Agreement on classification of clinical photographs of pigmentary lesions: exercise after a training course with young dermatologists

Simone Cazzaniga,[1,2] Lucia De Ponti,[3] Giorgio Maria Baratelli,[4] Salvatore Francione,[5] Carlo La Vecchia,[6,7] Anna Di Landro,[1] Andrea Carugno,[8,9] Marco Di Mercurio,[8] Lerica Germi,[10] Giampaolo Trevisan,[10] Mirko Fenaroli,[4] Claudia Capasso,[5] Michele Pezza,[5] Pietro Dri,[11] Emanuele Castelli,[12] Luigi Naldi[1,10]

[1]Centro Studi GISED, Bergamo, Italy; [2]Department of Dermatology, Inselspital University Hospital of Bern, Switzerland; [3]Italian League for the Fight Against Cancer (LILT), section of Bergamo, Italy; [4]LILT, section of Como, Italy; [5]LILT, section of Benevento, Italy; [6]LILT, section of Milan, Italy; [7]Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy; [8]Dermatology Unit, ASST Papa Giovanni XXIII Hospital, Bergamo, Italy; [9]Ph.D. Program in Molecular and Translational Medicine (DIMET), University of Milano Bicocca, Milan, Italy; [10]Dermatology Unit, San Bortolo Hospital, Azienda ULSS8 Berica, Vicenza, Italy; [11]Zadig, Scientific Publisher, Milan, Italy; [12]Sistemi e Progetti Software (SPS), Bergamo, Italy

Abstract

Smartphone apps may help promoting the early diagnosis of melanoma. The reliability of specialist judgment on lesions should be assessed. Hereby, we evaluated the agreement of 6 young dermatologists, after a specific training. Clinical judgment was evaluated during 2 online sessions, 1 month apart, on a series of 45 pigmentary lesions. Lesions were classified as highly suspicious, suspicious, non-suspicious or not assessable. Cohen's and Fleiss' kappa were used to calculate intra- and inter-rater agreement. The overall intra-rater agreement was 0.42 (95% confidence interval - CI: 0.33-0.50), varying between 0.12-0.59 on single raters. The inter-rater agreement during the first phase was 0.29 (95% CI: 0.24-0.34). When considering the agreement for each category of judgment, kappa varied from 0.19 for not assessable to 0.48 for highly suspicious lesions. Similar results were obtained in the second exercise. The study showed a less than satisfactory agreement among young dermatologists. Our data point to the need for improving the reliability of the clinical diagnoses of melanoma especially when assessing small lesions and when dealing with thin melanomas at a population level.

Introduction

Increasing the awareness of melanoma with the promotion of self-examination by informed people, and early access to dermatological advice for suspected lesions, are possible ways to anticipate the melanoma diagnosis and to improve survival, at a population level, in a sustainable way.[1,2]

Smartphones are largely available in the general population, and may be exploited to transfer clinical images taken by the patient, directly to a physician through an app.[3,4]

In spite of the fact that dermoscopy may improve the clinical classification of pigmentary lesions,[5] the simplest way to use the app for such a purpose by the general public is to transfer photographs of lesions as they appear macroscopically. We already did a validity study on an app called *Clicca il Neo*, comparing distant assessment of such kind of photographs with the direct clinical evaluation of original lesions. A small number of well experienced dermatologists with a high level of documented agreement participated in the study.[6] With the aim of expanding the number of collaborating dermatologists, we also enrolled young dermatologists with a limited level of clinical experience. We conduced a new agreement study on a set of photographs selected among those sent by app users during the above mentioned previous validity study. We evaluated the agreement after an online course aimed at improving the identification and classification of pigmentary lesions.

