Diagnostic capacity of skin tumor artificial intelligence-assisted decision-making software in real-world clinical settings

Cheng-Xu Li[1,2], Wen-Min Fei[1,2], Chang-Bing Shen[1,2,3], Zi-Yi Wang[1,2], Yan Jing[4], Ru-Song Meng[5], Yong Cui[1,2]

[1]Department of Dermatology, China-Japan Friendship Hospital, Beijing 100029, China;
[2]Graduate School, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing 100730, China;
[3]Hinda and Arthur Marcus Institute for Aging Research, Hebrew SeniorLife and Harvard Medical School, Boston, MA, USA;
[4]Department of Dermatology, The First Affiliated Hospital, Anhui Medical University, Hefei, Anhui 230032, China;
[5]Department of Dermatology, Specialty Medical Center of the Air Force, Chinese People's Liberation Army, Beijing 100142, China.

Abstract

Background: Youzhi artificial intelligence (AI) software is the AI-assisted decision-making system for diagnosing skin tumors. The high diagnostic accuracy of Youzhi AI software was previously validated in specific datasets. The objective of this study was to compare the performance of diagnostic capacity between Youzhi AI software and dermatologists in real-world clinical settings.

Methods: A total of 106 patients who underwent skin tumor resection in the Dermatology Department of China-Japan Friendship Hospital from July 2017 to June 2019 and were confirmed as skin tumors by pathological biopsy were selected. Dermoscopy and clinical images of 106 patients were diagnosed by Youzhi AI software and dermatologists at different dermoscopy diagnostic levels. The primary outcome was to compare the diagnostic accuracy of the Youzhi AI software with that of dermatologists and that measured in the laboratory using specific data sets. The secondary results included the sensitivity, specificity, positive predictive value, negative predictive value, F-measure, and Matthews correlation coefficient of Youzhi AI software in the real-world.

Results: The diagnostic accuracy of Youzhi AI software in real-world clinical settings was lower than that of the laboratory data ($P < 0.001$). The output result of Youzhi AI software has good stability after several tests. Youzhi AI software diagnosed benign and malignant diseases by recognizing dermoscopic images and diagnosed disease types with higher diagnostic accuracy than by recognizing clinical images ($P = 0.008$, $P = 0.016$, respectively). Compared with dermatologists, Youzhi AI software was more accurate in the diagnosis of skin tumor types through the recognition of dermoscopic images ($P = 0.01$). By evaluating the diagnostic performance of dermatologists under different modes, the diagnostic accuracy of dermatologists in diagnosing disease types by matching dermoscopic and clinical images was significantly higher than that by identifying dermoscopic and clinical images in random sequence ($P = 0.022$). The diagnostic accuracy of dermatologists in the diagnosis of benign and malignant diseases by recognizing dermoscopic images was significantly higher than that by recognizing clinical images ($P = 0.010$).

Conclusion: The diagnostic accuracy of Youzhi AI software for skin tumors in real-world clinical settings was not as high as that of using special data sets in the laboratory. However, there was no significant difference between the diagnostic capacity of Youzhi AI software and the average diagnostic capacity of dermatologists. It can provide assistant diagnostic decisions for dermatologists in the current state.

Keywords: Artificial intelligence; Skin tumor; Diagnostic accuracy

Introduction

In recent years, artificial intelligence (AI) has received unprecedented attention and has been researched and applied in many medical disciplines.[1-5] Dermatology is an intuitive morphological science that is especially suitable for AI-assisted diagnosis. AI holds great promise for the clinical application in the screening and the diagnosis of skin cancer. Recent studies have successfully demonstrated that dermatology AI based on deep learning algorithms can classify skin tumors at a dermatologist level and even its performance surpasses board-certified dermatologists.[6-8] Meanwhile, dermatology AI can provide high-quality medical services and alleviate the shortage and uneven distribution of medical resources.

Skin tumor, especially malignant skin tumor, is a prominent global public health problem. The incidence of the disease is increasing yearly, seriously affecting

Access this article online

Quick Response Code:

Website:
www.cmj.org

DOI:
10.1097/CM9.0000000000001002

Correspondence to: Prof. Yong Cui, Department of Dermatology, China-Japan Friendship Hospital, No. 2 Yinghua East Street, Beijing 100029, China
E-Mail: wuhucuiyong@vip.163.com

Chinese Medical Journal 2020;133(17)

Received: 26-04-2020 **Edited by:** Li-Shao Guo

human health.[9] Dermatologists usually make their diagnosis by the naked eye or are assisted by skin imaging methods such as dermoscopy, reflectance confocal microscopy, and very high-frequency skin ultrasound. Compared with the naked-eye examination, dermoscopy increases the sensitivity and specificity for the detection of skin tumors.[10] However, a recent study has shown that Chinese dermatologists have relatively low imaging diagnostic ability for skin tumors, and that the diagnostic ability of dermatologists in different regions is uneven.[11] Therefore, it is urgent to use dermatology AI to assist with the proper diagnosis.

