TEKNOLOGISK
INSTITUT
DMRI

Rapport

Vurdering af Meyn Footpadinspection System

Marchen Hviid

29. marts 2016
Proj.nr. 2003830
Version 05
MAHD/HNH

Sammendrag

Baggrund og formål

To danske fjerkræslagterier, her anført som: Slagteri A og Slagteri B har i 2013 købt Meyn Footpadinspection System til bedømmelse af trædepudesvidninger.

Projektets formål var at afprøve systemets evne til at klassificere slagtekyllingefødder, jf. den danske tre-trins-skala for trædepudeskader, og på den baggrund vurdere fordele og ulemper i forhold til den nuværende manuelle stikprøvekontrol.

Resultater – daglig drift

- Meyn-udstyret er driftsikkert og leverer tilstrækkelige data til beregning af en samlet flokscore til velfærdsplacering.
- Flokscoren fra Meyn-udstyret var generelt noget lavere end veterinærernes score i de undersøgte flokke.
- Der er dog mangler i dataintegration og udstyret er ikke sat op til at modtage information om flok-ID.

Resultater: 1-1 sammenligning

Begge fødder fra 320 kyllinger blev bedømt af to veterinærer fra Kødkontrollen, Meyn-udstyret og en reference. Meyn-udstyret gav lavere score (flere fik 0) end både referencen og veterinærerne. Graden af enighed mellem de manuelle bedømmelser var i gennemsnit 0,65.

Retningslinjer for bedømmelsen

Der er ikke fuldstændig overensstemmelse mellem, hvordan bedømmelsen foretages af udstyr (kun central trædepude) og veterinær (hvor skader på tæer og overgang mellem trædepude og tå også indgår). Det påvirker selvfølgelig graden af enighed mellem udstyr og veterinær.

Fordele ved udstyr

- Udstyret er driftsikkert og kan efter en optimering via projektet, med den nuværende opsætning og algoritme måle den centrale trædepude på 85-90% af alle kyllinger i en flok.
- Niveauet af alvorlige trædepudesvidninger er lavt i Danmark, og det kan derfor være en fordel, *at korrekt tilbagemelding og tidlig registering af stigende forekomst*, foregår på grundlag af målinger foretaget på en stor andel af de leverede kyllinger.
- Anvendes en optimal algoritme i visionsystemet kan hovedparten af alle trædepuder måles og vurderes efter de samme retnings-linjer; fremfor en lille stikprøve visuelt bedømt af eksperter.

Forbedringsmuligheder

- Algoritmerne i udstyret er udviklet på basis af manuelle bedømmelser, og der er fundet usikkerhed i hvordan reference-målingerne skal foretages, og hvorledes skala og bedømmelse skal

anvendes på kyllingefødderne. Det afspejler sig også i de niveauforskelle, som dette projekt har påvist.
- Udstyret er ikke endnu fuldt integreret i slagteriernes datasystem, og driftsdata fra udstyret kan derfor først benyttes i fuld omfang, når dette er på plads.

Anbefalinger

Meyn Footpadinspection System er et godt bud på et udstyr, som kan støtte veterinærkontrollen og slagterierne i at forbedre dyrevelfærden hos slagtekyllinger ved at monitere trædepudeskader og give feedback til producenten på et større datagrundlag for de enkelte kyllingeflokke.

Dette projekt har vist, at før udstyret eventuelt kan erstatte eller supplere den nuværende manuelle stikprøvekontrol, skal følgende iværksættes:

Installation

- Der skal installeres kommunikation mellem Meyn-udstyret og slagteriernes øvrige datasystemer. Ellers kan data ikke benyttes til de enkelte flokkes score for trædepudesvidninger, da dataopsamlingen bliver for usikker og med risiko for fejl.
- Der skal udarbejdes rutine for daglig kontrol og dataopsamling, herunder et fantom/referencemateriale til kontrol af udstyret.
- Der kan udvikles/installeres et program, som sikrer at et udsnit af billeder fra hver flok gemmes, hvis der senere skulle opstå tvivl om flokkens korrekte status.

Det skal vurderes, om udstyret med fordel skal gemme enkeltmålere-sultat af hver fod i databasen fremfor at anvende ringeste score ved forskel mellem fødder.

Trædepude-score

Der skal etableres konsensus om udførelse af referencebedømmelse af trædepudesvidning under danske forhold, så der er enighed om, hvilke niveauer og typer af svidninger Meyn-udstyret skal kalibreres til at måle. Undersøgelsen bør afklare:

- Er der forskel mellem 2-D og 3-D (dybde vurdering ved indsnit) i scoring for flokniveau og dermed opfølgningsfrekvens i problembesætninger.
- Kan arealberegning af svidning i forhold til trædepudeareal supplere/erstatte den nuværende scoring?
- Frekvens af tåsvidning og svidning i kant mellem trædepude og tæer samt mulig indflydelse af tå-/kantsvidning på den samlede flokscore.
- Bidrager svidninger på tæer i tilstrækkeligt omfang til at indgå i bedømmelse og udstyrets måling, i forhold til den kompleksitet det medfører fremfor kun at fokusere på en robust måling alene af trædepuden.

Algoritmer

- Der skal udvikles nye endelige algoritmer som tilpasses danske forhold og definitioner, herunder beslutte om/hvordan tåsvidninger skal håndteres.
- De nyudviklede algoritmer testes efter de samme 1:1 retningslinjer som anvendt i dette projekt i forhold til endelig reference definition.

Sortering

Meyn-udstyret er med den nuværende placering i den urene ende, ikke kaliberet til at bedømme 'rene' fødder, efter epidermis er fjernet.

Hvis Meyn-udstyret skal anvendes til tidlig frasortering af fødder efter kvalitet i forhold til markedskrav, skal det undersøges, om de installerede algoritmer er tilstrækkelige, eller om der kan udvikles nye algoritmer, hvor referencen er fødder bedømt, efter epidermis er fjernet.

Indhold

Baggrund og formål

For at forbedre dyrevelfærden i slagtekyllinger indførte Danmark i 2001 bedømmelser af trædepudesvidninger. Bedømmelsen er senest reguleret i bekendtgørelse nr. 757 af 23/6/2010, §11 stk. 1 og 2. Bedømmelsen blev indført efter inspiration fra Sverige, og det er da også det videnskabelige arbejde udført af Charlotte Ekfält (Foot-Pad Dermatitis in Broiler and Turkeys, Doctoral Thesis, 1998), som danner baggrund for de danske bedømmelser.

I dag foretager Kødkontrollen under Fødevarestyrelsen (FVST) en manuel stikprøvekontrol af kyllingefødder for trædepudeskader pr. flok leveret til de danske kyllingeslagterier (50 fødder fra den første tredjedel af flokken og 50 fødder fra den sidste tredjedel), og resultatet i form af en flokscore danner grundlag for opfølgning over for producenten.

Veterinær score

Bedømmelse af trædepudesvidning foretages på afskårne fødder efter vask og skoldning af den hele kylling, men uden at epidermis er fjernet. Fødderne bliver inddelt i kategorierne 0, 1 eller 2, afhængigt af graden af de konstaterede svidninger. Ud fra bedømmelserne beregnes en holdscore, hvor antal fødder med score 1 tildeles ½ point og antal fødder med score 2 tildeles 2 point.

Flok score

Derefter udregnes et flokscore (FC) – baseret på de enkelte fødders score:

FC = ½*antal(1) + 2*antal(2).

Denne formel anvendes også i Meyn-udstyret til at beregne flokscore, her vil det dog være procent af kyllinger med score 1 eller score 2 som indgår.

Flokscoren benyttes i opfølgning til producenterne, idet flokkene kan inddeles i en såkaldt velfærdskategori, hvor flokscore <40 er **grøn**, flokscore mellem 41-80 er **gul** og flokscore> 81 er **rød**. Gentagne **gule** og alle **røde** vil give anledning til ekstra kontrol og rådgivning i besætningen.