Materials and Methods

This was an agreement study, conducted after an online course, enrolling a total of six young dermatologists. The online course was organized in collaboration with the Italian League for the Fight Against Cancer (LILT) and the Scientific Publisher Zadig in Milan, in the period February-March 2021. The course was based on an atlas of pigmentary lesions and on several recognition exercises. At the end of the course, the reproducibility and consistency of the clinical judgments was evaluated during 2 online sessions, 1 month apart. During the sessions, the same series of 45 pigmentary lesions were presented with different orders, and participants were asked to classify them as highly suspicious, suspicious, non-suspicious or not assessable. TThese lesions were originally classified by consensus among three experienced dermatologists as highly suspicious (5 lesions, mainly thin melanomas, all confirmed histologically); suspicious, (10 lesions, either thin melanomas or atypical nevi, also documented histologically), non suspicious (25 lesions, clinically classified as a variety of melanocytic nevi or other benign pigmentary lesions), not assessable (5 lesions, where a need for a dermoscopic examination was considered as a pre-requisite).

Correspondence: Simone Cazzaniga, Centro Studi GISED, Via Torino 13, 4128 Bergamo, Italy.
Tel.: +39.035223753
E-mail: simone.cazzaniga@gised.it

Key words: teledermatology; skin cancer; melanoma; classification; agreement.

Contributions: SC, study design, data analysis, drafting of the work. LDP, GMB, SF, design, critical review. LN, CLV, study design, interpretation of data, drafting of the work, critical review. ADL, AC, MDM, LG, GT, MF, CC, MP, PD, EC, data collection, critical review. All authors approved the final version of the manuscript.

Conflict of interest: the authors declare no potential conflict of interest

Funding: the study was supported by a grant from the Italian League for the Fight Against Cancer (LILT) - CUP E62C20000150004.

Availability of data and material: data and materials are available by the authors.

Received for publication: 17 March 2022.
Accepted for publication: 25 April 2022.

Dermatology Reports 2023; 15:9500
doi:10.4081/dr.2022.9500

Publisher's note: all claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

Statistical analysis

For descriptive purposes, data was reported as means with standard deviations (SD) or absolute numbers with percentages for continuous and nominal variables respectively. Cohen's and Fleiss' kappa were used to calculate intra- and inter-rater agreement along with their 95% confidence intervals (CI). Kappa was interpreted as follows: <0 poor, 0.01-0.20 slight, 0.21-0.40 fair, 0.41-0.60 moderate, 0.61-0.80 substantial, 0.81-1.00 almost perfect agreement. The analyses were performed with SPSS software v.26 (IBM Corp, Armonk, NY).

Results

Demographics, phenotypic features and clinical characteristics of subjects and lesions considered are reported in Table 1. Most subjects were females (66.7%) with an average age of 39.8±14.0 years (mean ± SD). The most common phenotypic type was brown hair (68.9%) and eyes (55.6%). Lesions were mainly located on the legs (28.9%), anterior trunk (26.7%) or back (22.2%), with a diameter between 6-15 mm in 51.1% of cases and with a large portion of subjects (44.4%) reporting recent changes in the lesion. Table 2 shows the distribution of dermatologists' assessment in the first and second phase of the study. In the first phase 37.0% of lesions were judged as not assessable, 27.8% as non-suspected, 27.0% as suspected and 8.1% as highly suspected. The distribution of judgments on the same pictures in the second phase, after 1 month, was similar. More specifically, the overall intra-rater agreement, as assessed by Cohen's kappa, was 0.42 (95% CI: 0.33-0.50), varying from 0.12 to 0.59 on single raters (Table 3). When combing suspected and highly suspected lesions together, the overall kappa was 0.47 (95% CI: 0.39-0.56), ranging from 0.16 to 0.67. On the other side, the inter-rater agreement, as assessed by Fleiss' kappa, during the first phase was 0.29 (95% CI: 0.24-0.34) considering all the possible categories and 0.33 (95% CI: 0.28-0.39) combing suspected and highly suspected lesions together (Table 4). When considering the agreement for each category of judgment, kappa varied from 0.19 for not assessable to 0.48 for highly suspected lesions. Similar results were obtained in the second phase with kappa of 0.24 (95% CI: 0.19-0.29) and 0.30 (95% CI: 0.24-0.35) for all categories and for suspected and highly suspected lesions combined respectively.