Youzhi AI software (Shanghai Maise Information Technology Co., Ltd., Shanghai, China) is the AI decision-making system for skin tumors in China. Its data are based on the Chinese Skin Image Database (CSID), which is currently one of the largest skin image databases in China. Youzhi AI software was jointly developed by the CSID project team and Shanghai Maise Information Technology and was trained from a dataset including more than 200,000 dermoscopic images.[12] The training model utilized the GoogLeNet Inception v4 convolutional neural network architecture[13] as the basis. The segmentation branch was added as the output based on the classification branch. The training images were labeled and classified by professional dermatologists. The diagnostic accuracy of benign and malignant skin tumors reached 91.2%, and the diagnostic accuracy of disease types reached 81.4%, attaining the international level in recognition accuracy.[14,15]

The high accuracy of Youzhi AI software was previously validated using specific datasets in the research laboratory. However, studies have shown that the performance of AI decision support systems in the real world or using different datasets will be lower than the experimental settings.[16-19] The performance of Youzhi AI software in skin tumor diagnosis decision-making has not been evaluated using unfiltered clinical data in a real-world randomized comparative trial. The purpose of this study was to investigate the performance of Youzhi AI software in the clinical setting and to compare it with the performance of dermatologists. Since multiple dermatologists use the same images to test Youzhi AI software, we were also able to assess the repeatability of the system outputs.

Methods

Ethical approval

This study was approved by the Ethics Committee of China-Japan Friendship Hospital, and informed consent was signed by all patients.

Study population

A total of 2023 patients admitted to the Department of Dermatology of the China-Japan Friendship Hospital from September 2017 to June 2019 were initially included in our retrospective study. All lesions were surgically excised and pathologically examined because of equivocal dermoscopic findings, and the lesion was considered malignant tumor, or at the patient's request. Dermoscopy was performed pre-operatively. Exclusion criteria were as follows: (a) controversial cases with ambiguous histopathological reports; (b) the histopathological diagnosis results do not belong to skin tumor; (c) digital clinical images and/or dermoscopic images of skin tumors did not meet the diagnostic requirements, for example, a part of the image data was missing, the image was not clear enough to be diagnosed, or skin lesions covered by exogenous pigment cannot show its true colors; (d) other treatments such as photodynamic therapy before skin tumor resection/biopsy; (e) skin tumors beyond the recognition capability of Youzhi AI software; (f) low-quality images for other reasons. At last, 1438 patients met the inclusion requirements, and 106 patients were randomly selected from these 1438 patients by simple random sampling method in the study.

Image acquisition

All clinical images and dermoscopic images were taken by dermatologists from the China-Japan Friendship Hospital under standard illumination using the FotoFinder medicam® 1000 (FotoFinder Systems GmbH, Birnbach, Germany). This desktop dermoscope features continuous optical real-time zoom and autofocus, perfectly suitable for dermoscopy of skin, hair, and nails. Therefore, the quality of the image could be guaranteed.

Study design

The clinical images and dermoscopic images of the 106 patients were analyzed using the Youzhi AI software (system version 2.2.5). We validated that the first-level classification nodes of the Youzhi AI software algorithm were benign lesions and malignant lesions. Second, we validated that the second-level classification nodes of the Youzhi AI software algorithm were 14 types of skin tumors. So, the results of benign and malignant judgments and the disease types were given by the software. Investigators must be trained before the experimental operation for the reason that images need to be appropriately cropped during software analysis to ensure that the lesion was located in the center of the recognition area. The experiment was repeated five times by different investigators to minimize errors in the software results and verify the stability of the software diagnostic accuracy.

Meanwhile, 11 dermatologists were invited to participate in the experiment. The 11 dermatologists consisted of four physicians who passed the primary dermoscopy proficiency level test, four physicians who passed the intermediate dermoscopy proficiency level test, and three experts in the field of dermoscopy.

Initially, 212 clinical images and dermoscopic images of 106 patients were processed randomly, sequenced (random sequence mode, DR), and then diagnosed by 11 dermatologists. One week later, the clinical photos of 106 patients corresponded to the dermoscopic images one by one (match mode, DM) and were diagnosed by the same 11 dermatologists, giving benign and malignant judgments and disease types. All diagnostic results were then compared with their respective histopathological results, which were regarded as the reference standard.

Statistical analysis

The normality of the distribution of continuous variables was tested by the D'Agostino-Pearson test. Continuous variables with normal distribution were presented as mean ± standard deviation; non-normal variables were reported as median (interquartile range). Means of two continuous normally distributed variables with homogeneity of variance were compared by independent samples *t*-test (Student's *t* test). Mann-Whitney *U* test was used to compare the means of two groups of variables not normally distributed. The one-sample *t*-test was used to compare the diagnostic accuracy of AI software in the laboratory and this experiment. All given *P* values were two-tailed, and the criterion for significance was set at $P < 0.05$. Statistical analysis was performed with SPSS (version 20.0, IBM, Armonk, NY, USA) or GraphPad Prism (version 7.0, GraphPad software).

