Abstract

Purpose

It has previously been shown that the intervals between screening examinations for diabetic retinopathy can be optimized by including individual risk factors for the development of the disease in the risk assessment. However, in some cases, the risk model calculating the screening interval may recommend a different interval than an experienced clinician. The purpose of this study was to evaluate the influence of factors unrelated to diabetic retinopathy and the distribution of lesions for discrepancies between decisions made by the clinician and the risk model.

Methods

Therefore, fundus photographs from 90 screening examinations where the recommendations of the clinician and a risk model had been discrepant were evaluated. Forty features were defined to describe the type and location of the lesions, and classification and ranking techniques were used to assess whether the features could predict the discrepancy between the grader and the risk model.

Results

Suspicion of tumours, retinal degeneration and vascular diseases other than diabetic retinopathy could explain why the clinician recommended shorter examination intervals than the model. Additionally, the regional distribution of microaneurysms/dot haemorrhages was important for defining a photograph as belonging to the group where both the clinician and the risk model had recommended a short screening interval as opposed to the other decision alternatives.

Conclusions

Features unrelated to diabetic retinopathy and the regional distribution of retinal lesions may affect the recommendation of the examination interval during screening for diabetic retinopathy. The development of automated computerized algorithms for extracting information about the type and location of retinal lesions could be expected to further optimize examination intervals during screening for diabetic retinopathy.

Introduction

Diabetic retinopathy is a leading cause of blindness in the Western world. The disease is detected by screening which encompasses visual acuity measurement and fundus photography (Stefánsson et al. **2000**; Jeppesen & Bek **2004**; WHO **2008**; Bandello et al. **2013**). On the basis of the diabetes type and the severity of retinopathy, a fixed interval until the following screening examination is defined to ensure that no patient will develop vision-threatening complications during that period (Singer et al. **1992**; American Academy of Ophthalmology Retina Panel **2012**). However, this rule-based approach also implies that patients with slow disease progression will undergo a number of superfluous examinations with no consequences for the

management of the disease. A reduction in the number of these examinations requires algorithms that consider the patient's individual risk factors (Aspelund et al. 2011). A newly developed decision model considering risk factors such as sex, age at onset of diabetes mellitus, diabetes type, diabetes duration, HbA1c and blood pressure can allow a significant prolongation of the control interval without increasing the risk of developing vision-threatening retinopathy (Mehlsen et al. 2012). The predictive value of this model was found to be lowest for patients with more severe retinopathy, and it was concluded that the grader might have included other factors than the severity of retinopathy in the decision, such as the location of retinopathy lesions or conditions unrelated to diabetic retinopathy (Bek & Helgesen 2001; Hove et al. 2004, 2006).

Therefore, 90 randomly selected screening examinations, where the prediction of the model matched exactly or deviated significantly from the clinician's recommended interval, were identified. The fundus photographs obtained at these examinations were analysed to investigate the role of the location and number of retinopathy lesions and the presence of retinal lesions unrelated to DR for the assessment of screening intervals.

Materials and Methods

Data set

The study was based on 8987 consecutive examinations of both eyes performed on 3572 patients with diabetes attending the screening clinic for diabetic retinopathy at the Department of Ophthalmology, Aarhus University Hospital, between 1 January 2004 and 31 December 2007. Each screening examination started with a registration of body weight and height, previous medical history and family history of eye-related diseases. The patients were defined as having type 1 diabetes mellitus (age of onset <30 years requiring insulin within 1 year from the diagnosis) or type 2 diabetes mellitus (all other patients). Resting blood pressure was measured using an electronic sphygmomanometer (Omron M4, HEM-722c1-E Omron, Kyoto, Japan), and LogMAR visual acuity was measured according to ETDRS principles (Mehlsen et al. 2011). Subsequently, mydriasis was induced by tropicamide 1% (Alcon, Copenhagen, Denmark) and phenylephrine 2.5% eyedrops (SAD, Danish Hospital Pharmacies, Skanderborg, Denmark). After waiting for approximately 30 minutes, two 60° fundus photographs were taken from each eye, one centred at the fovea and another at the optic disc (OD). Before 1 March 2002, the photographs were documented on Kodak Ectachrome 64 colour diapositive film using a Canon CF 60UV fundus camera (Canon, Tokyo, Japan), and after this date, the photographs were documented using a digital unit (FinePix S1 Pro, FUJIFILM; Minato, Tokyo, Japan) mounted on the same fundus camera. On the basis of these data, a grader used a rule-based approach in accordance with general clinical guidelines (Kristinsson et al. 1995; American Academy of Ophthalmology Retina Panel 2012) to recommend one of the following fixed screening intervals: 3, 6, 12, 24 or +36 months ($I_{grader.}$).

