gamle metode (eller eventuelt en referencemetode), så er fortolkningen af d_i mere umiddelbar: Værdier over 0 betyder at den nye metode giver højere værdier end den gamle, og vice versa.

Figur 1 viser fx at Aeroset har en systematisk tendens til at give lidt højere værdier end Hitachi 917 ved P-Albumin over 500 umol/L, mens der er en bedre overensstemmelse mellem de to metoder ved lavere koncentrationer.

Grænserne for overensstemmelse og middelværdien af differencerne

Figur 2 er mage til Figur 1, bortset fra at der er vist 3 vandrette linier:

Figur 2



Den midterste røde linie viser middelværdien af differencerne, dvs. middel(d) = $\sum d_i/n$. Værdien er 8,6 umol/L med et 95% konfidensinterval fra 6,7 til 10,5 umol/L (ikke vist i figuren). Intervallet beregnes som gennemsnittet(d) ± SE(d) x $t_{97.5}(n-1)$, hvor SE(d) er standardfejlen på estimatet af middel(d) = SD(d)/$\sqrt{n}$, og $t_{97.5}(n-1)$ er 97,5-percentilen i en t-fordeling med n-1 frihedsgrader[4]. Da konfidensintervallet ikke inkluderer 0 betyder det, at middel(d) er statistisk signifikant forskellig fra 0, hvis man som sædvanligt anvender et signifikansniveau på 5%.

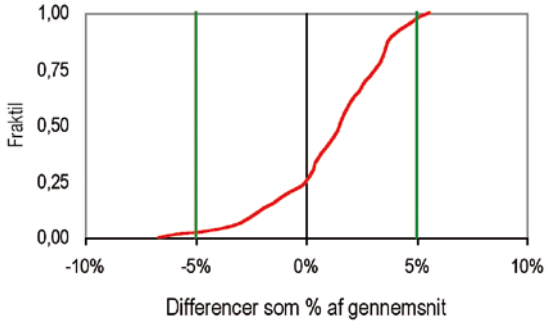
De to grønne linier er 95% grænserne for overensstemmelse *(limits of agreement)* mellem metoderne, dvs. de grænser indenfor hvilke i middel 95% af værdierne af d_i findes. Grænserne er estimeret som gennemsnit(d) ± 1,96 x SD(d). Bemærk, at der *ikke* er tale om konfidensintervaller, men om grænser der minder om referenceintervaller. Man kan eventuelt beregne konfidensintervaller for grænserne[2], men dette gøres sjældent i praksis.

Hvordan skal man nu fortolke disse resultater?

Et traditionelt udgangspunkt ville være at fokusere på det resultat, at middel(d) er statistisk signifikant større end 0 – og herudfra konkludere, at målemetode B giver højere værdier end målemetode A, dvs. at man har et problem, hvis man udskifter A med B.

Men er det en sufficient og faglig relevant fortolkning af resultaterne? Hvis man ser nærmere på værdierne af differenserne og på 95% grænserne for overensstemmelse, så er uoverensstemmelserne forholdsvis beskedne.

Figur 3



Figur 3 viser den kumulerede fordeling af differenserne i procent af de tilsvarende gennemsnit, dvs. værdierne 100 x d_i/gns_i. Fordelingen er højreforskudt i forhold til 0-linien, men ca. 95% af differenserne findes alligevel indenfor intervallet 0 ± 5%.

Hvis dette er acceptabelt, set fra en faglig, klinisk orienteret synsvinkel, så har man ikke noget problem. Og hér er en vigtig pointe, når man sammenligner målemetoder – man kan ikke bruge p-værdier eller andre aggregerede statistiske talstørrelser til at afgøre, hvorvidt eventuelle forskelle er betydningsfulde eller ej. Vurderingen er ikke en opgave for statistiker, men for en klinisk biokemiker!