Results

Pathological types of selected skin tumors

All the skin tumor diagnoses were based on the gold standard of histopathological examination and supported by history. Images with the following diagnosis were included in this study: among the malignant skin tumors and precancerous lesions, there were 4 cases of malignant melanoma, five cases of squamous cell carcinoma, 24 cases of basal cell carcinoma, and three cases of actinic keratosis. Among the benign skin tumors, there were 19 cases of nevus cell nevus, including ten cases of intradermal nevus, four cases of junctional nevus, and five cases of the compound nevus. Besides, there were 35 cases of seborrheic keratosis, four cases of hemangioma, six cases of dermatofibroma, and six cases of the epidermoid cyst [Table 1].

Measured results and abbreviations interpretations

We calculated the diagnostic accuracy of the Youzhi AI software and each dermatologist based on the histopathological results of the biopsy, including the diagnostic accuracy of benign and malignant (BMA) and the diagnostic accuracy of disease type (DTA). Therefore, the following results were obtained. (a) The BMA and DTA when dermatologists identified 212 randomly sequenced images, abbreviated as DR-BMA and DR-DTA. Among them, the diagnostic accuracy of dermatologists in recognizing dermoscopic images was abbreviated as DRD-BMA and DRD-DTA, and the diagnostic accuracy in recognizing clinical images was abbreviated as DRC-BMA and DRC-DTA, respectively. (b) The BMA and DTA when dermatologists identified 106 sets of matched dermoscopic and clinical images, abbreviated as DM-BMA and DM-DTA. (c) The BMA and DTA by Youzhi AI software abbreviated as AI-BMA and AI-DTA. The diagnostic accuracy of dermoscopic images by Youzhi AI software was abbreviated as AID-BMA and AID-DTA, the diagnostic accuracy of clinical images was abbreviated as AIC-BMA and AIC-DTA, respectively. The relationship between the above-measured results was shown in Figure 1. (d) The two diagnostic accuracies that we mentioned in the introduction part, the BMA and DTA measured in the laboratory through a specific dataset, were abbreviated as Lab-BMA and Lab-DTA. The values of the above-mentioned results were shown in Figure 2.

Table 1: Number and proportion of hispathological types of 106 selected lesions.

Histopathology	*n* (%)
Malignant melanoma	4 (4)
Squamous cell carcinoma	5 (5)
Basal cell carcinoma	24 (23)
Actinic keratosis	3 (3)
Intradermal nevus	10 (9)
Junctional nevus	4 (4)
Compound nevus	5 (5)
Seborrheic keratosis	35 (33)
Hemangioma	4 (4)
Dermatofibroma	6 (5)
Epidermoid cyst	6 (5)
Total	106 (100)

Comparison of diagnostic accuracy between Youzhi AI software and dermatologists

No matter in the BMA, or the DTA, there was no statistical difference in the diagnostic accuracy of Youzhi AI software and dermatologist under the two modes (AI *vs.* DR; AI *vs.* DM). The AID-BMA was not more accurate than the DRD-BMA, with no statistically significant difference ($P = 0.761$). However, in terms of disease types, the AID-DTA was higher than DRD-DTA (0.7642 [95% CI 0.7132–0.8000] *vs.* 0.6338 [95% CI 0.5513–0.7163]), and the difference was statistically significant ($P = 0.010$). Compared with dermatologists, there was no statistically significant difference in the diagnostic accuracy of clinical images in Youzhi AI software (AIC-BMA *vs.* DRC-BMA, $P = 0.476$; AIC-DTA *vs.* DRC-DTA, $P = 0.682$) [Table 2].

Evaluation of the diagnostic performance of dermatologists under different modes

To find out if different diagnostic modes (random sequence mode and match mode) had an impact on the diagnostic accuracy of dermatologists, we found that the DM-DTA was significantly higher than DR-DTA (0.7358 [95% CI 0.6821–0.7896 *vs.* 0.6141 [95% CI 0.5195–0.7086, $P = 0.022$) and DRD-DTA (0.6338 [95% CI 0.5513–0.7163], $P = 0.032$). There was no significant difference in the comparison between DR-BMA and DM-BMA ($P = 0.296$), DM-BMA and DRD-BMA ($P = 0.319$). Regardless of whether it was DM-BMA or DM-DTA, the diagnostic accuracy of dermatologists by matching mode was higher than that by clinical images only and had statistical significance (DM-BMA *vs.* DRC-BMA, $P = 0.023$, and DM-DTA *vs.* DRC-DTA, $P = 0.021$). Compared with DRC-BMA, DRD-BMA had higher