The interval recommended by the grader was compared to the outcome of a multinomial logistic regression model (I_{model}). The outcome of the model was defined as the five intervals (3, 6, 12, 24 or +36 months) considering the patient's individual risk factors: diabetes duration, gender, age at diagnosis of diabetes, haemoglobin A1c, the number of retinal haemorrhages and hard exudates counted on the macula centred 60° fundus photograph (Mehlsen et al. 2012).

Of the 8987 visits from 3572 patients, the prediction of the screening interval by the two methods was concordant in 6691 visits (74.4%) and discordant in the remaining 2296 visits (25.6%) from 1553 patients. To obtain a subset of observations with the highest information content with respect to the concordance and discordance of the two models, only the visits with a recommendation of the two longest (L, 24 or +36 months) or the shortest (S, 3 months) intervals, by I_{grader} and $I_{model,}$ respectively, were selected. This selection reduced the data set to 609 visits. In patients who had had multiple visits, one of these visits was selected at random to represent the patient in the data material. The resulting 426 visits from the same number of patients were allocated to four groups as shown in Table 1. Subsequently, a random sample of the observations was selected in order to avoid biases in the results due to the different sample sizes. The number of visits assigned by both the grader and the model to the longest ($L_{grader}L_{model}$) and to the shortest ($S_{grader}S_{model}$) intervals were standardized to equal the size of the group of visits assigned to the shortest interval by the grader but the longest interval by the model ($S_{grader}L_{model}$), consisting of 30 observations. The group of visits assigned to the longest interval by the grader but the shortest interval by the model ($L_{grader}S_{model}$) was discarded from the analysis due to the limited number of observations (4) that could have altered the results (Wei & Dunbrack **2013**) and due to the minimum risk that this case represents for patients in screening programmes for diabetic retinopathy. The balancing process resulted in 90 visits, 30 from each of the three groups.

Table 1. Number of visits for the four subgroups in the selection process

	Longest/shortest	**Final selection**
1. L_gL_m = longest interval by both the grader and the model – concordant, S_gS_m = shortest interval by both the grader and the model – concordant, S_gL_m = shortest interval by the grader and longest interval by the model – discordant, L_gS_m = longest interval by the grader and shortest interval by the model – discordant.		
L_gL_m	227	30
S_gS_m	165	30

	Longest/shortest	Final selection
S_gL_m	30	30
L_gS_m	4	Excluded
Total	426	90

Image analysis

Twenty-two of the 90 visits from the three groups ($S_{grader}L_{model}$, $L_{grader}L_{model}$, $S_{grader}S_{model}$) had been performed before the introduction of digital fundus photography, and the retina was therefore documented on diapositives. These diapositives were scanned using an HP G4025 scanner (Hewlett-Packhard, Palo Alto, CA, USA) in a resolution of 1200 PPI and True Colour format (24 bit, 256 shades of red, green, blue for a total of 16 777 216 colour variations).

A software tool was created using the open-source image-processing software imagej (US National Institutes of Health, Bethesda, MD, USA) to allow the marking of individual lesions in fundus photographs and store the co-ordinates of the central pixel in each lesion for later analysis. The first author marked the four types of lesions manually, that is dot haemorrhages/microaneurysms (DH/MA) with a diameter smaller than the diameter of the temporal venules at the crossing of the OD, large haemorrhages (all other haemorrhages), hard exudates and cotton wool spots. A second software tool was created to allow the marking of two circles around the fovea (C1 and C2) and three ellipses (E1, E2 and E3) delimiting areas with known accumulation of DR lesions (Hove et al. **2004**), and to save these markings in a separate file associated with each fundus photograph. The last author subsequently reviewed the grading of retinopathy lesions, and in case of discrepancy in the two gradings, his opinion was followed.

Features Selection

In each fundus photograph from each of the three groups, the number and the percentage of each of the four lesion types located within each of the five areas (altogether 40 features) were calculated, and when a lesion type was not present in the photograph, a value of −1 was assigned to the corresponding features.

