

Nasjonalt prosjekt for standardisering av holdbarhetsforsøk

Innledning

I 2014 ble det oppnevnt en arbeidsgruppe for holdbarhetsforsøk i regi av Norsk Klinisk-kjemisk kvalitetssikring (NKK), Bioingeniørfaglig institutt (BFI) og Norsk selskap for medisinsk biokjemi (NSMB).

Arbeidsgruppen har bestått av:

Kristin M Aakre (NKK), leder

Pål Rustad (NKK)

Heidi Eilertsen (BFI)

Torill Kalfoss (BFI)

Arne Åsberg (NSMB)

Ann Helen Kristoffersen (NSMB)

Gruppens mandat var å utarbeide en felles norsk protokoll for holdbarhetstesting som senere kunne gjøres tilgjengelig for det norske miljøet på NKK sine nettsider og det langsiktige målet var at holdbarhetsforsøkt utført ved norske laboratorier kunne gjøres tilgjengelig på nettsidene slik at informasjonen ble delt.

Gruppen har hatt jevnlig telefonmøter i løpet av siste halvår 2014 og første halvår 2015 og har utarbeidet en protokoll med tilhørende dataverktøy i Excel. Protokoll og dataverktøy ble testet ut på NKKs workshop på NKK møtet i 2015, og fikk gode tilbakemeldinger fra deltagerne. Protokoll og verktøy er justert i forhold til de innspill som kom på NKK møtet.

Arbeidsgruppen anbefaler nå at protokoll og dataverktøy (se vedlegg) gjøres allment tilgjengelig på nettsiden.

Protokoll for holdbarhetsforsøk

1. Definisjon av prøvematerialets holdbarhet

Arbeidsgruppen mener at holdbarhet er prøvematerialets evne til å beholde opprinnelig måleverdi innenfor definerte grenser når det lagres under definerte forhold i et definert tidsrom (1).

Med en holdbarhetsundersøkelse vil man undersøke om analyseresultatet av en komponent under oppbevaring endres så lite at det fortsatt kan brukes til formålet. Det er naturlig å studere både gjennomsnittlig endring i et visst antall prøver og endring i den enkelte prøve. Dermed får vi to kriterier for holdbarhet:

1. Gjennomsnittskonsentrasjonen skal endre seg mindre enn utgangsgjennomsnittet \pm tillatt systematisk avvik (tillatt bias).
2. Konsentrasjonen i den enkelte prøve skal endre seg mindre enn utgangsverdi \pm tillatt totalfeil.

Å bestemme tillatt bias og tillatt totalfeil er ingen enkel sak. I sjeldne tilfeller har helsemyndigheter og/eller fagorganisasjoner oppgitt grenser som man kan følge. Ellers anbefaler arbeidsgruppen at man bruker angivelser som er basert på vurdering av klinisk konsekvens av feilmåling, og - hvis slike vurderinger ikke er publisert - på data om biologisk variasjon. Prinsipper for å bestemme tillatt feil basert på data om biologisk variasjon er beskrevet av Fraser (2), mens tabeller over tillatt bias og tillatt totalfeil basert på slike data finnes i Ricos tabeller (3). Der er det også gitt referanser til originalartikler, som bør vurderes fordi beregningsgrunnlagets kvalitet er varierende.

Noen definisjoner brukt i dokumentet:

Sant gjennomsnitt: Den gjennomsnittskonsentrasjonen prøvematerialet har ved gitte betingelser og som vi ville målt i et forsøk som inkluderte et uendelig antall prøver.

Målt gjennomsnitt: Det gjennomsnittet vi måler i forsøket vårt og som alltid inkluderer en usikkerhet.

