ORIGINAL ARTICLE

Interrater and intrarater agreement of confocal microscopy imaging in diagnosing and subtyping basal cell carcinoma

D.J. Kadouch,[1,*,†] A.S.E. van Haersma de With,[1,†] Y.S. Elshot,[1,2] M. Peppelman,[3] M.W. Bekkenk,[1,4] A. Wolkerstorfer,[1] I. Eekhout,[5,6] C.A.C. Prinsen,[5] M.A. de Rie[1,4]

[1]Department of Dermatology, Academic Medical Center, Amsterdam, The Netherlands
[2]Department of Dermatology, Netherlands Cancer Institute, Amsterdam, The Netherlands
[3]Department of Dermatology, Radboud University Medical Center, Nijmegen, The Netherlands
[4]Department of Dermatology, VU Medical Center, Amsterdam, The Netherlands
[5]Department of Epidemiology and Biostatistics, Amsterdam Public Health (APH) Research Institute, VU University Medical Center, Amsterdam, The Netherlands
[6]Netherlands Institute for Applied Scientific Research (TNO), Leiden, The Netherlands
*Correspondence: D.J. Kadouch. E-mail: d.j.kadouch@amc.uva.nl

Abstract

Background Reflectance confocal microscopy (RCM) imaging can be used to diagnose and subtype basal cell carcinoma (BCC) but relies on individual morphologic pattern recognition that might vary among users.

Objectives We assessed the inter-rater and intrarater agreement of RCM in correctly diagnosing and subtyping BCC.

Methods In this prospective study, we evaluated the inter-rater and intrarater agreement of RCM on BCC presence and subtype among three raters with varying experience who independently assessed static images of 48 RCM cases twice with four-week interval (T1 and T2). Histopathologic confirmation of presence and subtype of BCC from surgical excision specimen was defined as the reference standard.

Results The inter-rater agreement of RCM for BCC presence showed an agreement of 82% at T1 and 84% at T2. The agreements for subtyping BCC were lower (52% for T1 and 47% for T2). The intrarater agreement of RCM for BCC presence showed an observed agreement that varied from 79% to 92%. The observed agreements for subtyping varied from 56% to 71%.

Conclusions In conclusion, our results show that RCM is reliable in correctly diagnosing BCC based on the assessment of static RCM images. RCM could potentially play an important role in BCC management if accurate subtyping will be achieved. Therefore, future clinical studies on reliability and specific RCM features for BCC subtypes are required.

Received: 30 July 2017; Accepted: 7 December 2017

Conflict of Interest Disclosures

None of the authors have any financial arrangements or potential conflict of interest related to this article.

Funding sources

No funding has been received for this article.

Introduction

The rising incidence of basal cell carcinoma (BCC) is causing a major burden on worldwide healthcare systems.[1] With the increasing use of effective non-surgical therapies for superficial BCC, histological subtype (i.e. aggressiveness) becomes more important in determining the most suitable BCC treatment.[2]

Current international guidelines recommend on performing a punch biopsy to confirm clinical diagnosis and divide between BCC subtypes.[3,4] However, non-invasive skin imaging techniques might be able to change the diagnostic pathway for patients suffering from BCC.[5,6] Of those techniques, in vivo reflectance confocal microscopy (RCM) seems very promising as the procedure enables inspection of the whole lesion while the morphologic features are similar to routine histology.[7] If RCM would be able to accurately diagnose and subtype BCC, not only

†Both authors contributed equally to this study.

the amount of painful invasive skin biopsies could be reduced but also the time delay between diagnosis and treatment, administrative workload and healthcare costs.[8,9] Yet prior to replacement of routine punch biopsies, a critical appraisal of the diagnostic RCM procedure is needed. An important risk of techniques such as RCM is that it relies on morphology-based assessment. Therefore, it is subject to interpretation bias.

The purpose of this study was to determine the inter-rater and intrarater agreement of RCM in correctly diagnosing and subtyping BCC based on static RCM images.