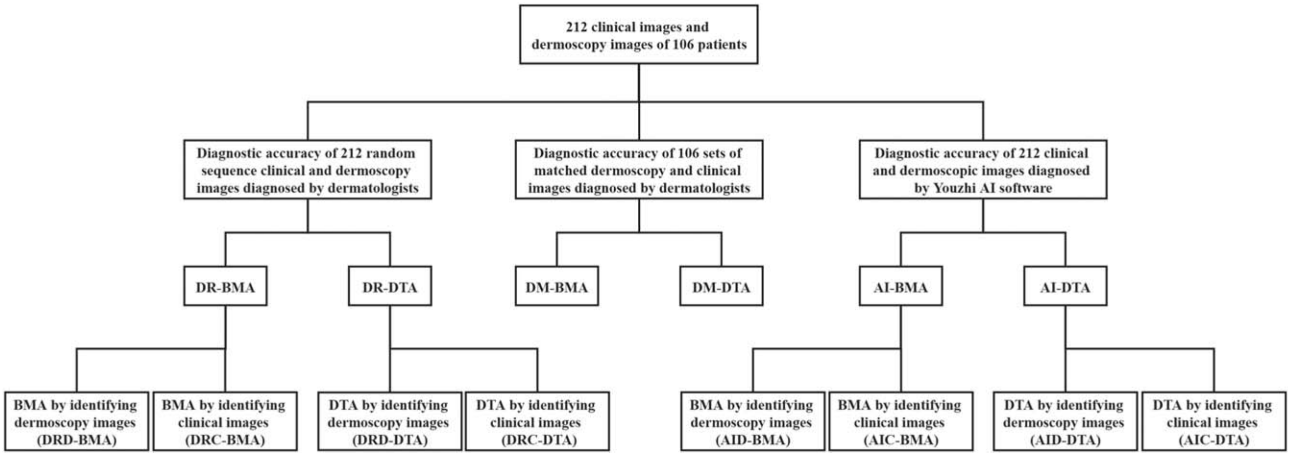
212 clinical images and dermoscopy images of 106 patients
Diagnostic accuracy of 212 random sequence clinical and dermoscopy images diagnosed by dermatologists
Diagnostic accuracy of 106 sets of matched dermoscopy and clinical images diagnosed by dermatologists
Diagnostic accuracy of 212 clinical and dermoscopic images diagnosed by Youzhi AI software
DR-BMA
DR-DTA
DM-BMA
DM-DTA
AI-BMA
AI-DTA
BMA by identifying dermoscopy images (DRD-BMA)
BMA by identifying clinical images (DRC-BMA)
DTA by identifying dermoscopy images (DRD-DTA)
DTA by identifying clinical images (DRC-DTA)
BMA by identifying dermoscopy images (AID-BMA)
BMA by identifying clinical images (AIC-BMA)
DTA by identifying dermoscopy images (AID-DTA)
DTA by identifying clinical images (AIC-DTA)

Figure 1: Relationship between all test results in the study. AI: Diagnostic accuracy of Youzhi AI software in recognizing dermatoscopic and clinical images; AIC: Diagnostic accuracy of clinical images recognition by Youzhi AI software; AID: Diagnostic accuracy of dermoscopic images recognition by Youzhi AI software; BMA: In term of diagnostic accuracy of benign and malignant tumors; DM: Diagnostic accuracy of dermatologists in recognizing 106 sets of matched dermoscopic and clinical images; DR: Diagnostic accuracy of dermatologists recognizing 212 randomly sequenced dermoscopic and clinical images; DRC: Diagnostic accuracy of dermatologists in recognizing clinical images; DRD: Diagnostic accuracy of dermatologists in recognizing dermoscopic images; DTA: In term of the diagnostic accuracy of disease types.

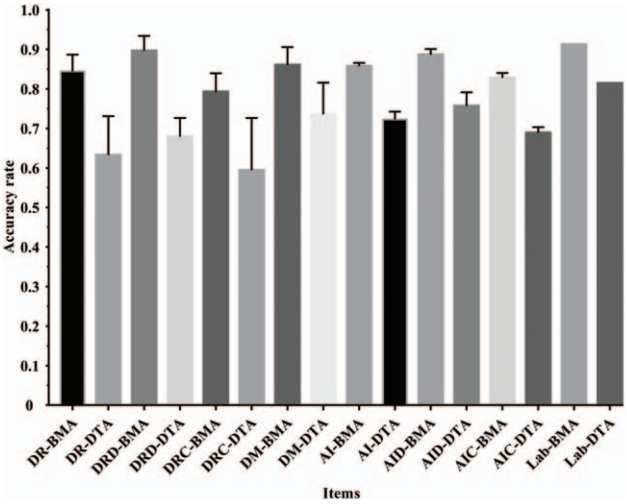

Figure 2: Accuracy for all measured results. AI: Diagnostic accuracy of Youzhi AI software in recognizing dermatoscopic and clinical images; AIC: Diagnostic accuracy of clinical images recognition by Youzhi AI software; AID: Diagnostic accuracy of dermoscopic images recognition by Youzhi AI software; BMA: In term of diagnostic accuracy of benign and malignant tumors; DM: Diagnostic accuracy of dermatologists in recognizing 106 sets of matched dermoscopic and clinical images; DR: Diagnostic accuracy of dermatologists recognizing 212 randomly sequenced dermoscopic and clinical images; DRC: Diagnostic accuracy of dermatologists in recognizing clinical images; DRD: Diagnostic accuracy of dermatologists in recognizing dermoscopic images; DTA: In term of the diagnostic accuracy of disease types; Lab: Diagnostic accuracy measured in the laboratory through a specific dataset.