2. Planlegging av holdbarhetsforsøk

I planleggingen av holdbarhetsforsøk må man tenke igjennom hvilke faktorer som er mest nyttige å teste. Dette kan være:

Prøvemateriale

Tidsintervall mellom prøvetaking og analysering

Oppbevaring ved ulike temperaturer (for eksempel kjøleskap, romtemperatur, temperatur over 25 grader)

Målemetode / Instrument

Reagens

Transport til laboratoriet (for eksempel postgang, hentetjeneste, rørpost)

Frysing (eventuelt ved ulike temperaturer)

Tining - flere sykluser

Sentrifugering – sentrifugeringshastighet, tid fra prøvetaking til sentrifugering, temperatur under sentrifugering

Prøverør

Oppbevaring i originalrør (med eller uten gel) eller avpipettert prøvemateriale

Luftfuktighet

Bruk av tørris

pH

Interferenter (f.eks. hemoglobin, lipemi, icterus og antikoagulasjon) som kan påvirke holdbarheten

Egenskaper hos pasienten (f.eks. ulike sykdomstilstander) som kan påvirke holdbarheten

De faktorene man mener er mest relevant for rutinedrift eller vitenskapelig (avhengig av hensikten med forsøket), bør inkluderes i forsøket. Man bør i størst mulig grad simulere en reell situasjon. Man bør alltid inkludere prøvemateriale med ulike konsentrasjoner (både fra friske og syke individer) slik at man får testet holdbarhet i hele det aktuelle måleområdet. Spesielt viktig er det å teste holdbarheten ved kliniske beslutningsgrenser.

3. To ulike metoder for å teste holdbarheten

3.1. Batch-metoden

Denne metoden egner seg spesielt hvis analytten er holdbar ved en spesiell oppbevaring (f.eks. nedfrysing). Man samler inn prøver fra ulike individer og deler prøvematerialet i like mange porsjoner som antall kombinasjoner av oppbevaringsbetingelser og oppbevaringstider. Når hver porsjon har vært oppbevart ved de gitte betingelser i den tilmålte tid, settes porsjonen i fryseboks, gjerne $-70\text{ }^{\circ}\text{C}$ eller kaldere, for å hindre ytterligere endringer. Alle porsjoner fra samme person analyseres samlet (i én «batch») for å minimere effekten av analytiske dag-til-dag-variasjoner. Hvis analytten ikke er holdbar i dypfrost prøvemateriale (for eksempel partikkelkonsentrasjonen av celler), kan man justere for analysemetodens faktiske dag-til-dag-variasjoner ved å etablere dagens gjennomsnittsnivå i et stabilt kontrollmateriale og justere analyseresultatene tilsvarende¹. Alle måleresultater omregnes (standardiseres) til prosent av utgangsverdien, som settes lik 100 %, men rådata bør også vurderes, se nedenfor.

3.1.1. Vurdering av gjennomsnittsendringer

For hvert oppbevaringstidsrom estimerer man det sanne gjennomsnittet ved å måle et gjennomsnitt og gjennomsnittets 90 % konfidensintervall. Konfidensintervallets grenser sammenliknes med akseptområdet, som er $100\% \pm$ tillatt bias. Man har 3 mulige utfall:

1. Hele konfidensintervallet ligger i akseptområdet. I så fall er dette holdbarhetskriteriet oppfylt. Det er minst 95 % sikkert at sann gjennomsnittlig endring er mindre enn tillatt bias.

¹ Hvis kontroller analyseres i hver serie, kan alle resultater i hver serie multipliseres med faktoren Mt/Ms der Mt er middelverdien av alle kontrollene i alle seriene og Ms er middelverdien av seriens kontrollverdier.

2. Hele konfidensintervallet ligger utenfor akseptområdet. Da er prøvematerialet ikke er holdbart. Det er minst 95 % sikkert at sann gjennomsnittlig endring er større enn tillatt bias.

3. Konfidensintervallet krysser en eller begge akseptgrensene. Dette er et tvilstilfelle, som eventuelt kan avklares med å undersøke flere prøver.