Discussion

This study shows a less than satisfactory agreement among dermatologists, with a limited clinical experience, when judging about pigmentary lesions even after a training course has been performed. In addition, the study indicates that the consistency of the judgment, i.e., intra-rater agreement, varies among dermatologists with some dermatologists being more consistent than others. There are few studies assessing the agreement of dermatologists not supported by dermoscopy when judging about pigmentary lesions.[7,8] Even if dermoscopy is recognized as a pre-requisite for a clinical diagnosis, the search for suspicious lesions is usually directed by a preliminary inspection of the skin.[9]

Notably, the kappa values obtained in our study are similar to those obtained in the few similar studies published also enrolling experienced dermatologists. For example, the rates of inter- and intra-observer agreement amongst dermatologists were moderate in a concordance study where evaluation was limited to facial lesions.[8] Even when assessing dermoscopic features the level of agreement among different observers is rather low,[10,11] and adding dermoscopy to the clinical evaluation translate into a limited increase in a correct diagnosis (according to the study of Carli et al. the improvement was not higher than 15%).[12] In a meta-analysis, dermoscopy translated into an improved diagnosis of melanoma only in the hands of experienced clinicians and especially when the diagnosis was made by a group of examiners in consensus.[5]

Conclusions

All in all, these data are of practical relevance, and point to the need for improving the reliability of the clinical diagnoses of melanoma especially when assessing small lesions and when dealing with thin melanomas at a population level and not in the context of pigment lesions clinics. To improve diagnostic reliability, assessment by an interconnected group of experts, so-called collective intelligence, has been proposed.[13] More feasible, is assessment in duplicate by two different observers with discordance being solved by consensus or third-party adjudication. Finally, given the promising diagnostic performance of machine learning algorithms, such as deep convolutional neural networks,[14] automatic computer-based procedures are worth being assessed for melanoma early diagnosis in a *real world* setting.

Table 1. Demographics, phenotypic features and clinical characteristics of subjects and lesions selected in the study.

	N=45	%
Gender		
Male	15	33.3%
Female	30	66.7%
Age (years)		
Mean, SD	39.8	14.0
Hair colour		
Black	5	11.1
Brown	31	68.9
Red	1	2.2
Blond	6	13.3
Other	2	4.4
Eye colour		
Black	2	4.4
Brown	25	55.6
Green	6	13.3
Light blue	11	24.4
Other	1	2.2
Lesion site		
Head/face/neck	4	8.9
Shoulders/armpits	4	8.9
Arms	2	4.4
Anterior trunk	12	26.7
Back	10	22.2
Legs	13	28.9
Lesion diameter		
<6 mm	19	42.2
6-15 mm	23	51.1
>15 mm	1	2.2
Unknown	2	4.4
Recent onset		
No	34	75.6
Yes	6	13.3
Unknown	5	11.1
Recent changes		
No	15	33.3
Yes	20	44.4
Unknown	10	22.2
Personal history of melanoma		
No	33	73.3
Yes	6	13.3
Unknown	6	13.3
Family history of melanoma		
No	30	66.7
Yes	8	17.8
Unknown	7	15.6
Sunburns in lifetime		
No	28	62.2
Yes	12	26.7
Unknown	5	11.1
Ongoing immunosuppressive therapies		
No	43	95.6
Yes	1	2.2
Unknown	1	2.2

SD, standard deviation.

Table 2. Distribution of dermatologists' assessment of lesions in the first and second phase of the study.

