TITLE

The utility of non-rule based visual matching as a strategy to allow novices to achieve skin lesion diagnosis

Authors

R Benjamin Aldridge [1a]

Dominik Glodzik [2]

Lucia Ballerini [2a]

Robert B Fisher [2]

Jonathan L Rees [1]

[1] Department of Dermatology, University of Edinburgh, Lauriston Building, Lauriston Place, Edinburgh, EH3 9HA

[2] School of Informatics, University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB

Funding from The Wellcome Trust (Reference 083928/Z/07/Z) and the Foundation for Skin Research (Edinburgh) . RBA[a] and LB[a] supported by Wellcome Trust.

Corresponding author

Professor Jonathan Rees

Department of Dermatology, University of Edinburgh, Level 1 Lauriston Building, Lauriston Place, Edinburgh, EH3 9HA

jonathan.rees@ed.ac.uk, Tel: 44-(0) 131 536 2041, Fax: 44-(0) 131 229 8769

ABSTRACT

Non-analytical reasoning is thought to play a key role in dermatology diagnosis. Considering its potential importance, surprisingly little work has been done to research whether similar identification processes can be supported in non-experts. We describe a prototype diagnostic support software, which we have used to examine the ability of medical students (at the beginning and end of a dermatology attachment) and lay volunteers, to diagnose 12 images of common skin lesions. Overall the non-experts using the software had a diagnostic accuracy of 98% (923/936) compared to the control group of 33% (215/648) (Wilcoxon $p<0.0001$). We have demonstrated, within the constraints of a simplified clinical model, that novices' diagnostic scores are significantly increased by the use of a structured image database coupled with matching of index and referent images. The novices achieve this high degree of accuracy without any use of explicit definitions of likeness or rule-based strategies.

Keywords

Non-analytical reasoning, Skin cancer, Electronic support software, Melanoma, Dermatology diagnosis

Principal Author

Professor Jonathan Rees

Department of Dermatology, University of Edinburgh, Level 1 Lauriston Building, Lauriston Place, Edinburgh, EH3 9HA

jonathan.rees@ed.ac.uk, Tel: 44-(0) 131 536 2041, Fax: 44-(0) 131 229 8769

INTRODUCTION

Understanding the cognitive skills involved in making a dermatological diagnosis may be important both for improving the education of doctors—whether experts or generalists— and also for enabling patients to detect early signs of skin disease. In this regard, as in the rest of this paper, we are thinking particularly of skin cancer and lesions that might be confused with skin cancer.

Despite the importance of the topic to dermatological practice there is only a handful of papers concerned with the psychological processes involved in dermatological diagnosis, notably those coming from the laboratory of Geoff Norman and colleagues (1-7). At the risk of some simplification, the processes involved in diagnosis can be viewed as being either explicit and based on conscious reasoning or, alternatively, as being implicit and holistic, and hidden from the conscious view of the diagnostician (8). This distinction in certain respects corresponds to the division between Type 1 and Type 2 decision making highlighted by Kahneman (for review see Evans (8)). To take a concrete example, in diagnosing a nodular basal cell carcinoma a clinician might argue that he applies a set of rules such as the presence of a pearly edge, telangiectasia and so on, or alternatively might 'at a glance' recognise features holistically that, from previous experience and learning, are characteristic of a basal cell carcinoma. Whilst in reality it seems likely that different processes might be used in different clinical situations, there is good evidence that much clinical reasoning and other forms of expertise is indeed holistic and that the clinician may not be privy to how he or she achieves the correct diagnosis (9-11). In the particular context of some medical expertise this form of reasoning has been labelled as "non-analytical reasoning" by Norman (2, 4, 5).

One issue raised by such insights is whether it is possible to build tools that might enhance non-analytical strategies such that, rather than apply explicit rules (e.g. the ABCD rules for melanoma diagnosis (12)), novices or learners might be able to match index cases with a database of images to achieve a diagnosis (or at least narrow the range of diagnostic uncertainty). In our experience many clinicians are very sceptical that such an approach might be useful. There is however some tentative evidence that such a matching strategy may work, although only to the extent that it has been demonstrated to be better than chance (13).