Classification

The photographs in the three groups were combined with the photographs from each of the two other groups to result in three-two-group data sets. For each of these new data sets, a Naïve Bayes classifier (John & Langley **1995**) implemented by the weka software (version 3-7-10, University of Waikato, Hamilton, New Zealand) was used to assess whether the 40 features could predict the photograph's original group (Hall et al. **2009**). The final classification of a Naïve Bayes classifier is produced by calculating the posterior probability of a photograph as belonging to one of two groups and by assigning it to the group with the highest probability. The results were reported as the percentages of the correctly and incorrectly classified observations. The false-positive rate was plotted against the true-positive rate to represent a receiver operation characteristics (ROC) curve. The ability of the classifier to discriminate photographs from the two-group data sets as belonging to either set was expressed by the area under the ROC curve (AUC), where a perfect discrimination would result in AUC = 1 and lack of discrimination would result in AUC = 0.5

Ranking

Subsequently, in each of the two-group data sets, an information gain value between zero and one was calculated for each of the 40 features using weka software, which expressed how this feature contributed to identifying a photograph as belonging to the correct of the original groups. The values were ranked from largest to smallest number in order to provide an overview of the contribution of individual features to the differentiation between the groups.

Re-evaluation of the photographs by a retina specialist

The photographs from all patients were blinded and were regraded by a retina specialist Toke Bek (TB) who noted possible lesions other than diabetic retinopathy that might have affected the grader's decision about control interval. The occurrence of such lesions was compared among the groups.

Cumulative representation of lesions

Finally, to visualize the overall distribution of DH/MAs, the images were resized so that the fovea (F) and the OD could be superimposed. For each lesion in each image, a bell-shaped two-dimensional Gaussian with unity height and a standard deviation of 0.05, corresponding to three times the average DH/MA diameter, was plotted to resemble the intensity curve of an average DH/MA lesion. Subsequently, from both the right and the left eye, all lesion-centred intensity curves from all maps were added and plotted on a template containing the five areas of interest. Finally, in the template from each eye, the accumulated intensity values were normalized and represented with a colour scale ranging from dark blue (0) to dark red (1).

Results

Table 2 shows the results of the 10-fold cross-validation. It appears that the 40 features could be used to obtain a satisfactory classification of photographs belonging to either $L_{grader}L_{model}$ or $S_{grader}S_{model}$ and photographs belonging to either $S_{grader}S_{model}$ or $S_{grader}L_{model}$ with AUC values larger than 0.9, whereas the

classification of photographs as belonging to either group $L_{grader}L_{model}$ or $S_{grader}L_{model}$ was unsuccessful with AUC values just above 0.6. This confirms that the number and location of the lesions cannot explain the clinician's selection of a short interval when the risk algorithm selects a long screening interval.

Table 2. Classification results for the three cases

L_gL_m versus S_gS_m		**S_gS_m versus S_gL_m**		**L_gL_m versus S_gL_m**	
1. L_gL_m = longest interval by both the grader and the model – concordant, S_gS_m = shortest interval by both the grader and the model – concordant, S_gL_m = shortest interval by the grader and longest interval by the model – discordant. Success rate and area under the ROC curve (AUC).					
Correctly classified 94.8%		Correctly classified 88.2%		Correctly classified 60.8%	
Incorrectly classified 5.1%		Incorrectly classified 11.5%		Incorrectly classified 39.2%	
AUC	0.972	AUC	0.965	AUC	0.631

Table 3 shows the 40 features ranked by the information gain value in the three-two-group data sets. It appears that the number of MA/DH within the studied regions with values higher than 0.63 was the most important feature for defining a photograph as belonging to the data set $S_{grader}S_{model}$ as opposed to one of the other data sets. Conversely, large haemorrhages were the most important feature for defining a photograph as belonging to either the $L_{grader}L_{model}$ or the $S_{grader}L_{model}$ data set, but the values were lower than 0.12. This implies that when the risk algorithm had assigned the longest interval and the clinician made a different choice, this choice had been almost independent of the retinopathy grade as expressed by the number of the lesions and their location within the five areas.

Table 3. The 40 features arranged according to information gain value for each comparison

Rank	L_rL_i–S_rS_i		S_rS_i–S_rL_i		L_rL_i–S_rL_i	
	Info. gain value	**Feature**	**Info. gain value**	**Feature**	**Info. gain value**	**Feature**
1. Features are named with the following rule: areaLabel + measure + lesionType, where areaLabel = [E1, E2, E3, C1, C2], measure = [*n*, p] (*n* = count, p = percentage) and lesionType = [sHae, lHae, cott, exu] (sHae = small haemorrhages, lHae = large haemorrhages, cott = cotton wool, exu = exudates).						
1	0.734	E3n_sHae	0.712	E2n_sHae	0.114	C1p_lHae
2	0.725	E2n_sHae	0.641	E1n_sHae	0.114	C1n_lHae
3	0.722	E1n_sHae	0.553	E3n_sHae	0.082	E1p_lHae
4	0.634	E3p_sHae	0.538	E2p_sHae	0.082	C2p_lHae
5	0.634	E1p_sHae	0.517	E2n_exu	0.082	E2n_lHae