diagnostic accuracy and statistical significance (0.7950 [95% CI 0.7445–0.8456] *vs.* 0.8962 [95% CI 0.8195–0.9233], respectively, $P = 0.01$), while there was no significant difference between DRC-DTA and DRD-DTA ($P = 0.527$) [Table 2].

Performance evaluation of Youzhi AI software

In this study, the diagnostic accuracy of dermoscopic and clinical images of Youzhi AI software was lower than that of the laboratory test, and the difference was statistically significant ($P < 0.001$). In terms of BMA and DTA, the diagnostic accuracy of Youzhi AI software in recognizing dermoscopic images was higher than that of the clinical images (AIC-BMA *vs.* AID-BMA, $P = 0.008$, and AIC-DTA *vs.* AID-DTA, $P = 0.016$) [Table 2]. The diagnostic accuracy of Youzhi AI software with five repeated measurements was analyzed by one-way analysis of variance (ANOVA) to test the stability of its output results. The difference between the results of the five measurements was not statistically significant.

Kruskal-Wallis test one-way ANOVA was used to compare the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F-measure, and Matthews correlation coefficient (MCC) of the AI-BMA, AIC-BMA, and AID-BMA, and only AIC-BMA and AID-BMA showed statistically significant differences, as shown in Table 3.

Discussion

Compared with naked-eye observation, dermoscopy can provide more details on skin lesions, increase the sensitivity and specificity of detecting skin tumors, help clinicians to distinguish benign and malignant lesions, and reduce unnecessary pathological biopsies.[10] However, dermoscopy also has its limitations. For example, the examination and diagnosis of dermoscopy require professional training and continuous practice to improve the ability. At the same time, some skin tumors lack specific dermoscopic features, and the exact diagnosis cannot be obtained by dermoscopic morphological observation alone, which needs to be supplemented with information such as medical history and clinical pictures. It was also reflected in our experimental results that when dermatologists observed both dermoscopic images and clinical photographs, the diagnostic accuracy was higher, especially in the diagnosis of specific disease types. It indicates that clinical images and dermoscopic images are complementary to each other in the diagnosis of skin tumor diseases

Table 2: Comparisons of diagnostic accuracy differences between different types of results.

Comparison	BMA						DTA					
(Former *vs*. Latter)	Former	95% CI	Latter	95% CI	Statistics	*P*	Former	95% CI	Latter	95% CI	Statistics	*P*
Youzhi AI Software and dermatologists												
AI *vs*. DR	0.8585 (0.8491–0.8656)	0.8471 to 0.8680	0.8332 ± 0.0699	0.7863 to 0.8802	24.5*	0.760†	0.7311 (0.7052–0.7359)	0.6977 to 0.7476	0.6141 ± 0.1407	0.5195 to 0.7086	15.0*	0.180†
AI *vs*. DM	0.8585 (0.8491–0.8656)	0.8471 to 0.8680	0.8602 ± 0.0456	0.8296 to 0.8909	22.0*	0.563†	0.7311 (0.7052–0.7359)	0.6977 to 0.7476	0.7358 ± 0.0801	0.6821 to 0.7896	27.5*	0.955†
AID *vs*. DRD	0.8868 (0.8727–0.9009)	0.8683 to 0.9053	0.8962 (0.8491–0.9340)	0.8195 to 0.9233	24.5*	0.761†	0.7642 (0.7217–0.7878)	0.7132 to 0.800	0.6338 ± 0.1228	0.5513 to 0.7163	5.5*	0.010†
AIC *vs*. DRC	0.8302 (0.8161–0.8396)	0.8130 to 0.8436	0.7950 ± 0.0753	0.7445 to 0.8456	21.0*	0.476†	0.6887 (0.6745–0.7028)	0.6701 to 0.7072	0.5943 ± 0.1619	0.4856 to 0.7031	23.5*	0.682†
Dermatologists under different modes												
DR *vs*. DM	0.8332 ± 0.0699	0.7863 to 0.8802	0.8602 ± 0.0456	0.8296 to 0.8909	1.1	0.296‡	0.6141 ± 0.1407	0.5195 to 0.7086	0.7358 ± 0.0801	0.6821 to 0.7896	2.5	0.022‡
DRD *vs*. DM	0.8962 (0.8491–0.934)	0.8195 to 0.9233	0.8602 ± 0.0456	0.8296 to 0.8909	45.0*	0.319†	0.6338 ± 0.1228	0.5513 to 0.7163	0.7358 ± 0.0801	0.6821 to 0.7896	2.3	0.032‡
DRC *vs*. DM	0.7950 ± 0.0753	0.7445 to 0.8456	0.8602 ± 0.0456	0.8296 to 0.8909	2.5	0.023‡	0.5943 ± 0.1619	0.4856 to 0.7031	0.7358 ± 0.0801	0.6821 to 0.7896	2.6	0.021‡
DRC *vs*. DRD	0.7950 ± 0.0753	0.7445 to 0.8456	0.8962 (0.8491–0.934)	0.8195 to 0.9233	22.5*	0.010†	0.5943 ± 0.1619	0.4856 to 0.7031	0.6338 ± 0.1228	0.5513 to 0.7163	0.6	0.527‡
Youzhi AI software in different conditions												
AI *vs*. Lab	0.8585 (0.8491–0.8656)	0.8471 to 0.8680	0.912		228.1	<0.001§	0.7311 (0.7052–0.7359)	0.6977 to 0.7476	0.814		80.3	<0.001§
AID *vs*. Lab	0.8868 (0.8727–0.9009)	0.8683 to 0.9053	0.912		133.1	<0.001§	0.7642 (0.7217–0.7878)	0.7132 to 0.800	0.814		48.4	<0.001§
AIC *vs*. Lab	0.8302 (0.8161–0.8396)	0.8130 to 0.8436	0.912		150.7	<0.001§	0.6887 (0.6745–0.7028)	0.6701 to 0.7072	0.814		103.3	<0.001§
AIC *vs*. AID	0.8302 (0.8161–0.8396)	0.8130 to 0.8436	0.8868 (0.8727–0.9009)	0.8683 to 0.9053	0*	0.008†	0.6887 (0.6745–0.7028)	0.6701 to 0.7072	0.7642 (0.7217–0.7878)	0.7132 to 0.8000	0.5*	0.016†