Methods

Study design and patients

This reliability study evaluated inter-rater and intrarater agreement using static images of 48 RCM cases among three raters (DK, YE and MP). The series of images were prospectively derived from clinically suspected BCC that were included in our recent randomized controlled trial that was performed between 3 February 2015 and 2 October 2015.[10] Consecutive eligible patients of 18 years and older with a clinically suspected, primary, untreated BCC, regardless of subtype and present for at least one month, were prospectively enrolled at the Department of Dermatology, Academic Medical Centre, University of Amsterdam (coordinating tertiary hospital), and the Department of Dermatology, the Netherlands Cancer Institute (participating tertiary hospital), Amsterdam, the Netherlands. We excluded patients with lesions not suitable for conventional surgical excision, lesions in a high-risk location of the face (H-zone and ears), lesions larger than 20 mm, recurrent BCC, macroscopic ulcerating lesions and those with basal cell naevus syndrome. The study was conducted according to the principles of the Declaration of Helsinki (Fortaleza, Brazil; October 2013) and in accordance with the Dutch Medical Research Involving Human Subjects Act (WMO). The research protocol has been approved by ethics committees at both centres (reference number: NL50112.018.14). All participants gave written informed consent prior to their participation in the study.

Study procedures

All clinically suspected BCCs were surgically excised directly after RCM imaging. Histopathologic confirmation of presence and subtype of BCC with the use of haematoxylin and eosin-stained sections taken from the excision specimen was defined as the reference standard. No punch biopsies were performed on the RCM cases.

Reflectance confocal microscopy imaging was performed according to a standardized protocol to diagnose clinically suspected primary BCC and to divide between subtypes during the trial period. A horizontal map of 4 × 4 mm (VivaBlock) was made at the level of the stratum corneum, stratum spinosum and papillary dermis. Vertical mapping (VivaStack) was performed by capturing a series of images of 0.5 × 0.5 mm in depth with steps of 4.5 μm. The mapping started at the top of the stratum corneum until the papillary dermis. Movies were made at the level of the dermal–epidermal junction to visualize capillary blood flow.

To differentiate between BCC subtypes, the following RCM criteria were used: presence of fine telangiectasia, multiple erosions, leaf-like structures, cords connected to the epidermis and epidermal streaming were characteristic features for superficial BCC. Basaloid island nests with peripheral palisading, clefting and increase in vascular diameter without cords connected to the epidermis were characteristic RCM features for nodular and micronodular BCC. The size and shape of the tumour nests allowed further distinction between these subtypes. The absence of small or big tumour islands as well as cords connected to the epidermis with dark silhouettes were characteristic features for infiltrative BCC.[11]

DK performed RCM imaging at the Academic Medical Center and YE at the Netherlands Cancer Institute. The VivaScope 1500® (VivaScope 1500®; Caliber ID, Henrietta, NY, U.S.A.) was used to acquire confocal images at both participating study centres. Each case for evaluation had horizontal optical RCM images at different levels of the skin, including one image at the granular spinous layer of the epidermis, one image at the basal layer of the epidermal and dermal–epidermal junction and one image at the superficial dermis. BCCs were divided into superficial, nodular and aggressive subtypes (i.e. micronodular, infiltrating or basosquamous).

Inter-rater and intrarater agreement

Inter-rater agreement was defined as the extent to which the interpretation of the selected RCM images is the same for repeated measurement by different persons on the same occasion.[12] Intrarater agreement was defined as the extent to which the interpretation of the selected RCM images is the same for repeated measurement by the same persons on different occasions.[12] Three raters (DK, YE and MP) independently reviewed all de-identified images of RCM cases twice, with a 4-week time interval. Raters reviewed the RCM images, without any information on clinical data, clinical photos or dermoscopic pictures. They were also blinded to their own previous interpretation and to each other's interpretation. Before the first and second assessments, RCM cases were shuffled and re-coded by a computer-based system (GraphPad Software, La Jolla, CA) to prevent identification. Diagnoses were recorded on standardized study forms including BCC presence (yes or no) and BCC subtype (superficial, nodular, aggressive or any combination). In addition, raters scored the images as easy, moderate or difficult to diagnose. At the time of the study, raters had between 1 and 5 years of experience with RCM.