At bruk av 90 % konfidensintervall gir minst 95 % sikkerhet i punkt 1 og 2 skyldes den ensidige testsituasjonen, dvs. at når den ene enden av 90 % konfidensintervallet til gjennomsnittet tangerer en av grensene for akseptområdet, er det bare 5 % sannsynlighet for at den sanne (men ukjente) gjennomsnittlige endring ligger på den andre siden av grensen.

3.1.2. Vurdering av endringer i den enkelte prøve

I tillegg til gjennomsnittsverdien for hver oppbevaringstid må man vurdere de enkelte måleresultatene, om de ligger innen akseptområdet, som er $100\% \pm$ tillatt totalfeil. Det er rimelig å kreve at 95 % av enkeltverdiene skal ligge innenfor grensene for tillatt totalfeil for hver oppbevaringstid. Det er nyttig å vurdere rådata for bedre å vurdere om eventuelle brudd på enkeltverdi-kriteriet er knyttet til prøver med konsentrasjoner av liten klinisk interesse eller der analysemetoden er minst presis. Hvis analysemetoden relativt er minst presis når konsentrasjonen er lav, og brudd på enkeltverdi-kriteriet kun forekommer i prøver med lav konsentrasjon, kan man vurdere om holdbarhetskriteriet bør justeres (gjøres mer romslig) for slike prøver, spesielt hvis prosentvis økt feilmåling ved lav konsentrasjon har liten eller ingen klinisk betydning. Vurdering av rådata er også viktig for å se om prøverekken fra en bestemt pasient ikke oppfyller enkeltverdi-kriteriet på bare ett, relativt tidlig, tidspunkt, men gjør det på alle andre tidspunkter. Slike resultater kan være et uttrykk for tilfeldige enkeltfeil, og de bør ikke få avgjørende betydning.

3.1.3. Samlet vurdering

For en bestemt oppbevaringsbetingelse må *begge* krav, både kravet til gjennomsnittsverdien og kravet til enkeltverdiene, være oppfylt for at prøvematerialet kan kalles holdbart.

3.1.4. Antall prøver

Med flere prøver får man smalere 90 % konfidensintervall rundt det målte gjennomsnittet og mindre sannsynlighet for feil konklusjon. Av hensyn til vurdering av gjennomsnittsendringer kommer man ofte langt med 10-30 ulike prøver, men av hensyn til å vurdere endringer i enkeltprøver skulle man ønske langt flere prøver. Av praktiske grunner foreslår arbeidsgruppen at antall prøver hovedsakelig tar hensyn til endringer i gjennomsnittet, men at det *ikke settes lavere enn 20* av hensyn til vurdering av enkeltverdier. Skulle det vise seg, etter at de 20 prøvene er undersøkt, at det målte gjennomsnittet ligger i akseptområdet, men 90 % konfidensintervallet krysser en av akseptgrensene, kan forsøket utvides med flere prøver for å innsnevre konfidensintervallet. Hvis gjennomsnittskriteriet er oppfylt, men mer enn 1 av 20 enkeltverdier ligger utenfor tillatt totalfeil, bør man også vurdere å utvide forsøket.