A scaleable vehicle in which to examine the utility of matching is by use of World Wide Web (WWW) browser based interfaces written in HTML/JAVA code. The WWW allows large numbers of images to be distributed at low cost and lends itself to the addition of computational engines that might, at a later date, allow a range of clinical variables to be added to enhance the possibility of success. Therefore in the present study we set out to examine experimentally whether non-experts can use a simple bespoke JAVA test interface to match index cases presented as a digital image with a range of images including those from the correct diagnostic class. In order to provide a reference level of competence we compared the results of such an approach with the diagnostic accuracy of a control groups of medical students before and after a dermatological attachment.

MATERIALS & METHODS

Software Image Selection

80 images from five diagnostic classes of commonly referred focal skin lesions were selected from the University of Edinburgh Dermatology Department's image library. The images comprised 14 haemangiomas, 23 seborrheic keratoses, 19 melanocytic naevi, 15 basal cell carcinomas and 9 squamous cell carcinomas. Images were chosen on the basis of technical quality of the images and because they were thought to be representative of a particular diagnostic class. These five diagnostic groups make up the majority of the lesions that are referred from primary care for specialist assessment. All the images had been collected using the same controlled fixed distance photographic setup; Canon (Canon, Japan) EOS 350D 8.1MP cameras, Sigma (SIGMA, Japan) 70mm f2.8 macro lens and Sigma EM-140 DG Ring Flash at a distance of 50cms. From these 80 images, 12 index lesions were randomly selected, with the remaining 68 images acting as referent images in the software image database.

Software Design

Our prototype software allows the user to make a direct visual comparison between a centralised index image and up to 12 surrounding referent images (see Figure I). The user then navigates through the library of referent images until they are happy that they have successful matched the index lesion to a similar referent image (or images). In this experiment the 68 referent images were arranged over 3 levels utilising a total of 18 different screens (1 screen for level 1, 5 screens for level 2, 12 screens for level 3). Irrespective of which index image was being tested, the referent images in the first level's screen were

identical for all matching attempts. It was only the subsequent second and third level screens' referent images that were determined by the individual user's image selection. The order in which these 5 second level and 12 third level screens were displayed and their relationship to a specific users image selection was predetermined by the experimenters and was kept constant for the duration of the experiment. The method employed for grouping the 68 images to the 18 screens and the relationship of a screen to a specific user interaction was based on the experimenters' opinion of visual similarity and to a lesser degree the lesions' underlying pathological diagnosis. If the user was unhappy with their selection at any stage of the process (prior to confirming their final match) the software allowed them to retrace their steps. As the screen-snap shots attest the software is very intuitive, nonetheless to demonstrate how to navigate through the software library and how to make a final diagnostic match we integrated into the software a short instructional video. This video, to avoid any potential bias, did not include images of skin lesions but demonstrated the key features of the software using simple pictures of differing shapes (circles, squares, crosses). A video demonstrating the version of the software tested is available on YouTube (Google, California)(14) .

Experiment 1

Similar to many UK medical schools the University of Edinburgh's undergraduate dermatology teaching program consists of an introductory series of 8 lectures followed by a two week clinical attachment incorporating 9 demonstration clinics (15, 16). All students that attended for their two-week clinical attachment over a 3-month period (November 2009- January 2010) were recruited into the study. In total 60 students were enrolled (4 batches of between 14-16 students). Other than 8 introductory lectures (one of which was

dedicated to skin cancer) none of the students had prior clinical experience of dermatology. Thirty-six (60%) of the students were female.

On the morning of Day 1 of the dermatology attachment (prior to seeing any patients), each batch of students was randomly split into two groups; the first group (in total 31 students) was asked to identify each of the 12 index images using the software and the second group (in total 29 students) was asked to identify the 12 test images by writing their diagnosis on an answer sheet. Forthwith the students using our prototype software to obtain a diagnostic match will be addressed as the 'software' group, and the students that provided written answers as the 'control' group. Test instructions were standardised across the batches of students. We were "generous" in what we accepted as correct answers for the control group, allowing spelling mistakes, incomplete terminology, abbreviations and lay terms. After the Day 1 test no score or feedback was provided to either group. Exactly the same experiment was repeated on the afternoon of Day 10 at the end of the students' dermatology attachment. The format of both the Day 1 and Day 10 experiments was identical, except that the introductory software video was not repeated to the software group on Day 10.

The 12 test images were presented to both groups of subjects in the same order and in an identical format. For both the software and control groups the skin lesion images were displayed using the same Apple iMac G5 20" monitors (Apple, California) with identical resolutions (1650x1050), calibrated for colour inconsistencies using the Pantone Huey Pro calibration (Pantone LLC, New Jersey). The experiments were all undertaken in a designated curtained room with similar ambient lighting conditions. No time restrictions were imposed for either group. Constructive feedback was only given after each batch of students had completed the Day 10 test during an additional tutorial.