Statistical analysis

For assessing inter- and intrarater agreement on BCC presence and BCC subtype, the percentages of observed agreement (i.e. $(a + d)/(a + b + c + d)$ and specific agreement (i.e. positive agreement (PA): PA = $2a/[2a + b + c]$PA = $2a/[2a + b + c]$ or PA = $a/[2a + (b + c)/2]$PA = $a/[2a + (b + c)/2]$; or negative agreement (NA): NA = $2d/[2d + b + c]$NA = $2d/[2d + b + c]$ or PA = $d/[d + (b + c)/2]$PA = $d/[d + (b + c)/2]$)) were calculated.[13] The proportion of specific agreement distinguishes agreement on positive or negative scores. To obtain an agreement parameter for three raters, all pairwise 2 × 2 tables (i.e. m $(m - 1)/2$) were summed and the b and c cells were averaged. Herewith, placement of the b- and c-cell values remains arbitrary. Subsequently, the observed agreement and specific agreement were calculated. We predefined an observed and/or specific agreement of more than 80% to be acceptable. In addition, the 95% confidence intervals were obtained by bootstrap resampling.

Results

In total, 288 RCM assessments were analysed; 48 RCM cases were reviewed two times by 3 raters. The reference standard of the 48 RCM cases revealed 40 BCC (83%), two actinic keratosis (4%), two Bowen's disease (4%), one squamous cell carcinoma (SCC) (2.0%), one naevus (2%), one solar lentigo (2%) and two other non-malignant inflammatory lesions (4%). Of the BCC, 17 (35%) had a superficial subtype, 17 (35%) had a nodular subtype and 6 (13%) had an aggressive subtype. Mixed subtypes were seen in nine (23%) of 40 BCC. Reference standard and BCC subtypes are summarized in Table 1.

Table 1 Tumour and patient characteristics of the 48 RCM cases

	RCM cases, n = 48 (%)
Age (years)	64 (39–84)
Sex	
Men	30 (62%)
Women	18 (38%)
Fitzpatrick skin type	
I	8 (17%)
II	30 (63%)
III	10 (20%)
BCC in medical history	
Yes	32 (67%)
No	15 (31%)
Immunocompromised*	
Yes	2 (4%)
No	46 (96%)
Tumour diameter (mm)	8 (3–15)
Tumour location	
Head/neck	8 (17%)
Trunk	31 (65%)
Arm	4 (8%)
Leg	5 (10%)
Number of BCC	40 (83%)
BCC subtype distribution†	
Superficial BCC	17 (43%)
Nodular BCC	17 (43%)
Aggressive BCC	6 (14%)
Number of non-BCC	8 (17%)
Actinic keratosis	2 (4%)
Bowen's disease	2 (4%)
SCC	1 (2%)
Non-malignant	4 (8%)

Continuous variables are expressed as mean (range) and categorical variables as n (%).
BCC, basal cell carcinoma.
*Patients who were taking immunosuppressive drugs such as oral steroids, methotrexate, ciclosporin for suppression of immunological disorder, or to prevent transplant rejection.
†This number represents the histologically confirmed basal cell carcinoma based on surgical excision specimen. Basal cell carcinoma subtype distribution according to the most aggressive subtype found at histology of surgical excision.

Description of RCM diagnosis at both reviewing sessions

At the first rating session (T1), the three raters diagnosed 'BCC presence' correctly (equally to reference standard) in 34–39 of 40 (mean 89%, range 85–98%) compared to 34–36 at the second rating session (T2) (mean 88%, range 85–92%) (Table 2). Of the correctly diagnosed BCCs, raters accurately diagnosed an aggressive subtype at T1 in 43–59% compared to 46–64% at T2. At T1, the raters scored 14–27% of the RCM cases as 'difficult to diagnose' compared to 13–40% at T2. Examples of RCM cases with good and poor inter- and intrarater agreement on BCC diagnosis and subtypes are shown in Figures 1 and 2.