Hyppighed

Hyppigheden af alvorlige problemer med trædepudesvidninger, som giver anledning til påbud, f.eks. at udarbejde en handlingsplan eller nedsætte belægningsgraden er relativ lav i Danmark. Opgørelser fra KIK databasen (Landbrug & Fødevarer) viste, at <10 % af flokke slagtet i 2011-2013 faldt i kategorien **rød**.

Omkostninger

Producenterne betaler i dag en afgift pr. hold til FVST, som dækker alle udgifter til kontrol, både bedømmelserne på slagterierne og den opfølgning/rådgivning som foretages i problembesætninger.

Automatisk overvågning

Meyn Footpadinspection System, som er baseret på vision, er indkøbt at to danske fjerkræslagterier til automatisk kontrol af alle fødder. Udstyret kan eventuelt overtage eller supplere den manuelle stikprøve, da udstyret kan monitorere skaderne på de fleste fødder, og vil sikre, at bedømmelsen er

stabil over tid og mellem flokke. Det vil øge slagteriernes præcision i den feedback, der gives til producenterne.

Formål

Projektets overordnede formål var at afprøve systemets evne til at klassificere slagtekyllingefødder, jf. den danske tre-trins-skala for trædepudeskader og på den baggrund vurdere fordele og ulemper i forhold til den nuværende manuelle stikprøvekontrol.

Delforsøg 1. Dokumentation og daglig drift

1.1 Meyn-udstyrets funktion og dokumentation af installation

Meyn-udstyret installeret i 2013

Meyn-udstyret blev installeret i 2013, og er i dag i funktion på to danske kyllingeslagterier. Algoritmerne i udstyret blev justeret i marts/april 2013, og algoritmerne i udstyret er baseret på bedømmelser af flere grupper á 100 fødder udført af veterinærerne på de 2 slagterier.

Udstyret er installeret på den urene del af slagtekæden efter kroppen er afskåret, så kun fødderne stadig hænger tilbage i bøjlerne, og endnu ikke er vippet af.

Udstyret er løbende blevet rengjort og vedligeholdt efter forskrifterne fra Meyn, og der er ikke blevet rapporteret betydende nedbrud. Slagterierne betegner udstyret som meget robust. Lyskilden skal skiftes hver 3. måned.

På et fælles møde (9. juni 2015) med projektgruppen og Meyn gruppen blev performance af udstyret og eventuelle justeringer drøftet. Især præsentationen af fødder i vision-udstyrets billedfelt skulle forbedres, og begge slagterier installerede guidebarer, som medførte en bedre vinkling af fødderne, som øgede andel af målte fødder markant.

Daglig kontrol

Ved start af slagtning kontrolles, at kamera optager billederne i den rigtige vinkel og position. Derudover er der endnu ikke udarbejdet en rutine for daglig kontrol og behandling af de opsamlede data. Der er desuden ikke leveret et fantom eller lignende referencemateriale til kontrol af udstyret efter nedbrud eller reparation.

Baggrundsplade

Meyn-personale har, ved test i løbet af projektet fremhævet, at det er vigtigt, at baggrundspladen holdes ren og fri for fjerrester og lignende, da det påvirker kvaliteten af billedet. Der er ikke fra udstyrsproducenten etableret procedure for renholdelse under produktionen, f.eks. automatisk spuling hver time eller lignende.

Udstyret tager et billede af begge trædepuder på en kylling og beregner trædepudesvidninger i tre klasser (**0, 1, 2**). Meyn-udstyret gemmer kun en score pr. kylling, og hvis fødderne ikke tildeles samme score, er det højeste score som gemmes og som dermed indgår i flokscoren.

Flokscoren beregnes på antal målte. Udstyret angiver desuden hvor mange kyllinger af flokken, det ikke var muligt beregne score på. Som standard gemmes følgende data fra hver flok:

Name	Start	End	NumberOf	Sco-red	NotSco-red	Flock Score	Score_0	Score_0_ Pct	Score_1	Score_1_ Pct	Score_2	Score_2_ Pct

Hvis der ikke er indtastet et "name" anvendes "I/O" som standard for manuel start af ny flok. Start/end er tidspunkter for manuel skift af flok.

Fletning af data

Meyn-udstyret er endnu ikke sat op til at modtage information om den flok-ID, der benyttes i de øvrige datasystemer hos Slagteri A og Slagteri B. For at flette de to typer af data – Vet.score fra KIK databasen og Meyn.score – var det nødvendigt, at flokkene blev identificeret vha. dato/tidskode og flokstørrelse.

I dag gemmes max 300 billeder i udstyret, og billederne overskrives, hvis der ikke manuelt tages kopi af billedfolderen, hvor de placeres default af udstyret. Der mangler at blive udviklet/installeret et program, som sikrer, at et udsnit af billeder fra hver flok gemmes, hvis der senere skulle opstå tvivl om flokkens korrekte status.

Formål

Ud fra eksisterende data blev stabilitet af de opsatte udstyr undersøgt. Undersøgelsen blev baseret på data fra Meyn-udstyrets brugergrænseflade og lagrede data. Dataopsamling blev foretaget af to omgange:

1) Før eventuelle ændringer for at kende basis niveau
2) Efter eventuelle ændringer for at kunne vurdere effekt af justeringer

Data på flokscore fra Meyn-udstyret sammenholdes med veterinær scoren i KIK databasen.

1.2 Resultater fra daglig drift

Resultaterne blev indsamlet i 2015. I tabel 1 er angivet antal flokke og hvor mange procent kyllinger af en flok, Meyn systemet kunne give en score før og efter justeringen. I algoritmen fra Meyn er indlagt, at hvis kun én fod er præsenteret i billedet indgår den, og hvis begge fødder er afbilledet og ikke er ens, er det den fod med højeste score, som indgår. Algoritmen er derudover baseret på areal af svidning og farve. Farvekorrektion sker også for at udgå, at skygger i billedet kommer til at tælle med som en svidning.

Begge slagterier savnede mulighed for direkte integration/interface fra Meyn til slagteriets øvrige datasystem, så floknummer kun skal indtastes ét sted. Det vil øge sikkerheden for at de data, der indgår i sammenligning mellem Meyn.score og Vet.score, også kommer fra samme flok. Aktuelt var det kun muligt at benytte tidsintervaller, og det kan medføre fejl, specielt ved delleverancer.

Materiale

Tabel 1. Antal flokke og scorede kyllinger pr. flok, data fra indledende driftskontrol.

	Antal flokke	Pct. kyllinger med Meyn score		
		Gnsn.	Min	Max
Slagteri A Før	85	65,1	44,5	93,0
Slagteri A Efter	69	87,0	47,9	99,1
Slagteri B Før	491	62,2	9,9	90,1
Slagteri B Efter	74	85,7	46,4	92,2

Den gennemsnitlige flokstørrelse var mellem 25.000 og 28.000 kyllinger med en variation fra <5.000 til ca. 80.000

Ved den første dataopsamling var der flere flokke fra Slagteri B, hvor antallet af kyllinger, som kunne måles var meget lav. Flokkene var tilfældigt fordelt over perioden, så der kunne ikke umiddelbart findes en forklaring. I de videre dataanalyser indgik de 65 flokke med mindre end 40 % scorede kyllinger ikke.

Justeringer

Følgende justeringer blev foretaget på det opsatte Meyn-udstyr mellem 1. (før) og 2.(efter) dataopsamling:

Slagteri A: Opretning af fødder for bedre præsentation, **ændret algoritme**, ny lyskilde, ny front

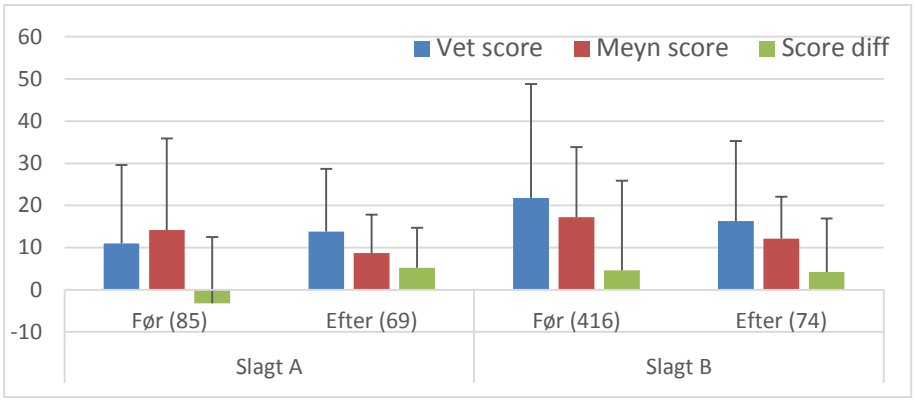
Slagteri B: Ny lyskilde, fokusering, opretning af fødder for bedre præsentation, ny front

Sammenligning af score

Sammenligning af de to metoder til trædepudesvidninger kan foretages på to måder, enten på basis af flokscoren (FC) eller på basis af overensstemmelse mellem de 3 velfærdsgrupper (**grøn**, **gul** og **rød**).