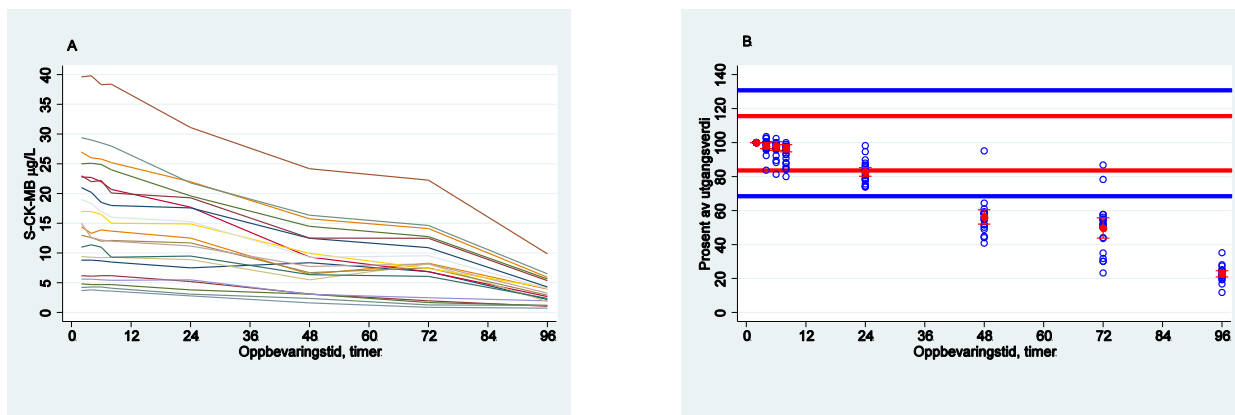
3.1.5. Eksempel

For s-CK-MB massekonsentrasjon ønsker man å undersøke holdbarhet ved romtemperatur, og velger å basere holdbarhetskriteriene på data om biologisk variasjon. Tillatt bias er 16,0 % og tillatt totalfeil er 31,2 % (3). Man vil bruke blodprøver fra 20 pasienter, som velges slik at s-CK-MB i prøvene fordeler

seg noenlunde jevnt i det klinisk relevante området. Serum fordeles i en serie av rør merket med oppbevaringstid, som er 2, 4, 6, 8, 24, 48, 72 og 96 timer etter *prøvetaking*, til sammen 8 porsjoner fra hver pasient. Konsentrasjonen i den porsjonen som er oppbevart i 2 timer etter *prøvetaking*, regnes som utgangsverdi, siden dette er den korteste, standardiserte oppbevaringstid man kan prestere. Ved endt oppbevaringstid i romtemperatur nedfrysres hvert fordelingsrør ved $-80\text{ }^{\circ}\text{C}$, og oppbevares ved den temperaturen inntil alle prøver er innsamlet og kan analyseres i samme analyseserie. Dersom det ikke er mulig å analysere alle prøvene i en serie, må i alle fall alle prøvene fra en pasient analyseres i samme serie. Man antar at ingen endring av s-CK-MB massekonsentrasjon finner sted ved $-80\text{ }^{\circ}\text{C}$.

Etter opptining og analysering vurderes resultatene grafisk i forhold til grensene for tillatt bias og tillatt totalfeil, som vist i figur 1A (rådata) og 1B (resultatene i prosent av utgangsverdi).

Resultatene tolkes slik: S-CK-MB er holdbar i prøver oppbevart i inntil 8 timer i romtemperatur, fordi i inntil 8 timers oppbevaring ligger hele 90 % konfidensintervallet for det målte gjennomsnittet i akseptområdet $100\% \pm$ tillatt bias og minst 95 % av alle enkeltverdier ligger i akseptområdet $100\% \pm$ tillatt totalfeil. For prøver oppbevart i 24 timer er kravet til enkeltverdiene oppfylt, men ikke kravet til gjennomsnittet. Ved å utvide prøveantallet kan man innsnevre konfidensintervallet og muligens tilfredsstille kriteriet, men sannsynligheten er liten da det målte gjennomsnittet ligger utenfor akseptgrensen. For de tre andre tidsrommene oppfylles ikke noe kriterium.



Figur 1. S-CK-MB i 20 prøver oppbevart i ulike tidsrom i romtemperatur. A: Rådata plottet mot oppbevaringstid. B: Resultater omregnet til prosent av utgangsverdi og plottet mot oppbevaringstid. De røde punktene viser gjennomsnitt (i prosent av utgangsverdi) for hvert tidspunkt, mens de røde, loddrette intervallene er 90 % konfidensintervall for gjennomsnittet, og de røde, horisontale linjene viser tillatt bias. De blå symbolene markerer enkeltverdier i prosent av utgangsverdi, og de blå, horisontale linjene viser tillatt totalfeil.