Inter-rater agreement of RCM images in diagnosing and subtyping BCC

With three raters, we created three 2 × 2 tables ($m(m - 1)/2 = 3 \times 2/2 \times 3$), representing the agreement between the raters: 1 vs 2; 1 vs 3; 2 vs 3. For BCC presence, calculating the proportions of observed agreement for the summed table with averaged b and c cells results in an observed agreement of 82% at T1 (95% CI = 74–92%) and 85% at T2 95% (CI = 76–92%]). The observed agreements for BCC subtype were lower (52% (95% CI = 42–63%) for T1 and 47% (95% CI = 35–58%) for T2).

The specific agreements on a positive score for BCC presence were high at both reviewing sessions 89% (95% CI = 82–94%) for T1 and 90% (95% CI = 84–95%) for T2), but the specific agreements on a negative score were lower (54% (95% CI = 30–71%) for T1 and 66% (95% CI = 40–82%) for T2). The specific agreements for BCC subtyping were also lower.

Table 2 Description of the three rates and their RCM diagnosis at both reviewing sessions

Raters	RCM experience (years)	BCC present, *n* = 40 (%)	Correct BCC subtype (%)	Difficult to diagnose RCM images (%)*
DK at T1	1	34 (85)	17 of 34 (50)	6 of 43 (14)
YE at T1	2	39 (98)	23 of 39 (59)	6 of 35 (17)
MP at T1	5	35 (88)	15 of 35 (43)	12 of 45 (27)
DK at T2	1	35 (88)	16 of 35 (46)	6 of 47 (13)
YE at T2	2	36 (92)	26 of 36 (64)	10 of 47 (21)
MP at T2	5	34 (85)	18 of 34 (53)	19 of 47 (40)

BCC, basal cell carcinoma.
*This item was not recorded by the raters in all 48 RCM cases at both reviewing sessions.

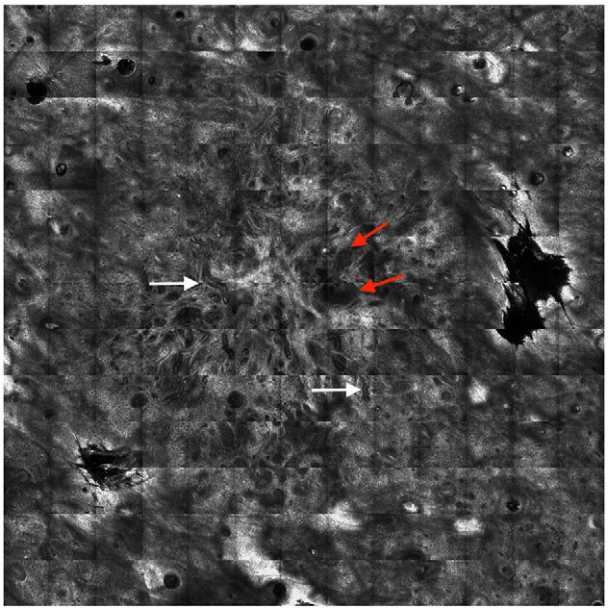

Figure 1 Example study case with good inter- and intrarater agreement. RCM overview image (mosaic) of the papillary dermis of a histology confirmed (excision specimen) nodular/micronodular mixed-type BCC on the left cheek. In the centre of the mosaic, an increase of (enlarged) blood vessels is seen (white arrows) in the presence of varying sized tumour nests (red arrows). All three raters accurately diagnosed BCC and recognized the most aggressive subtype (micronodular growth pattern) at both reviewing sessions (T1 and T2).