1.3 Sammenligning af flokscore fra henholdsvis stikprøve og Meyn-udstyr

Figur 1 viser gennemsnit og spredning for de to metoders FC, desuden er Vet.score – Meyn.score beregnet for de enkelte flokke, og gennemsnit er med i figur 1. Vet.score er den daglige stikprøve-bedømmelse udført af virksomhedens kødkontrolpersonale.



Antal flokke i opgørelsen står i parentes.

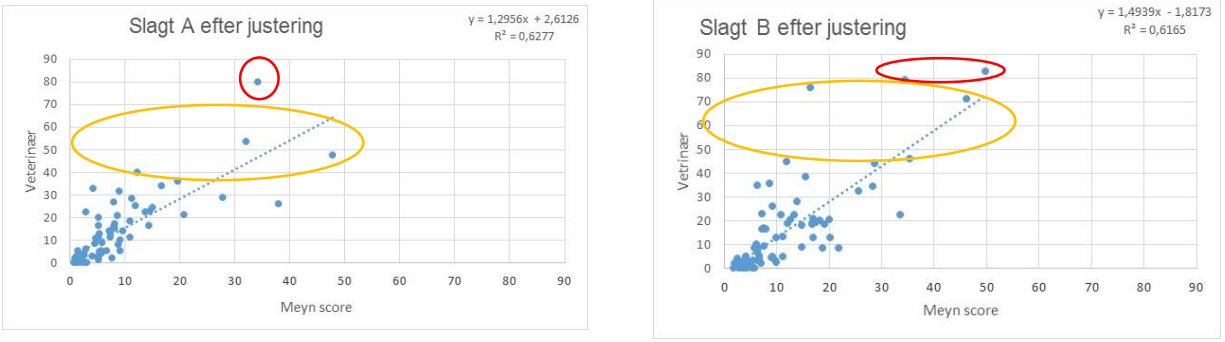
Figur 1. Vet.score og Meyn-score fra daglig produktion

Algoritmerne i Meyn-udstyret blev i 2013 udviklet hos Slagteri B og derefter implementeret og kontrolleret hos Slagteri A (intern rapport fra Meyn).

Beregninger på opgørelserne i figur 1 viste, at der ikke var signifikant forskel på Meyn-scoren og Vet.scoren i den 1. periode hos Slagteri A. Bemærk at recipe i Slagteri A's udstyr blev ændret mellem 1. og 2. rundes data-opsamling, da Meyns personale opdagede, at det var en forkert recipe, der var installeret i udstyret.

I de tre øvrige sammenligninger giver Meyn-udstyret en signifikant (Parret t-test) lavere score end Vet.scoren. Der er samtidig stor variation, idet spredningen på differencen mellem Vet.score og Meyn.score er 3 gange større end den gennemsnitlige difference.

I de tilgængelige data til sammenligning af Vet.score og Meyn.score er der ikke mange flokke med flokscore > 40. Det betyder, at sammenhæng (korrelation) ikke kan bestemmes særlig præcist, og at en statistisk sammenligning (enighed) af frekvens af velfærdsgrupper ikke er mulig. Meyn-udstyret fandt ikke de tre flokke fra de 2 slagterier med velfærdscore **2 – Rød**. Se figur 2.



Figur 2. Sammenhæng mellem Vet.score og Meyn-score
Data i **gul cirkel** er flokke med Vet.score > 40.
Data i **rød cirkel** er flokke med Vet.score > 80.

Meyn.scoren er signifikant lavere (parret T-test) end Vet.scoren på begge slagterier efter justering af Meyn-udstyret, og det afspejles også i figur 2.

Konklusion

Opretning af fødder ved hjælp af guidebarer umiddelbart før Meyn-udstyret bevirkede, at der kunne udregnes en trædepudescore for flere kyllinger i en flok. Frekvensen af scorede kyllinger blev øget fra 65 % til 85 % i gennemsnit af flokken.

Ændring af recipes i udstyret hos Slagteri A betød, at forskellen til flokscoren baseret på veterinær nærmede sig forskellen til Slagteri B.

Meyn-udstyret skal kunne kommunikere med de øvrige datasystemer på slagterierne, ellers kan data ikke benyttes til de enkelte flokkes score for trædepudesvidninger, da dataopsamlingen bliver for usikker og med risiko for fejl. Ligesom det ikke er muligt at give korrekt tilbagemelding til producenten

Anbefaling

Algoritmen/recipes i Meyn-udstyret skal justeres, så der er bedre overensstemmelse mellem flokscoren baseret på veterinær og flokscoren baseret på Meyn.

Sammenligning af flokscoren fundet med stikprøvekontrol og Meyn-udstyret er selvfølgelig vigtig, da en ændring af metode ikke bør ændre på den velfærdsgruppe, som flokken skal indplaceres i, forudsat metoden i øvrigt baserer sig på samme referenceprotokol. Da opgørelserne fra den daglige drift viste nogen uoverensstemmelse, blev delforsøg 2 sat i gang. I delforsøg 2 sammenlignes trædepudesvidning på de samme fødder scoret med henholdsvis udstyr og flere forskellige visuelle bedømmelser.

Delforsøg 2. Én til én sammenligning – Meyn-udstyr og veterinær score

Meyn-udstyret måler på de afskårne fødder, mens de stadig sidder i bøjlerne.

Det var ikke muligt at hænge fødder tilbage i bøjlerne, og få et retvisende billede af fødderne i Meyn-udstyret (oplysning fra Meyn). Derfor blev der ikke foretaget gentagne målinger på de samme kyllinger, f.eks. ved at gentage målingen på de fødder, der var bedømt af den daglige kontrol, eller ved at gentage målingerne på fødder som allerede var scoret.

Der er ikke tidligere gennemført én til én sammenligning mellem Meyn-udstyr og veterinærbedømmelser. Udstyrets algoritmer er udviklet på baggrund af hold á ca. 100 kyllinger, som så er scoret med Meyn-udstyr og af veterinær.

2.1 Fremgangsmåde

Forsøget blev gennemført hos både Slagteri A og Slagteri B. På begge slagtesteder indgik kyllinger fra tre flokke, og der blev udvalgt 1/3 forsøgskyllinger pr. flok. Der blev udtaget 160 kyllinger pr. slagteri, dvs. i alt 640 fødder.

Forsøgskyllingerne blev udtaget over én dag/slagteri i perioden kl. 7-11, og billedoptagelsen med Meyn-udstyret blev gennemført kl. 12-13.

Det var de lokale veterinærer/kødkontrolpersonale på de respektive slagterier som forestod bedømmelserne, hvilket indebærer mulighed for en evt. ukendt interkalibreringsbias mellem bedømmere. D1 & D2 for hver af de to slagterier er således forskellige bedømmere tilknyttet det enkelte slagteri.

Hele kyllinger til forsøget blev udtaget på slagtelinjen lige før afklipning af fødderne. Kyllingerne blev taget af bøjlen af veterinær (D1) i grupper á 20. Når der var udtaget ca. 20 kyllinger, blev begge fødder scoret af D1 i tre klasser: 0, 1 & 2, efter den samme skala som anvendes i den daglige kontrol. Det blev tilstræbt at der var ca. 1/3 kyllinger i hver score klasse.