3.2. Buksemetoden

Denne metoden anbefales først og fremst brukt når man er ukjent med om materialet kan oppbevares stabilt for referanse (f.eks. frysing). Prøver hentes fra rutinen slik at pasientprøver med patologiske verdier er lett tilgjengelig. Det ligger i metoden at prøvene må inkluderes på ulike tidspunkt (dager) og slik kompenseres dag til dag variasjon for analysemetoden. Rutineanalysen bør utføres relativt ofte og testen egner seg best for testing av relativt få modaliteter (f.eks. kun ulike tidsintervall, men med ellers like betingelser) for holdbarhet.

Metoden slik den beskrives her forutsetter at analytisk standardavvik (s) (eller variasjonskoeffisient, CV)

og kvalitetskrav for bias og tillatt totalfeil er kjent.

3.2.1. Beskrivelse

Analyser en fersk prøve fra rutine (prøve 1) og registrer resultatet. Sett prøven ved de betingelser som skal testes (f.eks. romtemperatur på benk i 3 dager). Neste dag analyseres en ny prøve (prøve 2) fra rutinen og settes ved testbetingelsene. Slik fortsetter utplukk av en prøve hver dag fra rutinen til forsøket avsluttes. Etter at prøve 1 har stått ved testbetingelsene i den tid som skal testes, analyseres denne på nytt og forskjellen fra første måling dividert på analytisk standard avvik beregnes ($d1/s$). Tilsvarende gjøres for prøve 2, 3 osv. etter at de har stått ved testbetingelsene og for hver dag beregnes $d2/s$, $d3/s$ osv. Dersom analysen kun utføres på virkedager kan man ta pause for inklusjon i perioden fredag til søndag uten at dette vil påvirke resultatene i stor grad.

Etter hvert som forsøket gjennomføres, føres summen av standardiserte avvik, $d1/s + d2/s + d3/s$ osv, i et skjema der det er markert forhåndsbestemte grenser, en grense for beslutningen *holdbar*, og en for beslutning *ikke holdbar* (disse to grensene ser til sammen ut som en bukse, derav navnet buksemetoden). Forsøket for test av gjennomsnittlig bias pågår inntil linjen som beskriver summen av alle d/s 'ene krysser en av grensene. Signifikansnivået for testen er $\alpha = 0,05$, dvs. at det bare er 5 % sjans for at man får konklusjonen *ikke holdbar* når det ikke er noen forandring (denne metoden er også designet slik at det bare er 5 % sjans for at man får konklusjonen *holdbar* når analysen egentlig er *ikke holdbar*, $\beta = 0,05$).

Hvis man har 1 s som kvalitetskrav for bias, vil forsøket kreve 15 – 38 prøver for å få konklusjonen *holdbar*, men forsøket kan i prinsippet avsluttes etter at første prøve er analysert hvis konklusjon er *ikke holdbar* (man bør imidlertid alltid analysere minst 20 prøver før noen konklusjon trekkes fordi man trenger så mange prøver for å få en rimelig vurdering av enkeltavvik – se under).

I sin originalversjon (4) anbefales en grense på ett (evt. 2) analytiske standard avvik (eller CV) som biasgrense for *holdbar/ikke holdbar*. Kvalitetskrav for gjennomsnittlig endring skal selvsagt velges uavhengig av hvilket statistisk verktøy man bruker, og i den versjonen av buksemetoden som er publisert på NKKs hjemmeside kan man selv velge kvalitetskrav. Dette er implementert etter anvisning fra Schneiderman et al. (5).

Fordi det i batch-metoden beskrevet over er tatt med en vurdering av enkeltavvik for hver prøve, er det også for denne metoden tatt med en slik vurdering. Tillatt totalfeil registreres i regnearket til høyre for TEa. Enkeltavvikene for prøvene er vist i et eget plot hvor tillatt totalfeil er avsatt som vannrette røde grenser. I rad 6 er vist hvor mange % av resultatene som har et avvik større enn tillatt totalfeil. Det kreves at ≥ 95 % av avvikene er mindre enn tillatt totalfeil for at prøvene skal være holdbare.