Values were shown as mean ± standard deviation, or median (interquartile range). **U* value, otherwise *t* value; †Mann-Whitney *U* test; ‡ Independent samples Student's *t* test; § One-sample *t*-test. CI: Confidence interval; AI: Diagnostic accuracy of Youzhi AI software in recognizing dermatoscopic and clinical images; AIC: Diagnostic accuracy of clinical images recognition by Youzhi AI software; AID: Diagnostic accuracy of dermoscopic images recognition by Youzhi AI software; BMA: In term of diagnostic accuracy of benign and malignant tumors; DM: Diagnostic accuracy of dermatologists in recognizing 106 sets of matched dermoscopic and clinical images; DR: Diagnostic accuracy of dermatologists recognizing 212 randomly sequenced dermoscopic and clinical images; DRC: Diagnostic accuracy of dermatologists in recognizing clinical images; DRD: Diagnostic accuracy of dermatologists in recognizing dermoscopic images; DTA: In term of the diagnostic accuracy of disease types; Lab: Diagnostic accuracy measured in the laboratory through a specific dataset.

Table 3: Indicators of the diagnostic capacity of Youzhi AI software in the diagnosis of benign and malignant skin tumors.

Characteristics	AI-BMA	AIC-BMA	AID-BMA	P^*
Sensitivity	0.7484 ± 0.0149 (0.7300–0.7669)	0.7110 ± 0.0169 (0.6901–0.7319)	0.7864 ± 0.0273 (0.7525–0.8203)	0.002
Specificity	0.9296 ± 0.0052 (0.9231–0.9361)	0.9060 ± 0.0107 (0.8928–0.9192)	0.9532 ± 0.0107 (0.9399–0.9665)	0.001
PPV	0.8750 ± 0.0098 (0.8628–0.8872)	0.8333 ± 0.0196 (0.8090–0.8577)	0.9167 ± 0.0196 (0.8923–0.9410)	0.001
NPV	0.8486 ± 0.0117 (0.8340–0.8631)	0.8257 ± 0.0120 (0.8109–0.8406)	0.8714 ± 0.0202 (0.8463–0.8965)	0.005
F-measure	0.8067 ± 0.0101 (0.7942–0.7942)	0.7673 ± 0.0164 (0.7469–0.7876)	0.8463 ± 0.0187 (0.8230–0.8695)	0.001
MCC	0.7004 ± 0.0162 (0.6804–0.7205)	0.6377 ± 0.0262 (0.6052–0.6702)	0.7634 ± 0.0296 (0.7267–0.8002)	0.001

Data are expressed as mean ± standard deviation (95% confidence interval). *Comparison of indicators of AIC and AID. AI: Diagnostic accuracy of Youzhi AI software in recognizing dermatoscopic and clinical images; AIC: Diagnostic accuracy of clinical images recognition by Youzhi AI software; AID: Diagnostic accuracy of dermoscopic images recognition by Youzhi AI software; BMA: In term of diagnostic accuracy of benign and malignant tumors; PPV: Positive predictive value; NPV: Negative predictive value; MCC: Matthews correlation coefficient.

and play a very important role in improving the accuracy of diagnosis.