Der blev sat ID-nummer på begge fødder, og kyllingerne blev derefter lagt i kasser, som stod rummet indtil dagens slagtninger sluttede (ca. 4 timer fra de første blev udtaget). Efter dagens slagtninger blev kyllingerne hængt tilbage på bøjlerne før afskæring af fødder, kroppene blev skåret fri og fødderne blev derefter scoret med Meyn-udstyret, og billederne blev gemt til eventuelle senere analyser.

Individuel Meyn score opsamling

Meyn-udstyrets score blev efterfølgende registreret af personer fra Meyn, så score pr. fod blev opsamlet. Den enkelte fods score kan aflæses på billedet, når udstyrets billedalgoritme benyttes, men det var ikke muligt at gemme registreringer fra begge fødder, da det i den eksisterende opsætning er kyllingens samlede score, som gemmes.

Efter Meyn-udstyret havde optaget billeder, blev fødderne taget af bøjlerne og samlet i kasser, så de igen kunne bedømmes efter samme princip (placering, lysforhold og skala) som de daglige stikprøver. Anden scoring blev foretaget af en ny veterinær (D2), og igen blev begge fødder fra hver kylling bedømt.

Efter scoring blev højre og ventre fod fra hver kylling samlet i en pose, frosset og efterfølgende sendt til Aarhus Universitet, Foulum. Her blev alle fødder i løbet af én dag igen bedømt af reference bedømmer (D3). D3 er ansvarlig for den overordnede kalibrering af veterinærerne fra slagterierne.

2.2 Resultater fra undersøgelsen

Data blev analyseret med proceduren: PROC FREQ, fra SAS Institute. I dette tilfælde benyttes Weighted Kappa Coefficient som effektmål, da der er tale om ordnede kategorier (0, 1, 2). (SAS/STAT® 9.22 userguide). Coefficienten går fra ***0*** (fuldstændig uenig) til ***1*** (fuldstændig enig).

I tabel 2 er antal af kyllingefødder, som indgår i beregningerne samlet fra de forskellige kilder.

Materiale

Tabel 2. Antal fødder pr. bedømmelse

		Meyn	D1	D2	D3
Slagteri A	Venstre	128	160	160	160
	Højre	122	160	160	160
Slagteri B	Venstre	149	160	160	154
	Højre	136	160	160	154

Der manglede bedømmelser på nogle fødder fra Meyn systemet. Det kan skyldes at fødderne ikke var præsenteret korrekt eller overlappede i billedet, så det ikke er muligt for algoritmen at finde den enkelte fods trædepude. Der er også frasorteret 6 par fødder fra D3's bedømmelse. Det skyldes fejlaflæsning af ID-nummer.

2.3 Sammenligning af veterinærer - D1 og D2 pr. slagteri

Da kyllingefødderne blev bedømt af to forskellige veterinærer kan en sammenligning af de to bedømmelser på hvert slagteri vise graden af

overensstemmelse mellem bedømmere og give et estimat for graden af enighed. Resultaterne fremgår af tabel 3 og 4.

Forsøgskyllingerne blev udvalgt og bedømt af D1 og D2 i det samme område, som de fødder der indgår i den manuelle flokscore, normalt udvælges. De to dommere havde ikke mulighed for at tale sammen om deres score, da bedømmelsen er foretaget forskudt; men skulle gennemføre bedømmelsen efter samme retningslinjer, som de plejede. Dog uden mulighed for at skære i foden og dermed udføre en 3-D bedømmelse.

Tabel 3. Slagteri A: Enighed mellem dommere - D1 og D2

Antal ***Række %*** ***Kolonne %***	**D2 (0)**	**D2 (1)**	**D2 (2)**	**Total**
D1 (0)	66 *55* *98,5*	53	1	120
D1 (1)	1	76 *69,7* *56,3*	32	109
D1 (2)	0	6	85 *93,4* *72,0*	91
Total	67	135	118	320

Kappa Coeffient (Graden af enighed) for Slagteri A's veterinærer D1 og D2 blev beregnet til *0,663*

Tabel 4. Slagteri B: Enighed mellem dommere D1 og D2

Antal ***Række %*** ***Kolonne %***	**D2 (0)**	**D2 (1)**	**D2 (2)**	**Total**
D1 (0)	108 *94* *88,5*	7	0	115
D1 (1)	10	118 *92,2* *67,1*	0	128
D1 (2)	4	51	22 *28,6* *100*	77
Total	122	176	22	320

Kappa Coefficient (Graden af enighed) for Slagteri B's veterinærer D1 og D2 er beregnet til *0,680*.

Score i klasser vil altid give diskussion af grænsetilfælde, og som det er kendt fra interkalibrering mellem bedømmere, kan der ikke forventes fuldstændig overensstemmelse mellem bedømmelserne. Det var dog kun få fødder (5 af 640), hvor den ene dommer gav 2 og den anden 0.

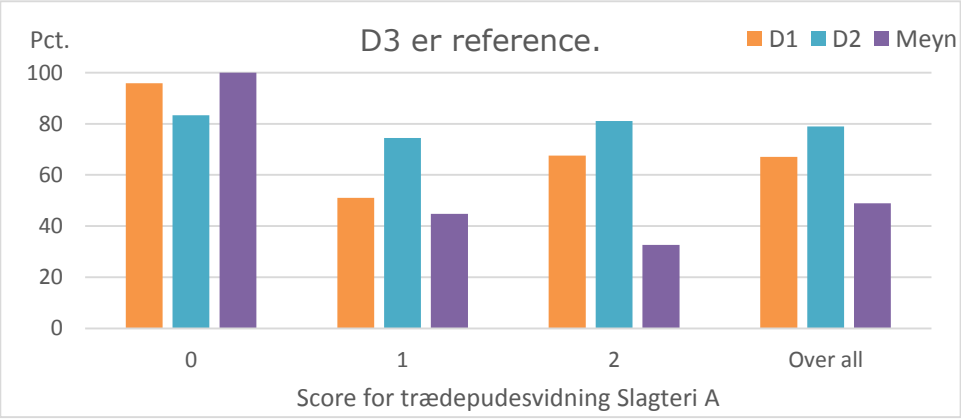
2.4 Referencens bedømmelse og overensstemmelse til Meyn, D1 og D2

Reference

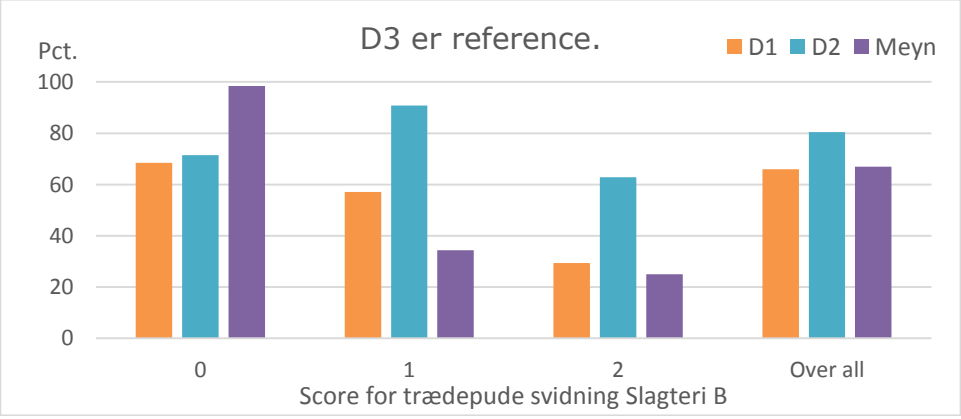
Referencen fra Aarhus Universitet (D3) udfører interkalibrering og sikrer dermed, at der tilstræbes ensartet bedømmelse af trædepudesvidninger på danske fjerkræslagterier. D3 er medtaget som reference i dette projekt, da bedømmelsen kan betragtes som en slags endeligt facit for dansk bedømmelse af trædepudesvidning.

I den efterfølgende databehandling betragtes D3's score som reference og de øvrige dommere og Meyn-udstyret sammenlignes derfor til denne. Data for de to slagterier blev behandlet hver for sig.