Intrarater agreement of RCM images in diagnosing and subtyping BCC

The observed agreements within the raters for BCC presence were 79% (95% CI = 67–90%) for rater DK, 92% (95% CI = 83–98%) for rater YE and 88% (95% CI = 77–96%) for rater MP. For BCC subtyping, the observed agreements within the raters were 56% (95% CI = 42–69%) for rater DK, 71% (95% CI = 58–83%) for rater YE and 57% (95% CI = 44–70%) for rater MP.

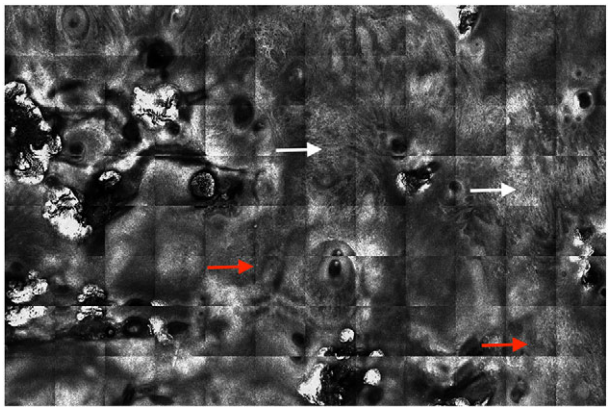

Figure 2 Example study case with poor inter- and intrarater agreement. RCM overview image (mosaic) of the spinous-granular layer of a histology confirmed (excision specimen) well-differentiated squamous cell carcinoma (SCC) on the right cheek. An atypical honeycomb is seen (white arrows) with round nuclear cells (red arrows). In the dermal papilla, enlarged round blood vessels were seen (not shown on mosaic). None of three raters were able to diagnose SCC at the reviewing sessions (T1 and T2). Furthermore, all three raters had a different diagnosis at T1 and T2 corresponding to a poor intrarater agreement.

Discussion

In this study, the inter-rater and intrarater agreement of RCM in correctly diagnosing and subtyping BCC was assessed based on static RCM images. Our results show that RCM is reliable in correctly diagnosing BCC. The observed inter-rater agreements for BCC presence were higher than 80% in both reviewing sessions. The observed intrarater agreement of the three raters for BCC presence ranged from 79% to 92%. This confirms previous findings on the usefulness of RCM in accurately diagnosing BCC.[14] As for subtyping BCC, we found that inter- and intrarater agreements were lower than 80%. The lower agreements for subtyping BCC seem consistent with the results of our recent diagnostic accuracy study.[15] Thus far, only two other studies have previously reported on subtype-specific *in vivo* RCM features.[16,17] The challenges for RCM users to accurately divide between BCC subtypes might be explained by the absence of studies that have reported on the reliability of individual subtype-specific RCM features. Furthermore, the limited detection depth of the RCM technique (up to 200 μm) remains a potential pitfall for accurate BCC subtyping.

This is the first prospective study that investigates the inter-rater and intrarater agreement of RCM in correctly diagnosing and subtyping BCC. Farnetani *et al.*[18] also previously reported on reproducibility of RCM feature recognition and accuracy of diagnosing skin cancer. In their retrospective web-based study,

Cohen's kappa was used to test the interobserver reproducibility of recognition of previously published RCM descriptors for melanoma and BCC. In line with their findings, we found RCM to be reliable in diagnosing BCC. However, Farnetani *et al.* did not report on the reliability of RCM in dividing BCC into subtypes. Besides, the use of Cohen's kappa is less informative for clinicians as it is considered to be a measure of reliability and not a measure of agreement.[19] In clinical practice, we are interested in inter-rater and/or intrarater agreement. Following the methods of De Vet and colleagues,[13] we therefore decided to calculate the proportion of observed and specific agreement instead of Cohen's kappa. We believe that this is one of the strengths of our study. Another study strength is the use of de-identified static RCM images to prevent interpretation bias of the raters as a result of clinical information.