3.2.2. Eksempel

For at man skal kunne bruke dette regnearket må man tillate bruk av makroer i Excel (se ”Fremgangsmåte” i regnearket).

For å gjøre det enkelt her, beskrives bare hvordan testen for 2 dager i kjøleskap utføres.

Fig 2. Registrering av analytisk standard avvik (eller CV) og kvalitetskrav.

Konstant analytisk CV for alle	S	1.00
		Krav 1 1.00
N		

Først må man svare på om analysen har ”Konstant analytisk CV for alle konsentrasjoner (J/N)” (Fig. 2) – antas CV å være mer konstant enn standardavvik, svarer man J (ja), ellers N (nei) – ved tvil, svar N². Som standardavvik eller CV benyttes laboratoriets totale variasjon (langtidsvariasjon).

Man må deretter registrere analytisk standardavvik hvis man har svart N på det forrige spørsmålet eller CV (husk å skrive % bak) hvis man har svart J.

Deretter må man velge ett eller to kvalitetskrav for bias som definerer hva som skal til for at analysen skal være holdbar. I eksempelet er kvalitetskravene for bias valgt til det samme og det dobbelte av det analytiske standard avvik for analysen (hhv. Krav 1 og Krav 2) og tillatt totalfeil (TEa) til 4 analytiske standard avvik. Kvalitetskravene oppgis i samme enhet som S (analytisk standardavvik)

Deretter starter forsøket:

Dag 1: Prøve 1, tappet samme dag, blir tatt ut av rutinen og resultatet, dvs. 4, blir registrert i celle B8 (fig. 3). Prøven blir deretter satt i kjøleskap.

Dag 2: Neste dag blir en ny prøve, prøve 2, tappet samme dag, plukket fra rutine, resultat registreres i celle B9 og prøven blir satt i kjøleskap.

Dag 3: Prøve 1 oppbevart i kjøleskap i 2 dager reanalyseres i rutine og resultat blir registret i kolonne C8. Hvis man skal teste ved flere tidsrom, settes prøven tilbake i kjøleskapet, ellers kastes prøven. Regnearket beregner d1/s til -1.0 (celle D7 i fig. 3) som plottes (Fig. 4). Prøve 3 tas ut fra rutine og resultat registreres i celle B10. Prøven blir satt i kjøleskap.

Dag 4: Prøve 2 oppbevart i kjøleskap i 2 dager reanalyseres i rutine og resultatet registreres i celle C8, summen av de to d/s beregnes i celle D8 til 0,0. Prøve 4 tas fra rutine, resultatet registreres i celle B11 og prøven settes i kjøleskap.

Slik fortsetter man til linjen som plottes (fig. 4) krysser enten den røde grensen med konklusjon *ikke holdbar* eller den grønne grensen med konklusjon *holdbar*.