I figurerne 3 og 4 er enighed mellem reference (D3) og D1, D2 og Meyn-udstyret afbilledet. Figurerne viser, hvor mange procent af reference-bedømmelsen, der var enighed om for hvert tilfælde, dvs. '0-0', '1-1', '2-2' og så den samlede enighed.



Figur 3. Slagteri A – opgørelse af enighed mellem reference og øvrige bedømmelser



Figur 4. Slagteri B – opgørelse af enighed mellem reference og øvrige bedømmelser

Der var stor enighed mellem reference og Meyn-udstyret i klasse 0, men meget ringere overensstemmelse mellem klasse 1 og klasse 2. Det påvirkede den samlede enighed, som dermed kun var 49 % for Slagteri A og 67 % for Slagteri B.

Overensstemmelse reference og veterinær

Med undtagelse af Slagteri A's klasse 0 var D2 mere enig med referencen på begge slagterier. En forklaring kan være, at D2's bedømmelse blev foretaget under samme betingelser som de daglige bedømmelser for trædepudesvidning, mens D1 skulle bedømme på hele kyllinger. Graden af enighed udtrykt med Kappa coefficienten, tabel 5, understøtter disse kommentarer.

Tabel 5. Graden af enighed (weighted Kappa) på trædepudesvidninger

Kappa koeeficient for:	Slagteri A	Slagteri B
D1/D3	0,617	0,554
D2/D3	0,734	0,663
Meyn test/D3	0,165	0,415

Da Footpad-bedømmelsen blev indført i Sverige i 1997 fandt Ekstrand et al, (citeret fra Charlotte Berg (1998)) gennemsnitlige kappa værdier på 0,86±0,15 i en undersøgelse, der omfattede 11 inspektører.

Den højeste grad af enighed i denne undersøgelse blev beregnet til *0,734* med et 95% konfidens interval på *[0,675 – 0,794]* (reference og D2-Slagteri A).

Graden af enighed mellem referencen og D1/D2 er dermed lavere i denne undersøgelse end tidligere rapporteret. Det kan eventuelt forklares af, at der også er en usikkerhed på, hvordan bedømmelsen skal foretages (se afsnit 3).

Graden af enighed kan bruges som et mål for kravet til udstyrets performance, og på begge slagterier var graden af enighed mellem referencen (D3) og Meyn-udstyret væsentligt lavere end mellem referencen og de 2 veterinærers score.

2.5 Ændrede algoritmer/recipes i Meyn-udstyret

Resultaterne af den første test viste lav grad af enighed mellem reference og Meyn-udstyr (tabel 5). Meyn (Cor Peiterse) udarbejdede derfor en ny algoritme baseret på referencedata (rapport i bilag 1). Billeder fra Slagteri A blev anvendt til at udvikle den nye recipe, og D3s bedømmelse blev brugt som reference.

Kommentarer til Bilag 1: Rapport fra Meyn

Principielt er det ikke optimalt at udvikle algoritmer på data fra et slagteri, og teste dem på et andet, da det kan medføre fejl, som skyldes andre faktorer end selve algoritmeberegningerne. Hermed kan det slagteri, som leverer data til algoritme-udviklingen, vise en bedre overensstemmelse mellem udstyrets score og bedømmelsen end et andet slagteri.

Den nye algoritme reducerer antal fødder i score 0 og øger antal fødder med score 2. I rapportens tabel 1 og 2 beregnes også en 'final' score for Meyn systemet som svarer til kyllingens værste score. Det kan give det indtryk, at niveauet for Meyn scoren er på niveau med reference.scoren. Det er dog ikke korrekt, da der ikke sammenlignes til en 'final' for reference.scoren.

Den score som er beregnet i bilag 1, tabel 1 og 2, kan ikke bruges i denne sammenligning, da materialet netop er udvalgt til en ligelig fordeling af de tre trædepudesvidningsklasser og dermed ikke kan sige noget om flokscore.

Konklusionen i Meyn-rapporten er derfor alene et udtryk for udstyrs-leverandørens beregninger.

DMRI modtog også de nye data beregnet på alle fødder (billeder) fra forsøget. Hvor billederne var analyseret med den nye recipe og igen tildelt værdierne **0**, **1**, **2**.

I tabel 6 er vist de beregnede Kappa koefficienter for referencen og de to bedømmere inkllusive 95% konfidens interval pr. slagteri, som et udtryk for graden af enighed.

Tabel 6. Kappa coefficient for den nye recipe

Kappa koeeficient for:	Slagteri A	Slagteri B
D3/Meyn ny	0,588 [0,516-0,660]	0,661 [0,583-0,739]
D1/Meyn ny	0,661 [0,594-0,729]	0,415 [0,340-0,490]
D2/Meyn ny	0,575 [0,507-0,644]	0,449 [0,364-0,534]

Den nye algoritme gav en større grad af enighed mellem D3 (referencen) og Meyn-udstyret gældende for begge slagtesteder. Enighed mellem Meyn-udstyret og Slagteri B's bedømmelser var til gengæld lav.

Der er velkendt at der altid findes bedre overensstemmelse mellem rådata og estimeret værdi på de data som algoritmen udvikles på. Det kan være en del af forklaringen på, at data/billederne fra Slagteri A viste bedre overensstemmelse end Slagteri B.

Konklusion

Én til én sammenligning af trædepudesvidning målt med Meyn og flere forskellige referencer viste ikke fuld overensstemmelse mellem karakteren på den enkelte fod, hverken mellem udstyr og de enkelte referencer, eller mellem referencer indbyrdes. Graden af enighed var i gennemsnit ca. 0,6 til

0,7 mellem referencer. Mellem Meyn-udstyrets optimerede algoritme og referencer var overenstemmelse lavere, fra ca. 0,4 til 0,7.

Metode og Guidelines for trædepudebedømmelser

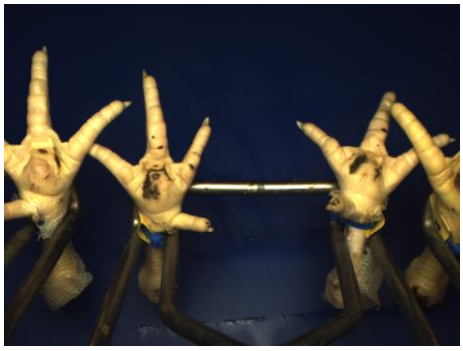
Trædepudesvidninger er udviklet for at have et redskab, som kan bruges som indikator for dyrevelfærd i fjerkræbesætningen.

De danske bedømmere anvender en skala, som er udviklet i Sverige. Retningslinjerne for bedømmelse af tåsvidningerne fremgår af denne kommentar fra Lotta Berg, 2015:
Till fotskadorna: i Sverige bedömer vi i praktiken bara den centrala trampdynan. I riktlinjerna står det att "Generellt skall foten som helhet bedömas. Främst är det dock skador på den centrala trampdynan som är viktigast, inte tårna." Jag har själv nästan aldrig sett någon fot som enbart har skador på tårna och inte alls på den stora trampdynan… Men om det är kycklingar som har fina centrala trampdynor men tydliga skador på tårna så ska de inte ha 0, utan få en 1:a. En liten fläck på en enstaka tå hade jag dock låtit passera – det "ska vara något" för att noteras alls.

3-D versus 2-D Tidligere interkalibrering af veterinærerne har vist, at der er forskel i gråzonerne: Når det skal afgøres, om foden er den værste 1'r eller den bedste 2'r, eller dårligste 2'er kontra bedste 3'er. I enkelte tilfælde vil en gennemskæring af trædepuden vise, hvor dyb svidningen er og det kan afgøre kategori, en såkaldt 3-D bedømmelse. Meyns visionudstyr kan kun måle på synlig overflade 2-D, til gengæld bedømmes en langt større andel af kyllingerne.