Limitations of our study include a selection bias of RCM cases. The series of images were derived from our recent randomized controlled trial that excluded lesions not suitable for conventional surgical excision, lesions in a high-risk location of the face (H-zone and ears), lesions larger than 20 mm, recurrent BCC, macroscopic ulcerating lesions and lesions of patients with basal cell naevus syndrome.[10] In addition, two different researchers (DK and YE) performed RCM imaging during the study period leading to a potential source of bias in acquiring the series of RCM images. Another study limitation includes the limited number of cases. In 40 of the 48 RCM cases, a BCC was histologically confirmed in surgical excision specimen. Of those, only six BCCs had an aggressive subtype.

In terms of external validity, it is important to emphasize that our study results are based on the interpretation of static RCM images that were acquired with the VivaScope 1500®. There is an important difference in diagnosing and subtyping clinically suspected BCC using real-time *in vivo* RCM combined with clinical information and dermoscopy compared to the blinded static RCM images that were assessed in our study. As demonstrated by Borsari *et al.*,[20] RCM should ideally be used as an add-on tool to clinical inspection and dermoscopy to increase accuracy in the diagnosis of skin cancer. Therefore, future research should be aimed at investigating the reliability of real-time RCM as it is expected to further improve RCM's inter-rater and intrarater agreement for diagnosing and subtyping BCC.

Reflectance confocal microscopy could potentially play an important role in the management of BCC if accurate subtyping will be achieved. We recommend on achieving international consensus on specific RCM features for subtyping BCC based on the results of large prospective clinical trials. For example, currently ongoing randomized controlled multicentre trial in Nijmegen, the Netherlands, that has been designed to investigate whether *in vivo* RCM can correctly identify the subtype of BCC.[21] Furthermore, using the more recently introduced flexible handheld VivaScope 3000® RCM (VivaScope 3000®; Caliber ID, Henrietta, NY, U.S.A.), clinically suspected BCC can be evaluated even faster. Previous studies already confirmed that the VivaScope 3000® is suitable in diagnosing BCC, including lesions on the more concave and convex high-risk head and neck areas.[22,23] It would be valuable to compare the reliability of the wide probe VivaScope 1500® with the VivaScope 3000® for accurately subtyping BCC.

In conclusion, our results show that RCM is reliable in correctly diagnosing BCC based on the assessment of static RCM images. RCM could potentially play an important role in BCC management if accurate subtyping will be achieved. Therefore, future clinical studies on reliability and specific RCM features for BCC subtypes are required.