Study phase	Judgment	Assessor												Total	
		1		2		3		4		5		6			
		N	%	N	%	N	%	N	%	N	%	N	%	N	%
I	Not assessable	22	48.9	20	44.4	16	35.6	20	44.4	8	17.8	14	31.1	100	37.0
	Non suspected	8	17.8	13	28.9	16	35.6	10	22.2	18	40.0	10	22.2	75	27.8
	Suspected	11	24.4	8	17.8	10	22.2	10	22.2	14	31.1	20	44.4	73	27.0
	Highly suspected	4	8.9	4	8.9	3	6.7	5	11.1	5	11.1	1	2.2	22	8.1
II (4 weeks after)	Not assessable	28	62.2	16	35.6	18	40.0	10	22.2	14	31.1	14	31.1	100	37.0
	Non suspected	3	6.7	14	31.1	13	28.9	17	37.8	17	37.8	10	22.2	74	27.4
	Suspected	12	26.7	10	22.2	10	22.2	11	24.4	12	26.7	17	37.8	72	26.7
	Highly suspected	2	4.4	5	11.1	4	8.9	7	15.6	2	4.4	4	8.9	24	8.9

Table 3. Intra-rater agreement between first and second phase of the study.

Assessor	Phase I judgment	Phase II judgment (4 weeks after)								Kappa (95% CI)*
		Not assessable		Non suspected		Suspected		Highly suspected		
		N	%	N	%	N	%	N	%	
1	Not assessable	19	67.9	0	0.0	3	25.0	0	0.0	0.49 (0.28, 0.71)
	Non suspected	5	17.9	3	100.0	0	0.0	0	0.0	0.54 (0.32, 0.76)**
	Suspected	4	14.3	0	0.0	7	58.3	0	0.0	
	Highly suspected	0	0.0	0	0.0	2	16.7	2	100.0	
2	Not assessable	12	75.0	4	28.6	3	30.0	1	20.0	0.53 (0.33, 0.72)
	Non suspected	2	12.5	10	71.4	1	10.0	0	0.0	0.56 (0.36, 0.76)**
	Suspected	2	12.5	0	0.0	5	50.0	1	20.0	
	Highly suspected	0	0.0	0	0.0	1	10.0	3	60.0	
3	Not assessable	12	66.7	2	15.4	2	20.0	0	0.0	0.59 (0.40, 0.78)
	Non suspected	5	27.8	11	84.6	0	0.0	0	0.0	0.67 (0.48, 0.85)**
	Suspected	1	5.6	0	0.0	7	70.0	2	50.0	
	Highly suspected	0	0.0	0	0.0	1	10.0	2	50.0	
4	Not assessable	7	70.0	10	58.8	3	27.3	0	0.0	0.37 (0.18, 0.57)
	Non suspected	2	20.0	7	41.2	1	9.1	0	0.0	0.45 (0.25, 0.64)**
	Suspected	1	10.0	0	0.0	6	54.5	3	42.9	
	Highly suspected	0	0.0	0	0.0	1	9.1	4	57.1	
5	Not assessable	1	7.1	5	29.4	2	16.7	0	0.0	0.12 (-0.07, 0.30)
	Non suspected	5	35.7	10	58.8	3	25.0	0	0.0	0.16 (-0.04, 0.36)**
	Suspected	7	50.0	1	5.9	5	41.7	1	50.0	
	Highly suspected	1	7.1	1	5.9	2	16.7	1	50.0	
6	Not assessable	7	50.0	3	30.0	4	23.5	0	0.0	0.35 (0.14, 0.56)
	Non suspected	4	28.6	5	50.0	1	5.9	0	0.0	0.41 (0.19, 0.62)**
	Suspected	3	21.4	2	20.0	12	70.6	3	75.0	
	Highly suspected	0	0.0	0	0.0	0	0.0	1	25.0	
Total	Not assessable	58	58.0	24	32.4	17	23.6	1	4.2	0.42 (0.33, 0.50)
	Non suspected	23	23.0	46	62.2	6	8.3	0	0.0	0.47 (0.39, 0.56)**
	Suspected	18	18.0	3	4.1	42	58.3	10	41.7	
	Highly suspected	1	1.0	1	1.4	7	9.7	13	54.2	

CI, confidence interval. *Cohen's kappa. **Kappa calculated combining suspected and highly suspected lesions together.