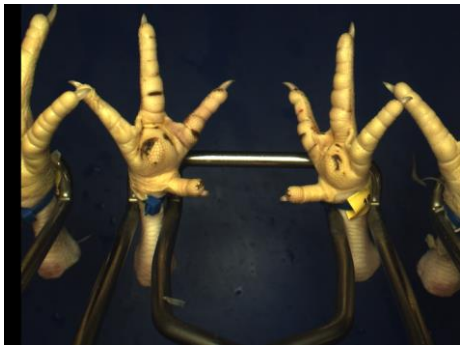
I denne undersøgelse var det kun den sidste bedømmer: Referencen, som havde mulighed, og D3 skar i ca. 35% af fødderne fra Slagteri A og 15% af fødderne fra Slagteri B.

En nærmere analyse af billederne fra Meyn-udstyret tydede på, at bedømmerne har skærpet opmærksomhed på rifter/svidninger mellem selve trædepuden og tæerne. Se billederne i figur 5A til 5D.



5A 5B

Figur 5A og B. Slagteri A - D1, D2 og D3 har bedømt begge til 2. Meyn har bedømt 0.



5C 5D

Figur 5C. Slagteri B – D1, D2 og D3 har bedømt begge til 2. Meyn har bedømt 0.

Figur 5D. Slagteri B – D1, D2 og D3 har bedømt den højre til 2. Meyn har bedømt 0.

På baggrund af dette foretog DMRI en visuel gennemgang af alle billederne opsamlet med Meyn-udstyret, hvor der var uoverensstemmelse mellem reference (D3), bedømmelse og Meyn-udstyret. Resultatet blev samlet i tabel 7.

Pr. billede blev noteret, om der kun var tåsvidninger, kun svidninger yderst/kant på trædepuden, kun var trædepudesvidninger eller svidning på både tå og trædepude.

Tabel 7. Samlet oversigt hvor der er uoverensstemmelse mellem Reference og Meyns oprindelige recipe

Slagteri A

D3-Meyn	antal	Kommentarer
0-1	0	
0-2	1	1 D3 har overset trædepudesvidning
1-0	60	6 stk. kun på tæer 9 stk. svidning yderst/kant på trædepude 32 stk. trædepudesvidning 13 stk. både trædepude og tåsvidning
1-2	2	2 stk. trædepudesvidning stort område D3 har fastlagt score efter skæring i trædepude
2-0	13	8 stk. trædepudesvidning 5 stk. både trædepude- og tåsvidning
2-1	58	30 stk. trædepudesvidning, heraf 3 store arealer 28 stk. både trædepude- og tåsvidning, heraf 4 store arealer

Slagteri B

D3-Meyn	antal	Kommentarer
0-1	2	1 trædepudesvidning som er overset af D3 1 lille plet som er blevet bedømt af Meyn
0-2	0	
1-0	56	1 fejl bedømt af reference – ingen svidning 1 stk. svidning kun på tæer 7 stk. svidning yderst/kant på trædepude 27 stk. trædepudesvidning 20 stk. både trædepude- og tåsvidning
1-2	2	2 stk. trædepudesvidning
2-0	7	2 svidning yderst på trædepude 1 stk. trædepudesvidning 4 stk. både trædepude- og tåsvidning
2-1	11	2 stk. trædepudesvidning 9 stk. både trædepude- og tåsvidning

Meyn-udstyret er med sin nuværende software udviklet til kun at måle trædepudesvidninger på den centrale del af foden og inddrager ikke svidninger på tæer eller svidninger yderst/på kant af den centrale trædepude.

På Sslagteri A skyldes 10 procent af de tilfælde, hvor der var uover-ensstemmelse mellem referencenter og Meyn, tå- og/eller kantsvidninger. Mens det var 13,5 % på Slagteri B.

Diskussion

Denne ekstra kontrol af billeder viste, at det ikke kun var på grund af svidninger uden for den centrale del af trædepuden, at referencen og Meyn-udstyret ikke bedømmer ens. Optællingen i tabel 7 viste også, at Meyn-udstyret generelt giver lavere score på fødder, hvor der kun ses svidning på trædepuden.

Udvikling af automatisk måleudstyr kræver gode referencedata, og især ved udvikling af algoritmer med efterfølgende kontrol af målingerne er det vigtigt, at der er konsensus om de manuelle bedømmelser. Ligesom ændrede normer for bedømmelserne kan have betydning for en efterfølgende kontrol af udstyret.

De algoritmer, som i dag er tilgængelige i Meyn-udstyret, kan forbedres. I overvejelserne om hvordan Meyn-udstyret skal foretage beregning af trædepudesvidning skal reference protokol også defineres mere præcist. Delprojekt 1 viste, at der ikke er god overenstemmelse mellem Meyn-udstyrets beregning af flokscore og den daglige kontrols flokscore, og delprojekt 2 viste, at bedømmelse af præcis samme fod også gav forskellige karakterer, hvor Meyn-udstyrets algoritme bedømte lavere end reference bedømmelserne.

Delprojekt 2 viste desuden, at graden af enighed på de manuelle bedømmelser var lavere end tidligere rapporteret i de indledende undersøgelser i Sverige. I de første videnskabelige undersøgelser blev benyttet en 6-trins skala. Det fremgår desværre ikke, om en 6-trins eller 3-trins skala giver den bedste overensstemmelse mellem dommere. Når der kun benyttes tre klasser, kan grænsetilfældene let blive placeret uens. Til gengæld er der ikke så mange.

Det kan også skyldes, at det foreløbig ikke er muligt at udarbejde en fuldstændig oversigt over trædepudesvidninger med tilhørende kategori. Ved en protokol som ikke kan/er udformet præcist, skal bedømmerne nødvendigvis anvende en grad af skøn.

I den danske velfærdskontrol på fjerkræslagterierne er der fra indførelsen benyttet den samme 3-trins skala.

Fordele og ulemper af automatisk overvågning af trædepudeskader i forhold til den nuværende stikprøvekontrol

Fordele

Fordele:

- Udstyret er driftsikkert og kan med den nuværende opsætning og algoritme måle den centrale trædepude på 85-90% af alle kyllinger i en flok.
- Niveauet af alvorlige trædepudesvidninger er relativt lavt i Danmark, og det kan derfor være en fordel for korrekt tilbagemelding til producenten, at målingen foretages på en stor andel af de leverede kyllinger.
- Med en optimal algoritme indlagt i ens installerede vision udstyr, som holdes under tæt kontrol, bliver hovedparten af alle trædepuder bedømt efter de samme retningslinjer med mindre risiko for stikprøvevariation.

Ulemper

Ulemper:

- Algoritmerne i udstyret er udviklet på basis af manuelle reference bedømmelser, og der er konstateret nogen usikkerhed om hvordan referencemålingerne skal foretages. Det afspejler sig også i de niveauforskelle som dette projekt har påvist.
- Software og algoritmer er, som de foreligger, ikke optimeret til at tage højde for svidninger på tæer. Det skal vurderes i hvilket omfang dette er nødvendigt.
- Udstyret er endnu ikke fuldt integreret i slagteriernes datasystem, og data fra udstyret benyttes aktuelt ikke i opfølgningsarbejdet.

MEMO

To Madchen Hviid
From :Cor Pieterse

CC :Louis van Steijn, Daniel Derksen

Subject :Performance Footpad camera system for Denmark

BILAG 1 New recipe for footpad measurement in Denmark

Introduction:

End October 2015 experiments were done to test the agreement between manual footpad measurement and automatic measurement with Meyn camera system. The results shows a difference for class 1 and class 2 which were too big. The impression is that, by changing the algorithm, the performance of the software should be better in accordance with the reference measurements. With the information from Denmark a new recipe is made especially for the Danish market.
Danish Meat research Institute (DRMI) is responsible for the research. During the test two local Veterinarians were involved by collecting 50 feet for class 0, 1 and 2. The whole chickens were taken off the slaughter line. After production the chickens were rehanged and after the rehanger between slaughter line and evisceration line the feet passed the camera for lesion detection. Finally the feet were collected again and packed with number for a final judgment by an extern reference. All images of the feet were saved for later use. The test is done in two different plants. The judgment of the reference is leading for the total experiment.

Material and Methods:

The collected images from the test from the first day (plant Slagteri A, Vinderup) are used to develop a new recipe. These data are quite good divided over the three classes and seems a good set to make a new recipe.
The score from the Danish expert is the reference to develop the new recipe. After determining the best recipe, the images from the second plant have been processed by the program using this new recipe. All scores are placed in an Excel file for further calculations by DRMI.