Rank	L_rL_i–S_rS_i		S_rS_i–S_rL_i		L_rL_i–S_rL_i	
	Info. gain value	**Feature**	**Info. gain value**	**Feature**	**Info. gain value**	**Feature**
6	0.634	C1n_sHae	0.489	E1p_sHae	0.082	E3n_lHae
7	0.634	C2n_sHae	0.486	E3p_sHae	0.082	E2p_lHae
8	0.634	E2p_sHae	0.484	C2n_sHae	0.082	C2n_lHae
9	0.634	C2p_sHae	0.401	E1p_exu	0.082	E1n_lHae
10	0.634	C1p_sHae	0.401	C1n_exu	0.082	E3p_lHae
11	0.550	E2n_exu	0.401	C1p_exu	0.079	C1n_cott

Rank	L_rL_i–S_rS_i		S_rS_i–S_rL_i		L_rL_i–S_rL_i	
	Info. gain value	**Feature**	**Info. gain value**	**Feature**	**Info. gain value**	**Feature**
12	0.540	C1p_exu	0.401	E3n_exu	0.079	E2n_cott
13	0.540	C1n_exu	0.401	E3p_exu	0.079	E2p_cott
14	0.540	E1n_exu	0.401	E2p_exu	0.079	E1p_cott
15	0.527	C2n_exu	0.401	C2n_exu	0.079	C1p_cott
16	0.468	E2p_exu	0.401	C2p_exu	0.079	E1n_cott
17	0.449	E3p_exu	0.401	E1n_exu	0.079	C2p_cott

Rank	L_rL_i–S_rS_i		S_rS_i–S_rL_i		L_rL_i–S_rL_i	
	Info. gain value	**Feature**	**Info. gain value**	**Feature**	**Info. gain value**	**Feature**
18	0.449	E3n_exu	0.389	C2p_sHae	0.079	E3p_cott
19	0.449	C2p_exu	0.389	C1n_sHae	0.079	E3n_cott
20	0.449	E1p_exu	0.389	C1p_sHae	0.079	C2n_cott
21	0.429	E2p_lHae	0.162	E2n_lHae	0	E1p_exu
22	0.429	E1n_lHae	0.162	E1p_lHae	0	E1p_sHae
23	0.429	E1p_lHae	0.162	E1n_lHae	0	E1n_exu

Rank	L_rL_i–S_rS_i		S_rS_i–S_rL_i		L_rL_i–S_rL_i	
	Info. gain value	**Feature**	**Info. gain value**	**Feature**	**Info. gain value**	**Feature**
24	0.429	E2n_lHae	0.162	E2p_lHae	0	C1p_exu
25	0.429	C1p_lHae	0.162	C1p_lHae	0	E1n_sHae
26	0.429	C2n_lHae	0.162	C2n_lHae	0	E3p_sHae
27	0.429	C2p_lHae	0.162	C2p_lHae	0	E3n_sHae
28	0.429	E3n_lHae	0.162	E3n_lHae	0	C2n_exu
29	0.429	E3p_lHae	0.162	E3p_lHae	0	C2p_exu

Rank	L_rL_i–S_rS_i		S_rS_i–S_rL_i		L_rL_i–S_rL_i	
	Info. gain value	**Feature**	**Info. gain value**	**Feature**	**Info. gain value**	**Feature**
30	0.429	C1n_lHae	0.162	C1n_lHae	0	C2p_sHae
31	0.307	E1p_cott	0.105	E1p_cott	0	C1n_exu
32	0.307	E3p_cott	0.105	E3p_cott	0	C1n_sHae
33	0.307	E3n_cott	0.105	E3n_cott	0	C1p_sHae
34	0.307	C1n_cott	0.105	C1n_cott	0	E2n_sHae
35	0.307	C1p_cott	0.105	C1p_cott	0	E3n_exu