Experiment 2

20 lay members of the public were recruited between May and July 2010. Mean age was 33 (Range 21-61). 75% were female. All but four had completed university education and the 20 subjects were employed in a wide range of different occupations (eg solicitor, accountant, teacher, secretary, chef). No volunteer had any personal experience of skin cancer nor had undergone any tuition in the identification of skin lesions.

The 20 subjects were provided with the same introductory video guide to the software as the students but no additional training. The experimental setup was identical to that undertaken by the students, with the same 12 test images and an identical version of the software (as described above). This group of subjects will subsequently be referred to as the 'lay' group.

Statistical analysis of all results was undertaken using R for Mac OS, V2.9.0 (17).

Ethics

NHS Lothian research ethics committee granted permission for the collection and use of the images, and additional permission for the use of medical students in this research was granted through the University's "Committee for the use of medical student volunteers".

RESULTS

Experiment 1

93% (112/120) of students completed both the Day 1 and Day 10 test. Student absence was evenly distributed across the four test groups; Day 1 control group (n=1), Day 10 control group (n=3), Day 1 software group (n=1), Day 10 software group (n=3).

At the start of their dermatology attachment (Day 1 test), out of the twelve test images, the control group correctly diagnosed a median of 1 image with a diagnostic accuracy of 16% (55/336), in the same Day 1 test the software group correctly identified a median of 12 images, resulting in a diagnostic accuracy of 99% (357/360). At the end of the students' dermatology attachment (Day 10 test) the control group correctly diagnosed a median of 6 images with a diagnostic accuracy of 51% (160/312) and the software group matched 12 images correctly, with a diagnostic accuracy of 99% (335/336). Results are shown graphically in Figure II.

Two sample Wilcoxon tests showed that the scores at Day 1 between the software and control group were significantly different ($p<0.0001$) as were the two groups scores at the end of the students' attachment on Day 10 ($p<0.0001$). Wilcoxon paired test showed that the control group's scores improved significantly ($p<0.0001$) over their attachment, whereas the software groups score did not appear to change ($p=0.582$).

There was no difference in test scores between the four batches of students or between the sexes. In addition we saw no particular pattern of results with respect to lesion type.

Experiment 2

The lay group, using the software, correctly identified a median of 12 images resulting in a diagnostic accuracy of 96% (231/240). See Figure II. Again there was no difference in test scores between the sexes or with respect to lesion type.

Two sample Wilcoxon tests showed that the student control group had significantly inferior diagnostic accuracy when compared to the lay group, at both the start and end of their dermatology attachment ($p<0.0001$).

DISCUSSION

Our results show clearly—within the constraints of a limited range of diagnostic possibilities and an image based approach—that medical students are able to utilise visual matching as a diagnostic strategy and achieve diagnostic scores that are higher than medical students who have completed a standard clinical dermatological attachment. This ability is not confined to medical students, as we subsequently went on to examine a group of non medical trained individuals who scored similarly. This success was therefore achieved by test subjects making matches on the basis of visual similarity without any attempt to apply any explicit rules of likeness based on dermatological knowledge. We believe these results are in keeping with the idea that promotion of non-analytical based reasoning strategies may be useful educationally useful for non-experts (18). There are however a number of limitations to our work and points worthy of further elaboration.

Immediately after completing their undergraduate dermatology teaching attachment, students' unaided diagnostic accuracy for common skin lesions was only 51%. Although it is mildly reassuring that the students improved their diagnostic acumen over the course of their two-week attachment, a final diagnostic accuracy of 51% is perhaps poor, although obviously any absolute score is dependent on the difficulty of the test set. This result is more sobering when one considers that the level of these students' diagnostic accuracy may reflect an artificially raised result; the students' achieved this level of accuracy after double-exposure to the 12 test images (the students had previously viewed, albeit without feedback, the 12 images during the first test on Day 1 of the attachment). In addition, as with the majority of UK undergraduate dermatological attachments, our students were in their penultimate clinical year, so it is probable that by the time they graduate a further drop off in

their diagnostic performance could be expected. However disappointing the students scores may seem, they are, in fact, not dissimilar to previous studies that have investigated the diagnostic accuracy of non-dermatologists after medical school training with colour images (19, 20).