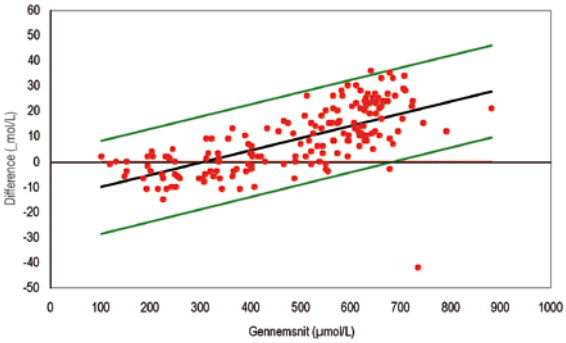
(Fortsætter side 26

(Fortsat fra side 25)

Analyser af 'skæve tal'

Figur 4 viser resultaterne af en anden analyse af data, som er mere korrekt i det aktuelle tilfælde, fordi der tydeligvis er en sammenhæng mellem differencerne og gennemsnittene af måleværdierne for de to metoder.

Figur 4



Det er gjort på følgende måde[2]:
Jeg har først estimeret parametrene β_0 og β_1 for den lineære regressionsmodel:

$$\textbf{forventet}(d_i) = \beta_0 + \beta_1 \times gns_i$$

Resultatet svarer til den sorte linie i Figur 4. Herefter har jeg beregnet residualerne for differencerne, dvs. $r_i = d_i - (\beta_0 + \beta_1 \times gns_i)$, og deres standardafvigelse, SD(r).
Når man herefter beregner værdier for

$$\beta_0 + \beta_1 \times gns \pm 1{,}96 \times SD(r)$$

fås de grønne 'skrå' grænser for overensstemmelse i Figur 4.

Man kan godt teste hypotesen $\beta_1 = 0$ som støtte for en vurdering af, om det er nødvendigt at bruge denne udvidede model, men som ovenfor er det i sidste ende den kliniske biokemikers vurdering af data der skal være afgørende.

Det er ikke helt sjældent, at differenserne spreder mere (eller mindre) ved høje måleværdier end ved lave værdier. Det kaldes variansheterogenecitet og giver problemer i alle parametriske modeller, hvis det er udtalt. Man kan i mange tilfælde dog afhjælpe problemet på en simpel måde, nemlig ved at logaritmere måleværdierne eller ved at bruge relative differenser, dvs. d_i/gns_i.

Hvis talværdierne er mere uregerlige, kan man bruge non-parametriske teknikker til karakteriseringen af overensstemmelsen mellem metoderne[2], men bør samtidig spekulere over *hvorfor* talværdierne er så uregerlige?

Hvor mange prøver skal man bruge til en metodesammenligning?

Der findes formler til beregning af det nødvendige antal, hvis man ønsker at kunne påvise en statistisk signifikant forskel af en given størrelse med en given statistisk styrke. Brugen forudsætter imidlertid at man har pålidelige estimater af metodernes varianser, og at fordelingen af differencerne er forholdsvis enkel[5]. NCCLS anbefaler mindst 40 prøver, dog uden at give nogen nærmere forklaring[6].

Mit råd, baseret på simpel erfaring, er – brug *mange*, helst over 75, og sørg for at værdierne er godt spredte i hele i måleområdet. Hvis man af økonomiske eller andre grunde bruger under 25 prøver, så risikerer man at få problemer med at afdække de reelle forskelle mellem målemetoderne.

Da vi skulle sammenligne en lang række gængse biokemiske analyser på en Hitachi 917 og en Aeroset, valgte vi at køre alle analyserne på omkring 175 tilfældigt udvalgte patientprøver. Det var rationelt, både med hensyn til analysearbejdet og håndteringen af data (der kunne hentes som blokke i instrumenternes computere), og strategien viste sig at give gode spredninger på alle målinger. Risikoen ved en sådan fremgangsmåde er dog, at man kun får et øjebliksbillede af forskellene mellem instrumenterne, dvs. ikke ser forskellene efter gentagne kalibreringer etc.

Enkeltbestemmelser, dobbeltbestemmelser eller flere?