Table 4. Inter-rater agreement in the first and second phase of the study.

Judgment	Phase I Kappa (95% CI)*	Phase II (4 weeks after) Kappa (95% CI)*
Not assessable	0.19 (0.11-0.26)	0.19 (0.12-0.27)
Non suspected	0.34 (0.26-0.41)	0.23 (0.15-0.30)
Suspected	0.29 (0.21-0.37)	0.25 (0.17-0.32)
Highly suspected	0.48 (0.41-0.56)	0.40 (0.32-0.47)
Total	0.29 (0.24-0.34)	0.24 (0.19-0.29)
	0.33 (0.28-0.39)**	0.30 (0.24-0.35)**

CI, confidence interval. *Fleiss' kappa calculated on total and on specific categories. **Kappa calculated combining suspected and highly suspected lesions together..

References

1. Farberg AS, Rigel DS. The Importance of Early Recognition of Skin Cancer. Dermatol Clin 2017;35:xv-xvi.
2. Coroiu A, Moran C, Bergeron C, et al. Operationalization of skin self-examination in randomized controlled trials with individuals at increased risk for melanoma: A systematic review. Patient Educ Couns 2020;103:1013-26.
3. Ana FA, Loreto MS, José LM, et al. Mobile applications in oncology: A systematic review of health science databases. Int J Med Inform 2020;133:104001.
4. Chuchu N, Takwoingi Y, Dinnes J, et al. Smartphone applications for triaging adults with skin lesions that are suspicious for melanoma. Cochrane Database Syst Rev 2018;12:CD013192.
5. Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. Lancet Oncol 2002;3:159-65.
6. Cazzaniga S, Castelli E, Di Landro A, et al. Mobile teledermatology for melanoma detection: Assessment of the validity in the framework of a population-based skin cancer awareness campaign in northern Italy. J Am Acad Dermatol 2019;81:257-60.
7. Roush GC, Barnhill RL, Ernstoff MS, Kirkwood JM. Inter-clinician agreement on the recognition of clinical pigmentary characteristics of patients with cutaneous malignant melanoma. Studies of melanocytic nevi, VI. Br J Cancer 1991;64:373-6.
8. Yu P, Li X, Huang Y, et al. Inter- and intra-observer agreement in dermatologists' diagnoses of hyperpigmented facial lesions and development of an algorithm for automated diagnosis. Skin Res Technol 2019;25:777-86.
9. Ontario Health (Quality). Pigmented Lesion Assay for Suspected Melanoma Lesions: A Health Technology Assessment. Ont Health Technol Assess Ser 2021;21:1-81.
10. Pizzichetta MA, Talamini R, Piccolo D, et al. Interobserver agreement of the dermoscopic diagnosis of 129 small melanocytic skin lesions. Tumori 2002;88:234-8.
11. Carrera C, Marchetti MA, Dusza SW, et al. Validity and reliability of dermoscopic criteria used to differentiate nevi from melanoma: a web-based International Dermoscopy Society study. JAMA Dermatol 2016;152:798-806.
12. Carli P, De Giorgi V, Naldi L, Dosi G. Reliability and inter-observer agreement of dermoscopic diagnosis of melanoma and melanocytic naevi. Dermoscopy Panel. Eur J Cancer Prev 1998;7:397-402.
13. Winkler JK, Sies K, Fink C, et al. Collective human intelligence outperforms artificial intelligence in a skin lesion classification task. J Dtsch Dermatol Ges 2021;19:1178-84.
14. Popescu D, El-Khatib M, El-Khatib H, Ichim L. New trends in melanoma detection using neural networks: a systematic review. Sensors (Basel) 2022;22:496.