Our results are also constrained by other features of our design. For obvious practical reasons, our testing relied on matching to an image not to a lesion on a real patient. There is still some uncertainty about the limitations of virtual versus real patients in this context, although we note that images are widely used in teaching and examination of clinical competence, and that if we think of our approach as a teaching tool for clinicians then virtual patients may be thought to at least supplement patient exposure. If such a matching tool is envisioned as a diagnostic support tool for the lay public (for instance in encouraging early presentation of suspicious pigmented lesions) then this limitation needs further exploration.

In the present studies we did not attempt to represent the whole of the complexity of dermatological morphology, rather focussing on a range of common lesions. Any performance figures clearly must, in a fundamental way, relate to the difficulty or atypicality of the test set. However, we would argue that our approach was that of proof-of-concept which, given the results, suggests further work is merited. The approach we have used based on only 80 images is, however, eminently scaleable and we are currently building software that will allow examination of several hundred images. Our view is that as the database increases in size it may become increasingly powerful assuming that we can order it in a way that is intuitive to the user. This can either be based on ordering of images based on

automatically extracted properties ('computer vision'), or user feedback, or some combination of the two (21-25).

That novices were able to identify skin lesions without any explicit definition of likeness or specific rule based analysis (such as the ABCD) makes our approach fundamentally different to most previous strategies to improve non-expert diagnosis. Whilst it is tempting to want to explore exactly what features of images users are actually matching to, this may neither be necessary or tractable. Ironically despite its appeal, in many situations there is clear evidence that exclusive rule based strategies may actually diminish diagnostic accuracy or decrease the utility of decision making (10, 18, 26-28).

Finally, whatever the insights our work provides into the relative different diagnostic strategies, we can envision two applications of our approach. The first would be as a teaching and learning tool for clinicians. Whilst we have not demonstrated any learning took place in our experiments, merely that we provided subjects with a software tool that enabled them to achieve something they would not have been able to achieve without the software, it is not difficult to imagine how such a system might be embedded with teaching material for clinicians. The second application is for the lay public, and the approach we describe might be considered an extension of the posters and leaflets that are used to educate the public about the warnings signs of skin cancer. Although many will be anxious about whether such approaches are safe, we note that 80% of internet users have already undertaken health related searches (29) and that there is some evidence that current strategies may actually worsen rather than improve diagnostic performance (18). It is surely better to examine experimentally how such approaches might improve matters rather than make unwarranted assumptions about how humans are able to categorize skin lesions.

ACKNOWLEDGMENTS

The work was supported by The Wellcome Trust (Reference 083928/Z/07/Z) and the Foundation for Skin Research (Edinburgh). We are also grateful to the advice and assistance given by Karen Roberston and Yvonne Bisset (Department of Dermatology, University of Edinburgh) regarding the photographic capture and preparation of the digital images.

REFERENCES

1. Norman G, Eva K, Brooks L, Hamstra S. Expertise in Medicine and Surgery. In: Ericsson KA, Charness N, Feltovich PJ, Hoffman RR, editors. The Cambridge Handbook of Expertise and Expert Performance. New York: Cambridge University Press; 2006. p. 339-354.
2. Norman GR, Rosenthal D, Brooks LR, Allen SW, Muzzin LJ. The development of expertise in dermatology. Arch Dermatol. 1989;125:1063-1068.
3. Brooks LR, Norman GR, Allen SW. Role of specific similarity in a medical diagnostic task. Journal of Experimental Psychology: General. 1991;120:278-287.
4. Norman G, Brooks LR. The Non-Analytical Basis of Clinical Reasoning. Adv Health Sci Educ Theory Pract. 1997;2:173-184.
5. Norman G, Young M, Brooks L. Non-analytical models of clinical reasoning: the role of experience. Medical education. 2007;41:1140-1145.
6. Jackson R. The importance of being visually literate. Observations on the art and science of making a morphological diagnosis in dermatology. Arch Dermatol. 1975;111:632-636.
7. Gachon J, Beaulieu P, Sei JF et al. First prospective study of the recognition process of melanoma in dermatological practice. Arch Dermatol. 2005;141:434-438.
8. Evans JS. Dual-processing accounts of reasoning, judgment, and social cognition. Annu Rev Psychol. 2008;59:255-278.
9. McLaughlin K, Rikers RM, Schmidt HG. Is analytic information processing a feature of expertise in medicine? Adv Health Sci Educ Theory Pract. 2008;13:123-128.
10. Kulatunga-Moruzi C, Brooks LR, Norman G. Using comprehensive feature lists to bias medical diagnosis. J Exp Psychol Learn Mem Cogn. 2004;30:563-572.