Det anbefales at lave dobbeltbestemmelser eller flere, og at anvende den ekstra information i metodesammenligningen[2]. Det centrale er, at grænserne for hvor god en overensstemmelse der kan ses mellem to metoder naturligvis afhænger af metodernes "overensstemmelse med sig selv", dvs. deres præcision.

Skal man fjerne 'outliers'?

Nej, men hvis der findes datapunkter, som tydeligt skiller sig ud fra de øvrige, bør man gentage data-

(Fortsætter side 28)

(Fortsat fra side 26)

analyserne uden disse, for at se i hvilket omfang de påvirker vurderingen af overensstemmelsen mellem målemetoderne. Og i øvrigt undersøge, om der kunne være specielle grunde til at der optræder store forskelle, fx at den ene metode er mere følsom for visse interfererende substanser end den anden.

Hvis man vil korrigere resultaterne for en ny målemetode

Den viste sammenligning af målinger af P-Albumin på to forskellige instrumenter gav ikke anledning til seriøse overvejelser om at omregne (korrigere) resultaterne fra det nye instrument (Aeroset), for at undgå et lille spring i måleniveauet.

Hvis vi ville, kunne vi have brugt formlen:

B_i **(korrigeret)** $= \alpha_0 + B_i \times \alpha_1$

hvor α_0 og α_1 er parameterestimaterne fra en lineær regressionsanalyse af A_i som funktion af B_i.

Det kan også gøres på andre måder, men uanset hvordan man gør, så vær opmærksom på at der skal indregnes nye komponenter i et eventuelt usikkerhedsbudget for den korrigerede metode, nemlig usikkerhederne på parameterestimaterne. Det kan være lidt kompliceret[7].

En lille notits om 'parameter' og 'variabel'

Det er meget almindeligt, også blandt specialister i klinisk biokemi, at sige 'biokemisk parameter', fx med henvisning til P-Natrium. Jeg vil gerne benytte lejligheden til at forsøge at få udryddet denne forvirrende sproglige fejl – P-Natrium er noget man kan måle og vis værdi varierer fra individ til individ, så det er en biokemisk *variabel*. En parameter er noget helt andet, nemlig en kvantitet der karakteriserer en fordeling af måleværdierne af variable – fx en middelværdi, en standardafvigelse eller en regressionskoefficient.

Da kliniske biokemikere er berømte (eller berygtede) for at insistere på korrekte betegnelser indenfor vores eget område, så bør vi også følge andres råd:

> "Misuse of medical terms is rightly deprecated. Like other language errors it leads to confusion and the loss of valuable distinction. Misuse of nonmedical terms should be viewed likewise"[8]

Referencer

(1) Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1(8476):307-310.

(2) Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999; 8(2):135-160.

(3) Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet 1995; 346(8982):1085-1087.

(4) Statistics with confidence. 2nd ed. London: British Medical Journal Books; 2000.

(5) Magari RT. Evaluating agreement between two analytical methods in clinical chemistry. Clin Chem Lab Med 2000; 38(10):1021-1025.

(6) Krouwer JS, Tholen DW, Garber CC, Goldschmidt HMJ, Kroll MH, Linnet K et al. Method comparison and bias estimation using patient samples; approved guideline - second edition. EP9-A2[22]. 2002. NCCLS; Wayne, Pennsylvania, USA.

(7) Borg L, Kristiansen J, Christensen JM, Jepsen KF, Poulsen LK. Evaluation of accuracy and uncertainty of ELISA assays for the determination of interleukin-4, interleukin-5, interferon-gamma and tumor necrosis factor-alpha. Clin Chem Lab Med 2002; 40(5):509-519.

(8) Altman DG, Bland JM. Statistics notes: variables and parameters. BMJ 1999; 318(7199):1667.

Taksigelse

Statistiker, ph.d. Jørn Attermann, Institut for Epidemiologi og Socialmedicin, Århus Universitet, takkes mange gange for rettelser og kommentarer til manuskriptet. ■