A study has shown that it is necessary to establish a specific dataset for skin diseases of different regions and races, to build a high-performance and highly stable computer-aided diagnosis system for skin disease.[20] The Youzhi AI software is developed based on the CSID data, which is a skin image database of Chinese. The training model mainly applies GoogLeNet Inception v4, which has a more uniform simplified architecture and more inception modules,[13] and has better performance and accuracy than Inception v3. After several measurements, the diagnostic accuracy of Youzhi AI software output is relatively stable.

However, as with many related research results, the diagnostic accuracy of the Youzhi AI software was reduced in practical work. There are several reasons for this. The first reason is that the difficulty of dermoscopic images used in these two software performance tests is different. Dermoscopic images used in laboratory tests are more typical and have more dermoscopic features required for software recognition, while dermoscopic images collected in clinical work tend to be less typical, increasing the diagnostic difficulty for dermatologists and software in this experiment. Another possible explanation for the decrease in performance is the fact that the test images come from different sources. The images of the two tests were collected by different dermoscopic devices in different medical institutions. Meanwhile, the acquisition period of dermoscopic images is relatively wide; the color calibration of the devices, the pressure, and tilt angle applied by the operator on the skin of the patients using dermoscopic devices, and so forth may lead to the generation of low contrast images and the loss of details.[17,21] In addition, laboratory testing is to identify dermoscopic image models. This experiment also tested the clinical image model, which is relatively immature, affecting the overall diagnostic performance of the software.

In 2019, we conducted a web survey of Chinese dermatologists' attitudes toward AI. The survey results showed that almost Chinese dermatologists considered the role of AI in "assisting dermatologists in their daily diagnosis and treatment activities." AI should be implemented in secondary hospitals and skin tumors in the future.[22] Our results show that even on the premise that the diagnostic accuracy of Youzhi AI software was reduced, there was no significant difference between its accuracy and the average level of dermatologists, and Youzhi AI software was even better than dermatologists in recognizing the type of skin tumor diagnosed by dermoscopic images. Moreover, the diagnostic accuracy of dermatologists with primary dermoscopic diagnostic ability was not as good as that of Youzhi AI software. It has been observed that young inexperienced dermatologists and family physicians have great difficulties in correctly assessing skin lesions, with significant decreases in sensitivity and specificity.[23] These findings guided the application of Youzhi AI software in dermatology. Therefore, Youzhi AI software is particularly suitable for providing assistant diagnostic decision-making for Chinese family physicians (general practitioners), primary dermatologists, and doctors with poor dermoscopic diagnostic capabilities.

Our study is the first to test the performance of Youzhi AI software using images taken under real clinical conditions. Nevertheless, there are several limitations to our study. First, this study had a retrospective design and involved only a single institution. The number of doctors participating in the test was not large enough. In the future, we will integrate the Youzhi AI software with a cloud-based multi-hospital collaboration platform to further test the diagnostic capacity of the software through a large sample and multicenter randomized controlled trial. The purpose is to further improve the software diagnostic ability, to do a good job in assisting the doctors in diagnosis, and to bring greater benefits to patients. Second, in the laboratory test, the total number of test data images accounts for 5% to 10% of the training (modeling) data, so the number of test images is nearly 3000. However, this is too much for dermatologists. Some dermatologists who participated in this study reflected that the number of images was slightly extensive, and there might be fatigue in the later stage of the experiment, which may affect the diagnostic accuracy. It will be more reasonable to design the number of test images and test methods in the future.

In conclusion, we verified the diagnostic accuracy of Youzhi AI software in the real clinical environment, although its performance is not as good as that tested in the laboratory using a specific dataset. However, as a computer-aided diagnostic system for skin tumor diseases, Youzhi AI software can effectively assist doctors to judge the benign and malignant skin tumors and specific types of diseases more accurately, particularly in medical institutions not experienced in dermoscopy. These results may have some reference value for further improvement of the performance of a computer-aided diagnostic system.

Funding

This study was supported by grants from the Fundamental Research Funds for the Central Universities (No. 3332019163), and the Beijing Municipal Science and Technology Commission Medicine Collaborative Science and Technology Innovation Research Project (No. Z191100007719001).

Conflicts of interest

None.