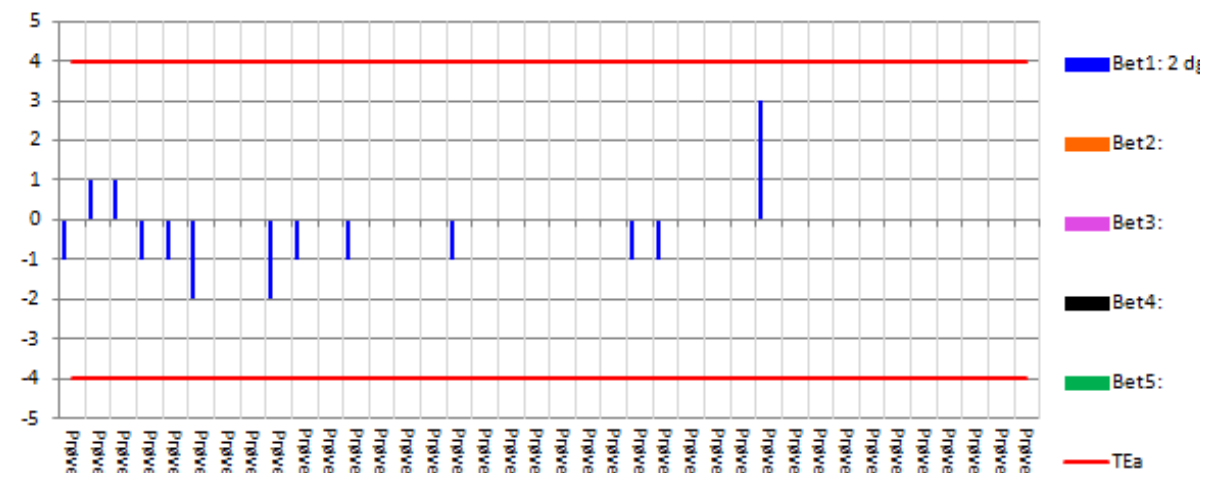
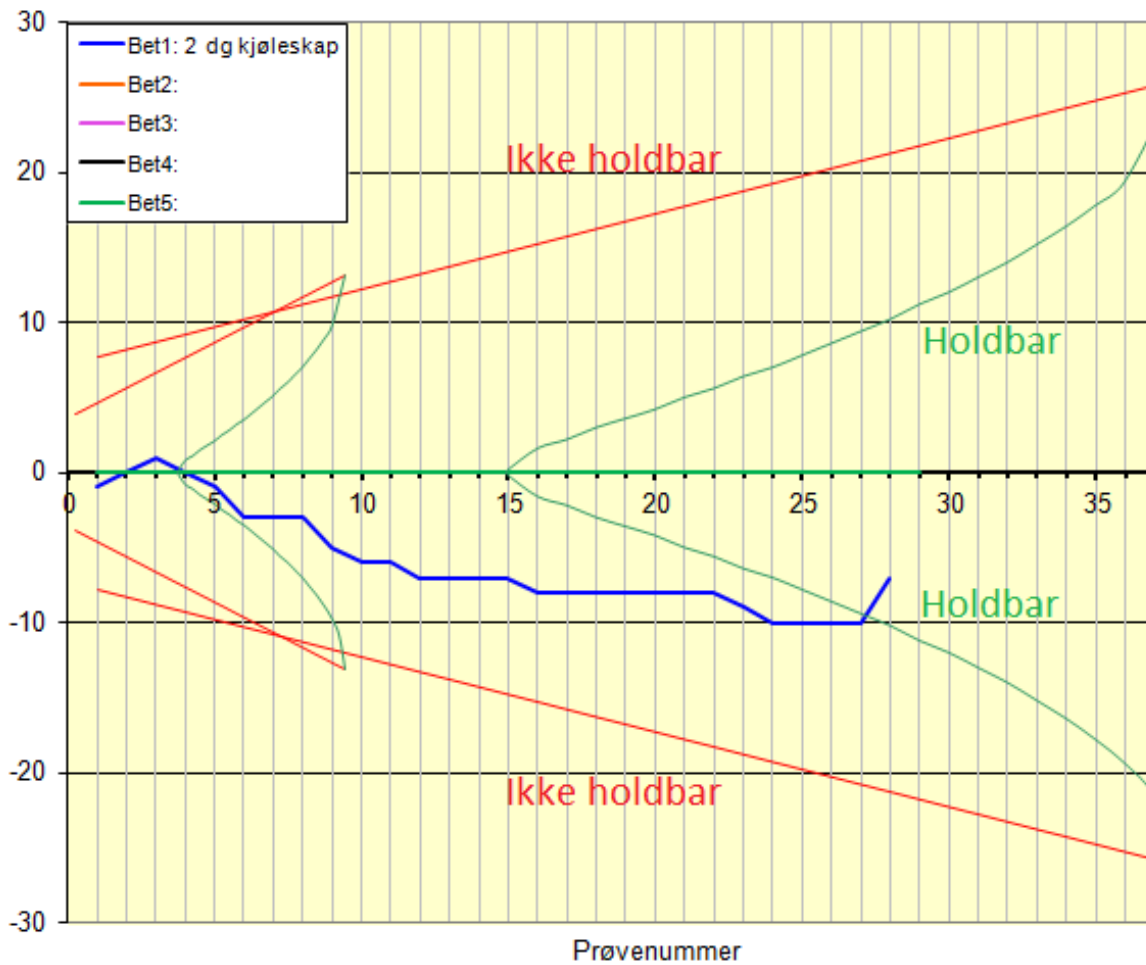
Hvis man ønsker å inkludere enkeltresultaters avvik i forhold til tillatt totalfeil, bør man fortsette forsøket til minimum 20 prøver er analysert. Nederst i fig. 4 ser man at ingen av de 28 enkeltavvikene har overskredet kravet for tillatt totalfeil (TEa).