Result/ Discussion:

With the new recipe there is found a good agreement between the manual score and automatic camera score.

The recipe has been developed using the data from plant Slagteri A Aars. After finding an agreement the recipe is set for the second plant without changing the borders.
Table 1 and 2 show the results.
The agreement in percentage for the three classes is good, score 0 in table 1 is about 3 % higher and in table 2 nearby the same as for the reference. The agreement for score 1 and score 2 are also close to the reference score which result in an small difference in the end score. Finally the end score is important.

plant Slagteri A Vinderup

			score 0		score 1		score 2		end score
		total	amount	%	amount	%	amount	%	
Refe-rence	**right**	**160**	**35**	**21,88**	**71**	**44,38**	**54**	**33,75**	**89,7**
	left		**37**	**23,13**	**66**	**41,25**	**57**	**35,63**	**91,9**
camera	right	136	63	46,32	55	40,44	18	13,24	46,7
test	left	149	68	45,64	62	41,61	19	12,75	46,3
	final	158	57	35,63	76	47,5	24	15	53,8
camera	right	138	43	31,16	53	38,41	39	28,26	75,7
new	left	149	53	35,57	57	38,26	66	44,30	107,7
	final	**157**	**39**	**25,3**	**66**	**41,1**	**52**	**33,5**	**87,6**

Table 1: The result of camera and reference measurement. The first rows are the data from the reference, then the results from recipe used during the test and finally the results from the new developed recipe.

plant Slagteri B Aars

			score 0		score 1		score 2		end score
		total	amount	%	amount	%	amount	%	
Refe-rence	**right**	**157**	**80**	**50,96**	**66**	**42,04**	**11**	**7,01**	**35,0**
	left	**157**	**85**	**54,14**	**58**	**36,94**	**14**	**8,92**	**36,3**
camera	right	122	94	77,05	24	19,67	4	3,28	16,4
test	left	128	102	79,69	22	17,19	4	3,13	14,8
	final	143	104	72,73	33	23,08	6	4,20	19,9
camera	right	123	66	53,66	51	41,46	6	4,88	30,5
new	left	127	83	65,35	34	26,77	10	7,87	29,1
	final	**146**	**76**	**52,05**	**57**	**39,04**	**13**	**8,90**	**37,3**

Table 2: The result of camera and reference measurement. The first rows are the data from the reference, then the results from recipe used during the test and finally the results from the new developed recipe.

Conclusion:

Within this experiment a good agreement is found between manual and automatic measuring the footpad lesions with the new recipe. The measurement of the reference is guiding the result of the camera system. The judgment of the feet from the reference is very important and forms the basis for the whole experiment. A small control is made based on the score of the reference and camera system special when there are found differences (In the chapter “Remarks” some examples are given and possible reason is given for these difference).

The reference has judged all feet and by doubts an incision is made in the lesion to control the depth of the lesion. This way of handling is not possible by automatic scoring of the lesion by camera and in practice time consuming by manual measuring.
The biggest difference will be found for lesion around the borders of 2 classes.
Especially class 1 is critical as class 1 is surrounded by two classes. The score of a border lesion can be different between manual and camera measurement. The judgment by the software is partly based on the surface of the lesion. Is the border for example 575 pixels and the lesion is 576 pixels the score is class 2 however if the surface of the lesion is 475 pixels the score is class 1. A human inspector cannot see these small differences.
Overall there is found a good agreement between manual score of the reference and the automatic footpad score. Finally the end score is the most important value.

Remarks:

The data from Denmark is placed in an Excel file besides the original dataset to get a better overview. In the set data is missing from feet number 103, 104 and 111 for plant Slagteri B Aars. After developing the new recipe we looked in the data why there were some remarkable differences. There are several examples were the toe’s were included in the manual measurement.
During several meetings is has been mentioned and agreed that Denmark followed the Swedish score method and score cart. During several test in the past this method was used.
In the Swedish method only the footpad is measured. So the software is developed on this method and find only the footpad were the lesion can be found.
There are limitations of the software. By measuring almost the whole flock a wrong classification (positive and negative) will be compensated in most cases and shall be divided over all groups.
A wrong classification of the reference will overrule a good classification of the camera system

Following the Swedish method the software tries to find the center of the footpad and from there build up the measuring area. We do not look at the lesions on the toe because this part is excluded from measurement (I can understand that the severe lesion on the toe can also be painful).

								camera during test				new recipe		
		Fodpar	**Hscore**	**H-cut**	**Vscore**	**V-cut**		right	left	**final**		right	left	**final**
1	2015-10-20-13h15m47s22_0.tif	1	1	n	1	n		1	0	**1**		0	0	**0**
22	2015-10-27-12h58m23s38_0.tif	22	2	j	2	j		0	0	**0**		1	0	**1**
48	2015-10-20-13h24m45s15_N.tif	48	1	j	1	j			0	**0**			0	**0**

Table 1: The different score, image name, result of the reference, result of camera during test and finally with the new recipe.

The results of examples from figure 1 to 3 of reference and camera score are given in table 1. Figures 1 to 3 gives different examples were probably the toe's were included in measurement.