References

1. Hogarty DT, Mackey DA, Hewitt AW. Current state and future prospects of artificial intelligence in ophthalmology: a review. Clin Exp Ophthalmol 2019;47:128–139. doi: 10.1111/ceo.13381.
2. Deyer T, Doshi A. Application of artificial intelligence to radiology. Ann Transl Med 2019;7:230. doi: 10.21037/atm.2019.05.79.
3. Diaz O, Dalton J, Giraldo J. Artificial intelligence: a novel approach for drug discovery. Trends Pharmacol Sci 2019;40:550–551. doi: 10.1016/j.tips.2019.06.005.
4. Kothari S, Gionfrida L, Bharath AA, Abraham S. Artificial intelligence (AI) and rheumatology: a potential partnership. Rheumatology 2019;58:1894–1895. doi: 10.1093/rheumatology/kez194.
5. Hekler A, Utikal JS, Enk AH, Solass W, Schmitt M, Klode J, *et al.* Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. Eur J Cancer 2019;118:91–96. doi: 10.1016/j.ejca.2019.06.012.
6. Fujisawa Y, Otomo Y, Ogata Y, Nakamura Y, Fujita R, Ishitsuka Y, *et al.* Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumor diagnosis. Br J Dermatol 2018;180:373–381. doi: 10.1111/bjd.16924.
7. Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, *et al.* Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. Eur J Cancer 2019;113:47–54. doi: 10.1016/j.ejca.2019.04.001.
8. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, *et al.* Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–118. doi: 10.1038/nature21056.
9. Gordon R. Skin cancer: an overview of epidemiology and risk factors. Semin Oncol Nurs 2013;29:160–169. doi: 10.1016/j.soncn.2013.06.002.
10. Yélamos O, Braun RP, Liopyris K, Wolner ZJ, Kerl K, Gerami P, *et al.* Usefulness of dermoscopy to improve the clinical and histopathologic diagnosis of skin cancers. J Am Acad Dermatol 2019;80:365–377. doi: 10.1016/j.jaad.2018.07.072.
11. Shen C, Shen X, Li C, Meng R, Cui Y. Assessment of imaging diagnosis ability of skin tumors in Chinese dermatologists. Chin Med J 2019;132:2119–2120. doi: 10.1097/CM9.0000000000000389.
12. Li C, Shen C, Xue K, Shen X, Jing Y, Wang Z, *et al.* Artificial intelligence in dermatology: past, present, and future. Chin Med J 2019;132:2017–2020. doi: 10.1097/CM9.0000000000000372.
13. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the impact of residual connections on learning. arXiv 2016;1602.07261.
14. Li CX, Shen CB, Cui Y. Research and application of dermatological artificial intelligence based on multi-dimensional skin image database of the Chinese population (in Chinese). Robot Ind 2018;23:96–102. doi: CNKI:SUN:JQRY.0.2018-06-021.
15. Shen CB, Li CX, Shen X, Jing Y, Wang ZY, Xue K, *et al.* Development and application of dermatological artificial intelligence products based on big data of skin image (in Chinese). Chin Digit Med 2019;14:22–25. doi: CNKI:SUN:YISZ.0.2019-03-009.
16. Dreiseitl S, Binder M, Vinterbo S, Kittler H. Applying a decision support system in clinical practice: results from melanoma diagnosis. AMIA Annu Symp Proc 2007;191–195.
17. Dreiseitl S, Binder M, Hable K, Kittler H. Computer versus human diagnosis of melanoma: evaluation of the feasibility of an automated diagnostic system in a prospective clinical trial. Melanoma Res 2009;19:180–184. doi: 10.1097/CMR.0b013e32832a1e41.
18. Zhao X, Wu X, Li F, Li Y, Huang W, Huang K, *et al.* The application of deep learning in the risk grading of skin tumors for patients using clinical images. J Med Syst 2019;43:283. doi: 10.1007/s10916-019-1414-2.
19. Lin H, Li R, Liu Z, Chen J, Yang Y, Chen H, *et al.* Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial. EClinicalMedicine 2019;9:52–59. doi: 10.1016/j.eclinm.2019.03.001.
20. Xie B, He X, Zhao S, Li Y, Su J, Zhao X, *et al.* XiangyaDerm: a clinical image dataset of asian race for skin disease aided diagnosis. Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention 2019;22–31. doi: 10.1007/978-3-030-33642-4_3.
21. Grana C, Pellacani G, Seidenari S. Practical color calibration for dermoscopy, applied to a digital epiluminescence microscope. Skin Res Technol 2005;11:242–247. doi: 10.1111/j.0909-725X.2005.00127.x.
22. Shen CB, Li CX, Xu F, Wang ZY, Shen X, Gao J, *et al.* Web-based study on Chinese dermatologists' attitudes towards artificial intelligence. Ann Transl Med 2020;8:698. doi: 10.21037/atm.2019.12.102.
23. Jaworek-Korjakowska J, Kleczek P. Automatic classification of specific melanocytic lesions using artificial intelligence. Biomed Res Int 2016;2016:8934242. doi: 10.1155/2016/8934242.

How to cite this article: Li CX, Fei WM, Shen CB, Wang ZY, Jing Y, Meng RS, Cui Y. Diagnostic capacity of skin tumor artificial intelligence-assisted decision-making software in real-world clinical settings. Chin Med J 2020;133:2020–2026. doi: 10.1097/CM9.0000000000001002