Fig 3. Dataregistrering for buksemetoden. Ref: Resultat 1. analysedag, R: Resultat 2. analysedag, Kum: sum av standardiserte avvik.

	A	B	C
1	Holdbarhet, buksemetode		
2			
3		0-still plott	
4			
5	Prøve nr	Bet1: 2 d i kjøleskap	
6	%>TEa	0 %	
7		Ref	R
8	Prøve 1	4	3
9	Prøve 2	5	6
10	Prøve 3	4	5
11	Prøve 4	6	5
12	Prøve 5	8	7
13	Prøve 6	6	4
14	Prøve 7	3	3
15	Prøve 8	5	5
16	Prøve 9	7	5
17	Prøve 10	6	5
18	Prøve 11	7	7
19	Prøve 12	15	14
20	Prøve 13	12	12
21	Prøve 14	9	9
22	Prøve 15	13	13
23	Prøve 16	15	14
24	Prøve 17	9	9
25	Prøve 18	7	7
26	Prøve 19	12	12
27	Prøve 20	15	15
28	Prøve 21	12	12
29	Prøve 22	15	15
30	Prøve 23	9	8
31	Prøve 24	7	6
32	Prøve 25	13	13
33	Prøve 26	20	20
34	Prøve 27	15	15
35	Prøve 28	17	20

² Man kan også evt. dele opp forsøket ved å velge konsentrasjoner i et område med med konstant s eller CV. Fra originalartikkelen omtales imidlertid valget på denne måten: ”These two models (konstant s eller konstant CV) were used in computer simulation studies... and it was found that the additive model (konstant s) was quite effective for all practical uses.”

Figur 4. Eksempel på resultater av holdbarhetstesting for eksempelet over (2 dg i kjøleskap), der konklusjonen er *holdbar* etter testing av 4 prøver mot Mål 1 (den minste buksa), og *holdbar* etter test med 28 prøver mot Mål 2 (den største buksa). Nederst er vist enkeltavvik hvor maksimalt 5 % av avvikene kan være utenfor de røde grensene (TEa; tillatt totalfeil)



4. Referanser

1. Åsberg A, Solem KB, Mikkelsen G. Prøvematerialets holdbarhet – kriterier og vurderinger. *Klinisk Biokemi i Norden* 2011;23(4):34-8.
2. Fraser CG. *Biological variation: From principles to practice*. Washington: AACC Press, 2001.
3. Desirable Specifications for Total Error, Imprecision, and Bias, derived from intra- and inter-individual biologic variation. <http://www.westgard.com/biodatabase1.htm> (07.11.2014).
4. Thiers RE, Wu GT, Reed AH, Oliver LK. Sample stability: a suggested definition and method of determination. *Clinical Chemistry* 22: 176-183, 1976.
5. Schneiderman, M.A. Armitage, P. A family of closed sequential procedures. *Biometrika* 49, 41 (1962). <http://www.clinchem.org/cgi/content/abstract/22/2/176>