Rank	L_rL_i–S_rS_i		S_rS_i–S_rL_i		L_rL_i–S_rL_i	
	Info. gain value	**Feature**	**Info. gain value**	**Feature**	**Info. gain value**	**Feature**
36	0.307	E2n_cott	0.105	E2n_cott	0	C2n_sHae
37	0.307	C2n_cott	0.105	C2n_cott	0	E2p_exu
38	0.307	C2p_cott	0.105	C2p_cott	0	E2n_exu
39	0.307	E2p_cott	0.105	E2p_cott	0	E2p_sHae
40	0.307	E1n_cott	0.105	E1n_cott	0	E3p_exu

The re-evaluation of the fundus photographs showed that the images from 17 patients contained conditions unrelated to DR that the grader might have thought should be kept under closer observation (six with a suspected tumour, three with exudative age-related macular degeneration (AMD), two with atrophic AMD, one with geographic atrophy, one with an optociliary shunt, one with a macular hole, one with congenital hypertrophy of the retinal pigmented epithelium, one with cataract and one with vascular pathology of unknown type). All these cases belonged to the $S_{grader}L_{model}$ data set, accounting for 57% of the patients in this group.

Discussion

The development of decision models to individualize the screening interval for diabetic retinopathy should be validated to ensure that all patients requiring shorter screening intervals or treatment are detected (Fu et al. **2016**; Laatikainen et al. **2016**) Furthermore, the algorithm should not recommend longer intervals than the clinician would, ensuring early detection of vision-threatening lesions. In a previously developed decision model, it was found that with a consideration of risk factors such as sex, age, age of onset of diabetes mellitus, HgbA1c and blood pressure, the average control interval could be prolonged 2.8 times for patients with type 1 diabetes and 1.2 times for patients with type 2 diabetes without losing a patient with vision-threatening changes (Mehlsen et al. **2012**). However, the testing of the model also identified a group of patients in whom the algorithm recommended a considerably longer control interval than the grader. It was concluded that this limitation might be due to factors unrelated to diabetic retinopathy or the location of lesions, which has been shown to be an independent risk factor for progression of the disease to a treatment-requiring stage (Hove et al. **2004**, **2006**), but which was not included in the model. Therefore, the purpose of the present study was to investigate the role of these factors for the clinician's recommendation of shorter control intervals than the decision model.

The methodological approach was dual. A classification approach was used to assess whether features describing lesion type and location could be used to predict agreement or not among the grader and the model. The selected features could successfully predict the agreement on short screening intervals. However, when the model had recommended a long screening interval, the features describing the location were unable to predict whether the grader had suggested a short or a long screening interval. The results of the re-evaluation of the fundus photographs by a retina specialist suggest that the imperfections of the model were unrelated to diabetic retinopathy, but more to interference of other potential vision-threatening diseases or fields of interest of the grader in the clinical decision-making. Therefore, an optimization of the screening model requires that these cases with factors unrelated to diabetic retinopathy are excluded from the data material before the algorithm calculates a screening interval based on relevant risk factors.

The purpose of the second approach was to identify which of the features were most relevant for the differentiation between the observation where the grader and the model had or had not agreed on a screening interval. To study causes of discrepancies between screening intervals recommended by the grader and the risk model, a group of patients who had been recommended long screening intervals was included and in whom only few or no retinopathy lesions were identified. This group also inherently contributed little information about the distribution of lesions. Therefore, it is likely that the role of the regional distribution of retinopathy lesions might be further elucidated by studying patient groups selected for this information to be present in the photographs.

The fact that the number of DH/MA located in E1, E2 and E3 was the most important feature for the distinction of short and long screening intervals agreed by both the grader and the model corroborates previous studies that the total number of these lesions is predictive for the progression of diabetic retinopathy (Kohner et al. **1999**; Sjølie et al. **2011**), but suggests that the predictive value is related to lesions developing in specific areas of the fundus. This might be clinically relevant if retinopathy grading could be limited to an identification of lesions in these areas. This hypothesis should be tested in a prospective study.

The significance of the percentages of DH/MA located within C1, C2, E1, E2 and E3 for the differentiation of high- and low-risk patients suggests that the pattern of distribution of these lesions in the ocular fundus has specific characteristics. This hypothesis is supported by a plot of all DH/MA identified in the high-risk patients included in the present study showing a preponderance of lesions temporal in the macular area (Fig. 1). However, this distribution should be described in more detail including a large number of observations and considering the influence of known risk factors for the development and progression of the disease, of which the arterial blood pressure has been shown to correlate with lesions being located around the temporal vascular arcades (Bek & Helgesen **2001**).