11. Norman G. Building on experience--the development of clinical reasoning. N Engl J Med. 2006;355:2251-2252.
12. Friedman RJ, Rigel DS, Kopf AW. Early detection of malignant melanoma: the role of physician examination and self-examination of the skin. CA: A Cancer Journal for Clinicians. 1985;35:130-151.
13. Brown N, Robertson K, Bisset Y, Rees J. Using a structured image database, how well can novices assign skin lesion images to the correct diagnostic grouping? J Invest Dermatol. 2009;129:2509-2512.
14. YouTube (Full URL to follow)
15. Burge S. Teaching dermatology to medical students: a survey of current practice in the U.K. Br J Dermatol. 2002;146:295-303.
16. Davies E, Burge S. Audit of dermatological content of U.K. undergraduate curricula. Br J Dermatol. 2009;160:999-1005.
17. R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
18. Girardi S, Gaudy C, Gouvernet J, Teston J, Richard M, Grob J. Superiority of a cognitive education with photographs over ABCD criteria in the education of the general population to the early detection of melanoma: a randomized study. Int J Cancer. 2006;118:2276-2280.
19. Hansra NK, O'Sullivan P, Chen CL, Berger TG. Medical school dermatology curriculum: are we adequately preparing primary care physicians? Journal of the American Academy of Dermatology. 2009;61:23-29.e1.
20. Federman DG, Kirsner RS. The abilities of primary care physicians in dermatology: implications for quality of care. Am J Manag Care. 1997;3:1487-1492.

21. Ballerini L, Li X, Fisher R, Rees J. A Query-by-Example Content-Based Image Retrieval System of Non-Melanoma Skin Lesions. In: Caputo B, Müller H, Syeda-Mahmood T, Duncan J, Wang F, Kalpathy-Cramer J, editors. Lecture Notes in Computer Science vol 5853. Berlin / Heidelberg: Springer; 2010. p. 31-38.

22. Ballerini L, Li X, Fisher R, Aldridge B, Rees J. Content-Based Image Retrieval of Skin Lesions by Evolutionary Feature Synthesis. In: Di Chio C, Cagnoni S, Cotta C, Ebner M, Ekárt A, Esparcia-Alcazar A, Goh C-K, Merelo J, Neri F, Preuß M, Togelius J, Yannakakis G, editors. Lecture Notes in Computer Science vol 6024. Berlin / Heidelberg: Springer; 2010. p. 312-319.

23. Li X, Aldridge B, Rees J, Fisher R. Estimating the ground truth from multiple individual segmentations with application to skin lesion segmentation. Proc Medical Image Understanding and Analysis. 2010;101-106.

24. Li X, Aldridge B, Ballerini L, Fisher R, Rees J. Depth Data Improves Skin Lesion Segmentation. In: Yang G-Z, Hawkes D, Rueckert D, Noble A, Taylor C, editors. Lecture Notes in Computer Science vol 5762. Berlin / Heidelberg: Springer; 2009. p. 1100-1107.

25. Laskaris N, Ballerini L, Fisher RB, Aldridge B, Rees J. Fuzzy description of skin lesions. In: Manning DJ, Abbey CK, editors. Proceedings of SPIE vol 7627. Bellingham: SPIE; 2010. p. 762717-1 to 762717-10.

26. Norman G, Eva K. Diagnostic error and clinical reasoning. Medical education. 2010;44:94-100.

27. Norman G. Dual processing and diagnostic errors. Adv in Health Sci Educ. 2009;14 Suppl 1:37-49.

28. Allen SW, Brooks LR, Norman GR, Rosenthal D. Effect of prior examples on rule-based diagnostic performance. Res Med Educ. 1988;27:9-14.

29. White R, Horvitz E. Cyberchondria: studies of the escalation of medical concerns in web search. ACM Transactions on Information Systems (TOIS). 2009;27:1-37.

FIGURES

Figure I Legend

Figure I: Screen 'snap-shots' taken from the software showing how a correct diagnostic match could be made for index/test image 11 (a seborrheic keratosis). The boxes highlight the user's selections at each of the 3 levels. A full range of screen 'snap-shots' are available on (http://www.reestheskin.eu/). In addition a video of the software is available on YouTube (14).