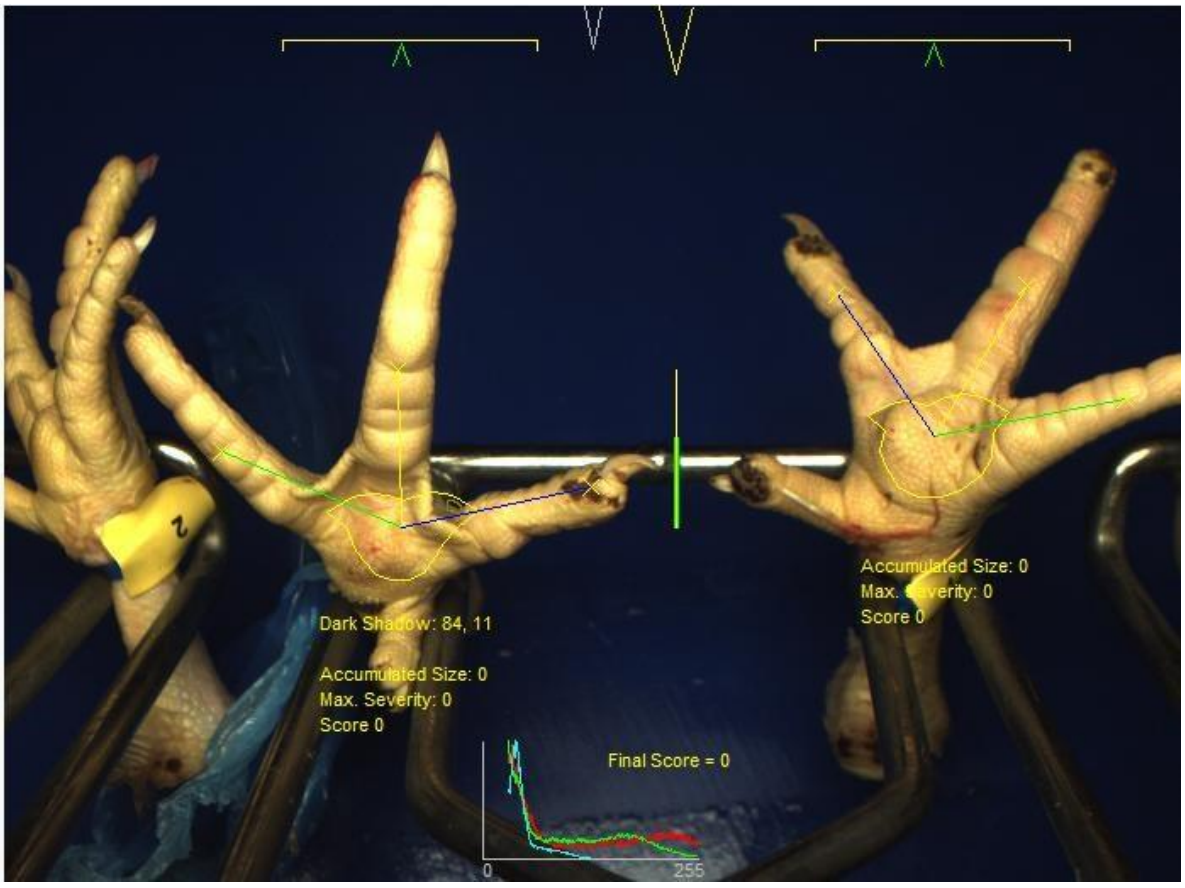
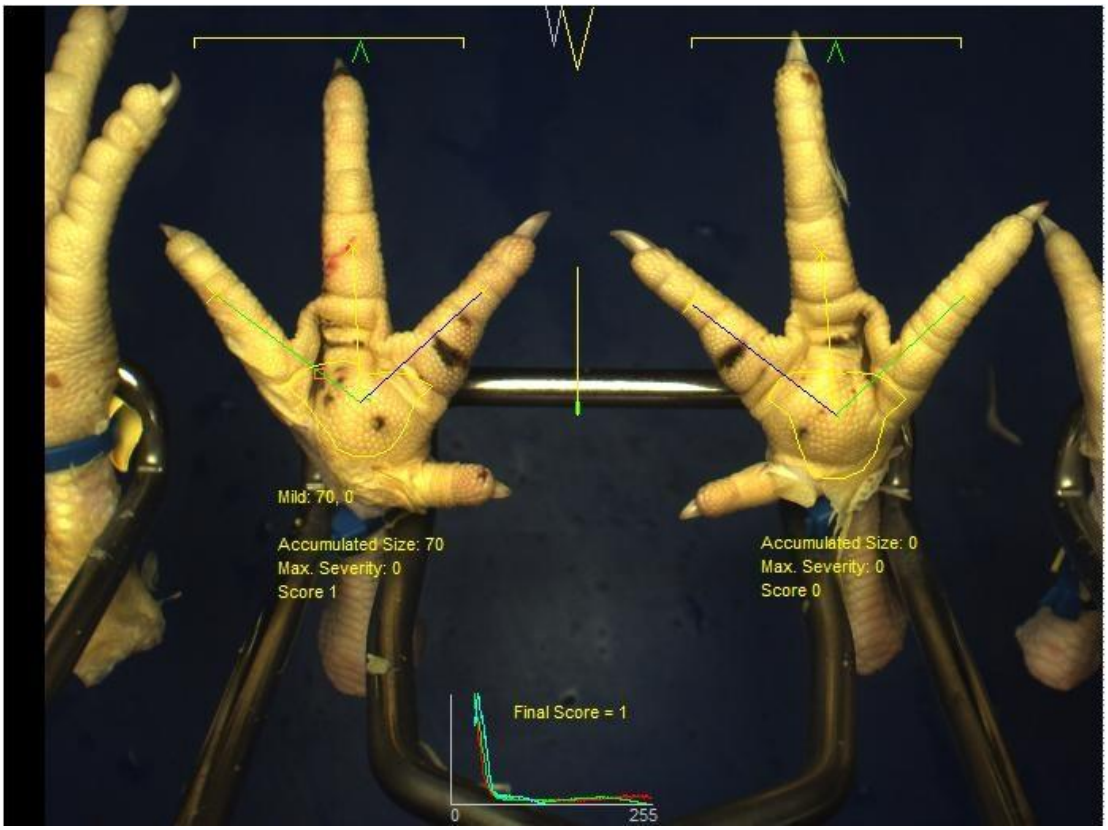


Fig. 1. Image 2015-10-20-13h15m47s22_0.tif and follow-number 1 from plant

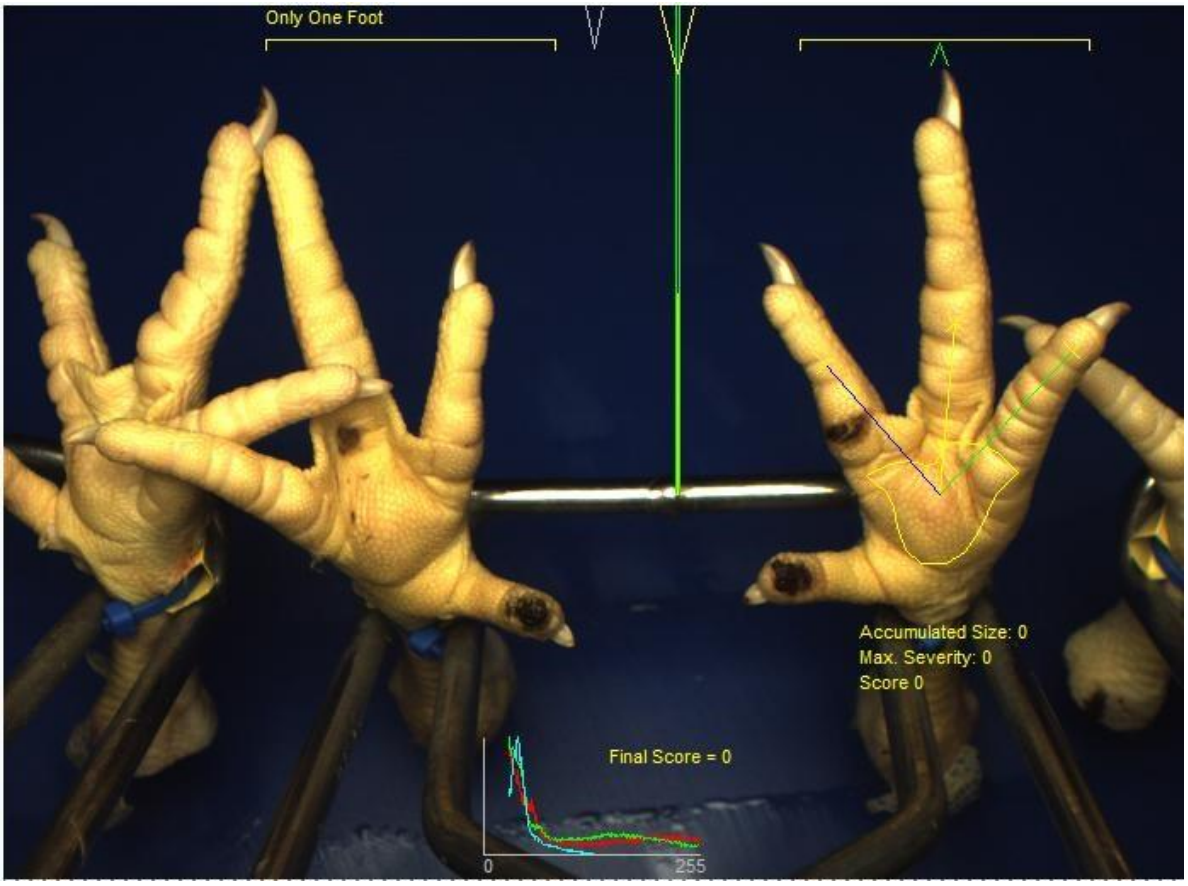
Slagteri A Vinderup. An example of a different score between the program and the reference. The foot centre is clean for right foot and a small lesion on the left foot at the edge of the footpad and toe. On the toe are lesions visible. Based on the footpad the score is 0. The reference scores 1 for both feet probably based on the lesion on the toe. (With the old recipe the right foot was scored as 1 caused by shadow between two toe's.) This is corrected with the new recipe.

December 15, 2015



Example 2 with image 2015-10-27-12h58m23s38_0.tif and follow-number 22 from plant Slagteri B Aars. Maybe also a wrong judgment. The reference scores for both feet score 2 and the software with new recipe scores respectively 1 and 0. Probably the lesions on the toe's are responsible for the difference.

December 15, 2015



Example 3 is image 2015-10-20-13h24m45s15_N.tif with follow-number 48 from plant Slagteri A Vinderup. The reference scores for both feet 1 while the software scores 0. The footpad is clean, but just outside of the footpad on two toe's (right and left foot) were lesions. All lesions are outside of measuring area for the software. (The software did not find the right foot probably caused through a toe from another chicken) and score on left foot score 0.