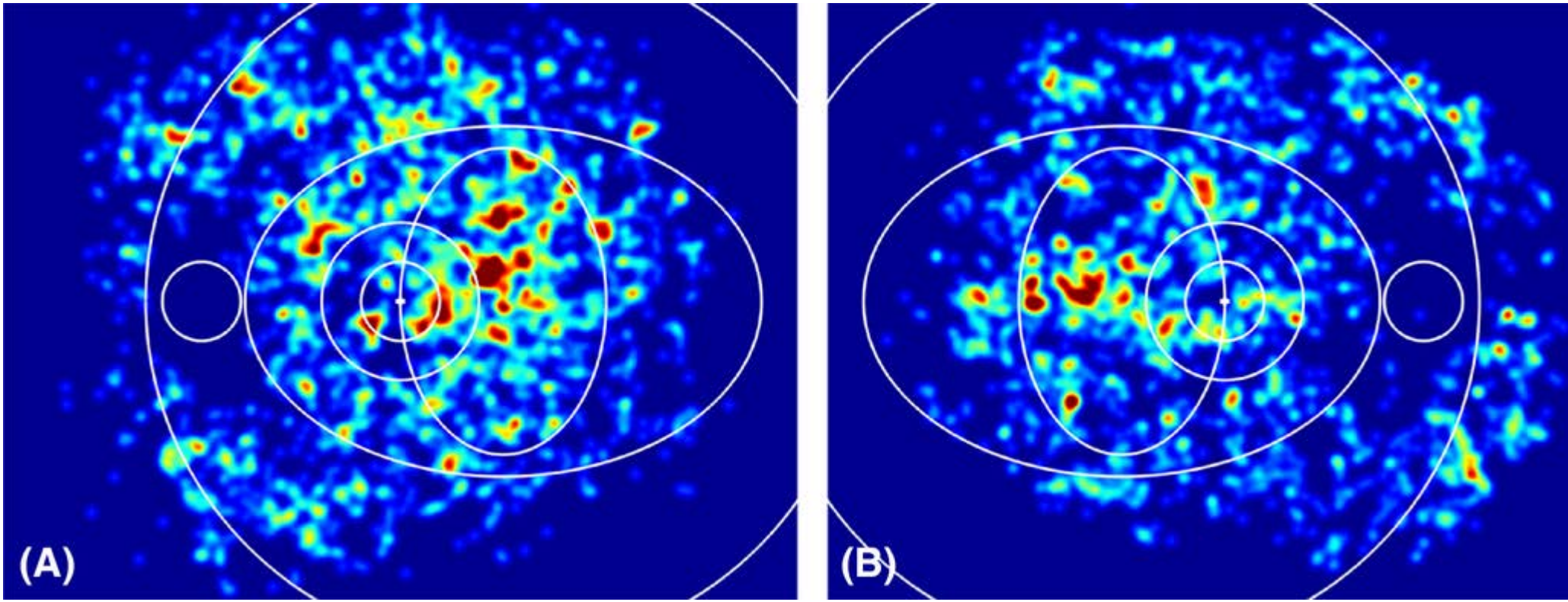

Figure 1.

The cumulated distribution of microaneurysms/dot haemorrhages in the studied fundus photographs, ranging from dark red corresponding to the most frequent occurrence of lesions temporally in the macular area to blue corresponding to no occurrence of lesions. Left eye to the left (a) and right eye to the right (b).

The fact that the location of larger haemorrhages had a lower predictive value suggests that the studied patients had not included a sufficient number of patients with severe retinopathy where the occurrence of these lesions temporal in the macular area was included as part of the definition of preproliferative diabetic retinopathy (Bek **2013**). The inclusion of patients with more advanced stages of diabetic retinopathy might also increase the number of vascular changes (Englmeier et al. **2004**; Broe **2015**) and cotton wool spots, the frequency of which increases with the thickness of the retinal nerve fibre layer (Kim et al. **2011**), but in which also the shape of the lesions has been shown to have diagnostic value (Jaworski **2000**). Similarly, the distribution of hard exudates might have particular interest for the study of the development of diabetic maculopathy.

In conclusion, the study has shown that features unrelated to diabetic retinopathy may affect the grader's recommendation of examination intervals in a screening programme for diabetic retinopathy. Additionally, the results are consistent with previous suggestions that the regional distribution of diabetic retinopathy lesions may improve the calculation of the risk for progression to vision-threatening diabetic retinopathy (Aspelund et al. **2011**; Mehlsen et al. **2012**), but this should be investigated further. The development of automated computerized algorithms for extracting this information on larger data sets can be expected to advance the understanding of the distribution of lesions and its relation to the progression of the disease.