Figure I

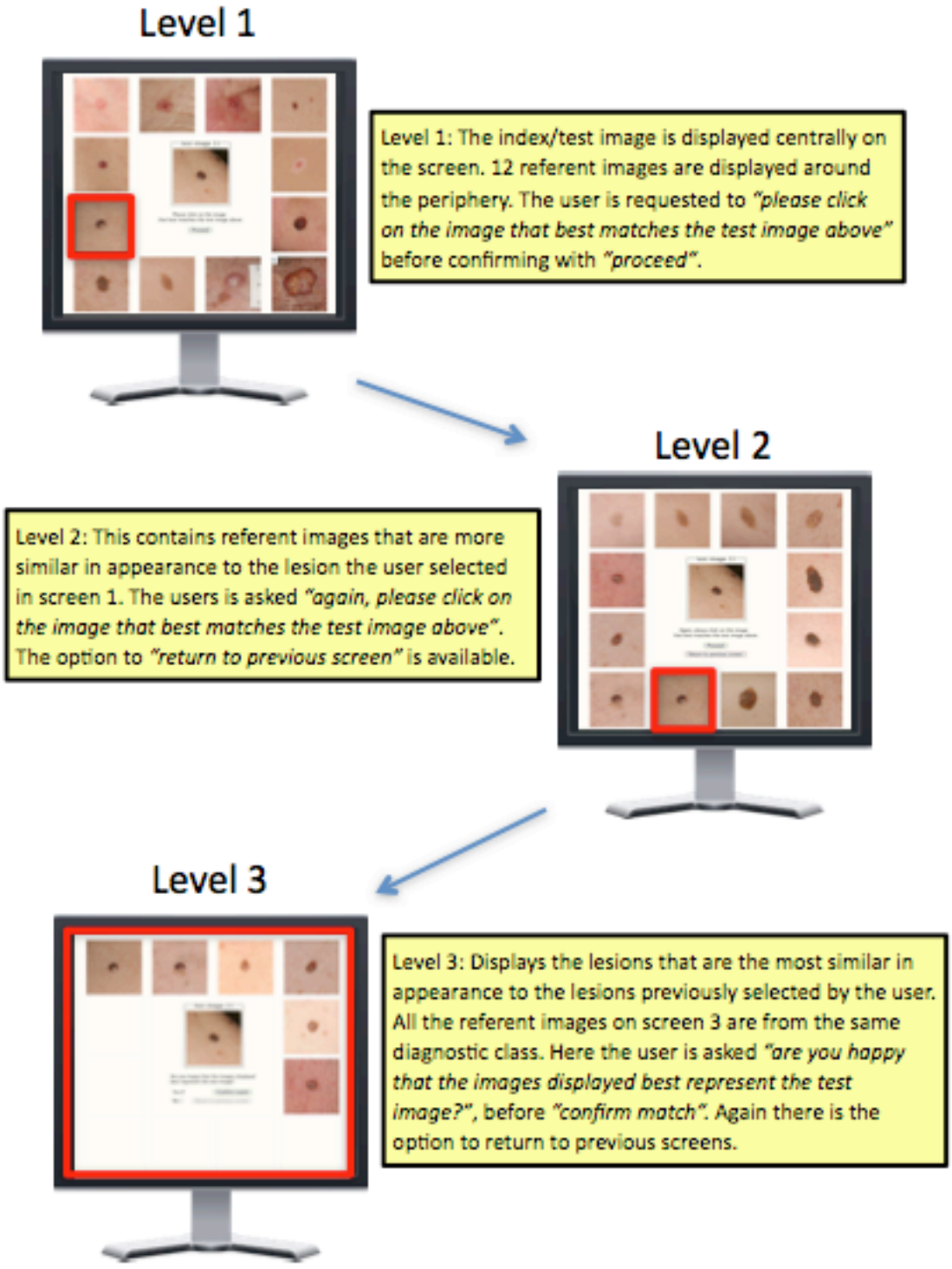
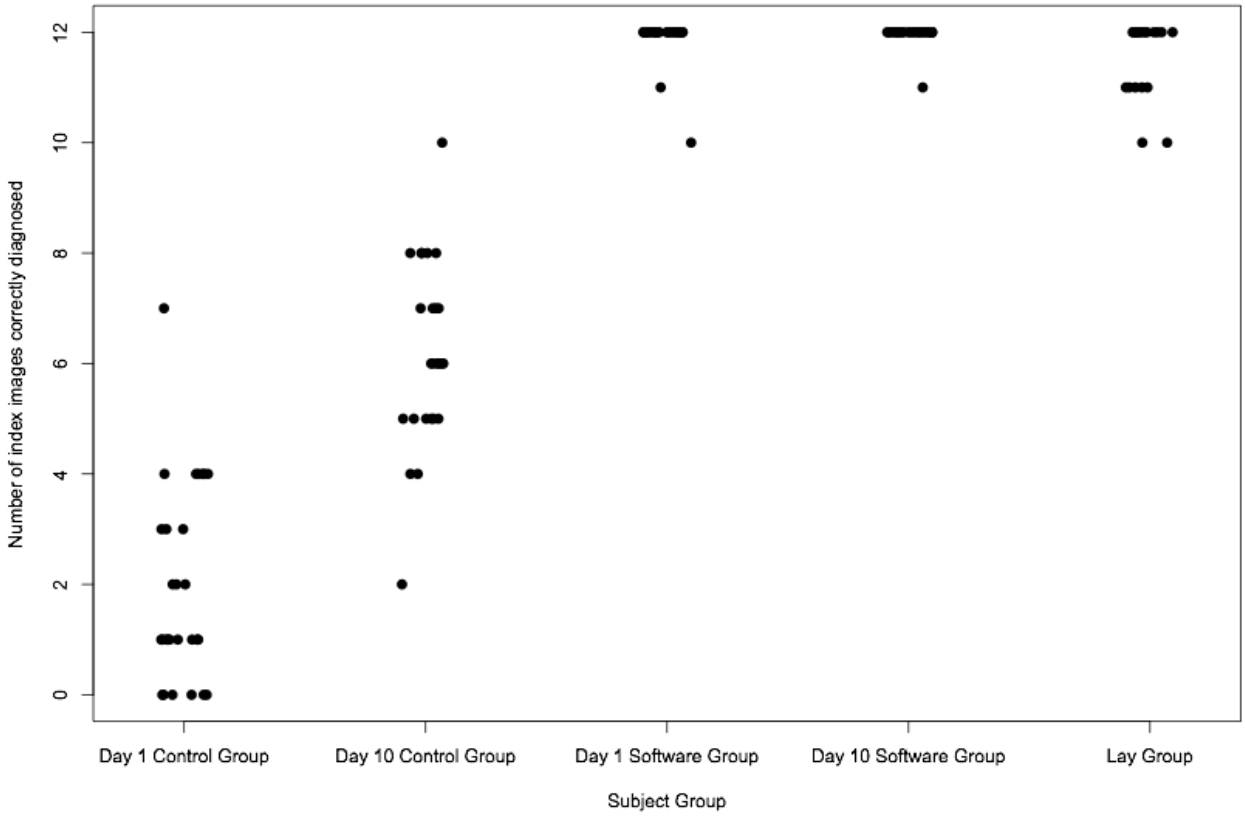