References

1 Verkouteren JA, Ramdas KH, Wakkee M, Nijsten T. Epidemiology of basal cell carcinoma: scholarly review. *Br J Dermatol* 2017; **177**: 359–372.
2 Kelleners-Smeets NW, Mosterd K, Nelemans PJ. Treatment of low-risk basal cell carcinoma. *J Invest Dermatol* 2017; **137**: 539–540.
3 Trakatelli M, Morton C, Nagore E *et al.* Update of the European guidelines for basal cell carcinoma management. *Eur J Dermatol* 2014; **24**: 312–329.
4 Telfer NR, Colver GB, Morton CA. British Association of D. Guidelines for the management of basal cell carcinoma. *Br J Dermatol* 2008; **159**: 35–48.
5 Rossi AM, Sierra H, Rajadhyaksha M, Nehal K. Novel approaches to imaging basal cell carcinoma. *Future Oncol* 2015; **11**: 3039–3046.
6 Giavedoni P, Puig S, Carrera C. Noninvasive imaging for nonmelanoma skin cancer. *Semin Cutan Med Surg* 2016; **35**: 31–41.
7 Rajadhyaksha M, Grossman M, Esterowitz D, Webb RH, Anderson RR. In vivo confocal scanning laser microscopy of human skin: melanin provides strong contrast. *J Invest Dermatol* 1995; **104**: 946–952.
8 Hoorens I, Vossaert K, Ongenae K, Brochez L. Is early detection of basal cell carcinoma worthwhile? Systematic review based on the WHO criteria for screening. *Br J Dermatol* 2016; **174**: 1258–1265.
9 Edwards SJ, Mavranezouli I, Osei-Assibey G, Marceniuk G, Wakefield V, Karner C. VivaScope(R) 1500 and 3000 systems for detecting and monitoring skin lesions: a systematic review and economic evaluation. *Health Technol Assess* 2016; **20**: 1–260.
10 Kadouch DJ, Elshot YS, Zupan-Kajcovski B *et al.* One-stop-shop with confocal microscopy imaging versus standard care for surgical treatment of basal cell carcinoma: an open label, non-inferiority, randomized controlled multicenter trial. *Br J Dermatol* 2017; **177**: 735–741.
11 Kadouch DJ, Wolkerstorfer A, Elshot Y *et al.* Treatment of basal cell carcinoma using a one-stop-shop with reflectance confocal microscopy: study design and protocol of a randomized controlled multicenter trial. *JMIR Res Protoc* 2015; **4**: e109.
12 Mokkink LB, Terwee CB, Gibbons E *et al.* Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) checklist. *BMC Med Res Methodol* 2010; **10**: 82.
13 de Vet HCW, Dikmans RE, Eekhout I. Specific agreement on dichotomous outcomes can be calculated for more than two raters. *J Clin Epidemiol* 2017; **83**: 85–89.
14 Que SK, Grant-Kels JM, Longo C, Pellacani G. Basics of confocal microscopy and the complexity of diagnosing skin tumors: new imaging tools in clinical practice, diagnostic workflows, cost-estimate, and new trends. *Dermatol Clin* 2016; **34**: 367–375.
15 Kadouch DJ, Leeflang MM, Elshot YS *et al.* Diagnostic accuracy of confocal microscopy imaging versus punch biopsy for diagnosing and subtyping basal cell carcinoma. *J Eur Acad Dermatol Venereol* 2017; **31**: 1641–1648.

16 Longo C, Lallas A, Kyrgidis A *et al.* Classifying distinct basal cell carcinoma subtype by means of dermatoscopy and reflectance confocal microscopy. *J Am Acad Dermatol* 2014; **71**: 716–24e1.
17 Peppelman M, Wolberink EA, Blokx WA, van de Kerkhof PC, van Erp PE, Gerritsen MJ. In vivo diagnosis of basal cell carcinoma subtype by reflectance confocal microscopy. *Dermatology* 2013; **227**: 255–262.
18 Farnetani F, Scope A, Braun RP *et al.* Skin cancer diagnosis with reflectance confocal microscopy: reproducibility of feature recognition and accuracy of diagnosis. *JAMA Dermatol* 2015; **151**: 1075–1080.
19 de Vet HC, Mokkink LB, Terwee CB, Hoekstra OS, Knol DL. Clinicians are right not to like Cohen's kappa. *BMJ* 2013; **346**: f2125.
20 Borsari S, Pampena R, Lallas A *et al.* Clinical indications for use of reflectance confocal microscopy for skin cancer diagnosis. *JAMA Dermatol* 2016; **152**: 1093–1098.
21 Peppelman M, Nguyen KP, Alkemade HA *et al.* Diagnosis of basal cell carcinoma by reflectance confocal microscopy: study design and protocol of a randomized controlled multicenter trial. *JMIR Res Protoc* 2016; **5**: e114.
22 Cinotti E, Jaffelin C, Charriere V *et al.* Sensitivity of handheld reflectance confocal microscopy for the diagnosis of basal cell carcinoma: A series of 344 histologically proven lesions. *J Am Acad Dermatol* 2015; **73**: 319–320.
23 Castro RP, Stephens A, Fraga-Braghiroli NA *et al.* Accuracy of in vivo confocal microscopy for diagnosis of basal cell carcinoma: a comparative study between handheld and wide-probe confocal imaging. *J Eur Acad Dermatol Venereol* 2014; **29**: 1164–1169.