Level 1
Level 1: The index/test image is displayed centrally on the screen. 12 referent images are displayed around the periphery. The user is requested to *"please click on the image that best matches the test image above"* before confirming with *"proceed"*.
Level 2
Level 2: This contains referent images that are more similar in appearance to the lesion the user selected in screen 1. The users is asked *"again, please click on the image that best matches the test image above"*. The option to *"return to previous screen"* is available.
Level 3
Level 3: Displays the lesions that are the most similar in appearance to the lesions previously selected by the user. All the referent images on screen 3 are from the same diagnostic class. Here the user is asked *"are you happy that the images displayed best represent the test image?"*, before *"confirm match"*. Again there is the option to return to previous screens.

Figure II Legend

Figure II: Plot of all 60 students' scores by group and test date, and the 20 lay novices' scores. The maximum score of 12 is achieved by identifying all the test images correctly. Day 1 control group (n=28, median score 1), Day 10 control group (n=26, median score 6), Day 1 software group (n=30, median score 12), Day 10 software group (n=28, median score 12). Lay group score (n=20, median score 12).

Figure II


Number of index images correctly diagnosed
0
2
4
6
8
10
12
Day 1 Control Group
Day 10 Control Group
Day 1 Software Group
Day 10 Software Group
Lay Group
Subject Group