Dermatology Practical & Conceptual

Artificial Intelligence versus Human Assessment in the Treatment of Upper Facial Wrinkles with Abobotulinum Toxin A

Kadir Küçük[1], Dilara İlhan Erdil[1], Fatmanur Hacinecipoğlu[1], Gökçen Çelik[1], Selda Pelin Kartal[1]

1 Dermatology and Venerology, Etlik City Hospital, Ankara, Tukey

Key words: Artificial intelligence, Dermatology, Facial wrinkling, Abobotulinum toxin A, Aesthetic medicine

Citation: Küçük K, İlhan Erdil D, Hacinecipoğlu F, Çelik G, Kartal SP. Artificial Intelligence versus Human Assessment in the Treatment of Upper Facial Wrinkles with Abobotulinum Toxin A. *Dermatol Pract Concept*. 2026;16(1):5978. DOI: https://doi.org/10.5826/dpc.1601a5978

Accepted: July 26, 2025; **Published:** January 2026

Funding: None.

Competing Interests: None.

Authorship: All authors have contributed significantly to this publication.

Corresponding Author: Kadir Küçük, MD, TC Sağlık Bakanlığı Ankara Etlik Şehir Hastanesi Dermatoloji Kliniği, Yenimahalle, Ankara, Türkiye. ORCİD: 0000-0001-6633-2292. E-mail: drkadirkucuk@hotmail.com

ABSTRACT

Introduction: Botulinum toxin is widely used to treat upper facial wrinkles, and its efficacy is typically assessed through photographic comparisons and standardized scales. Artificial intelligence (AI) is increasingly being integrated into aesthetic dermatology for objective wrinkle evaluation.

Objectives: This study aimed to compare human and AI-based assessments of pre- and posttreatment of upper facial wrinkles and evaluate their consistency and treatment effectiveness.

Methods: A total of 228 individuals (204 females, 24 males) who received abobotulinum toxin for glabellar, forehead, and lateral canthal wrinkles were analyzed using pre- and posttreatment photographs. Wrinkles were assessed by four human raters using the 5-point Merz scale and Global Aesthetic Improvement Scale (GAIS). AI evaluations were conducted using Haut.AI Face Skin Metrics 2.0, a pre-trained machine learning platform.

Results: AI had better error rates than humans for age prediction. The AI and human assessments showed high agreement for static and dynamic wrinkle evaluations ($P<0.001$). Posttreatment analysis indicated significant wrinkle reduction in both the human and AI assessments ($P<0.001$). Human assessment of GAIS scores was negatively correlated with wrinkle reduction ($P<0.001$). The treatment effects measured by AI and human raters showed a weak-to-moderate correlation.

Conclusion: AI-based assessments align well with human evaluations and can detect posttreatment improvements. However, the treatment effect did not correlate well with human evaluations. AI can serve as an objective tool for evaluating botulinum toxin treatment outcomes and complementing human assessments. However, there is still a need for a gold standard method to evaluate aesthetic improvement and harmony.

Introduction

Botulinum toxin has long been effective in treating upper facial wrinkles [1]. The efficacy of the treatment is assessed through pre- and post-procedure follow-ups using comparative analysis of photographs and wrinkle depth measurement scales [2-4]. Validated scales are commonly used to assess baseline appearance and treatment outcomes; however, they rely on subjective human interpretation, which introduces variability and bias. This subjectivity may limit the standardization and consistency in evaluating treatment efficacy [5].

Artificial intelligence (AI) is a rapidly evolving technology that has been used in medicine. Various algorithms have been used in diagnostic and therapeutic processes. AI has emerged as a promising tool for addressing subjectivity and the lack of standardized measures to evaluate aesthetic outcomes, thereby overcoming the limitations of human perception. By providing objective and quantifiable assessments, AI offers a more reliable approach for documenting the success of aesthetic interventions and understanding their impact. It can be utilized in aesthetic dermatology to recognize facial aesthetic concerns and reassess treatment outcomes [6,7].

Currently, AI-based platforms assist physicians in the dermatological evaluation of the skin by identifying and grading conditions such as dryness, acne, wrinkles, pigmentation disorders, erythema, and signs of photoaging. Some systems also offer treatment-planning functionalities and can simulate the expected posttreatment outcomes. Following therapeutic interventions, objective parameters such as perceived age and wrinkle severity are commonly assessed using pre- and posttreatment photographs, with comparisons drawn between AI-generated outcomes and human evaluations [8,9]. However, the degree of change is often interpreted solely based on absolute pre- and posttreatment scores without a direct comparison between the difference in scores observed by AI and human perception. Incorporating an analysis of the parallelism between these could provide a novel and valuable perspective for evaluating treatment efficacy in dermatological practice.

Objectives

This study aimed to compare upper facial wrinkles following abobotulinum toxin therapy and analyze changes through human and artificial intelligence evaluations.

Materials and Methods

This study was conducted at a tertiary hospital and evaluated patients who received abobotulinum toxin for glabellar, forehead, and lateral canthal wrinkles between November 2022 and May 2024. This study was approved by the local ethics committee (AEŞH-BADEK-2024-046).

Evaluation of Patients Using Photographs

A total of 228 participants were evaluated based on follow-up photographs, following the acquisition of informed consent. All analyses were performed anonymously using the designated platform. Patient data were not stored on any external server and were accessible only to the user. No identifiable personal information was transmitted or retained beyond the local environment, thereby ensuring full compliance with data privacy standards. Static and dynamic facial lines were documented using five standardized photographs taken from front and side views. Images were captured using a smartphone (iPhone 12 Pro; Apple Inc., Cupertino, CA, USA) and categorized as follows: first photo for static facial lines, second for glabellar frowning lines, third for dynamic forehead lines, and fourth and fifth for dynamic lateral canthal lines. To ensure the standardization of image acquisition, all photographs were taken by the same researcher. The photo sessions were conducted in a controlled environment, using the same room and consistent lighting conditions. The camera angles were selected according to the requirements of the platform, specifically the frontal and 45-degree lateral views. During each session, both static and dynamic facial expressions of the participants were guided and controlled by the researcher to maintain uniformity.

AI analysis was conducted using a pre-trained machine learning platform (Haut.AI Face Skin Metrics 2.0). The AI platform utilizes a convolutional neural network (CNN) for age estimation and wrinkle analysis. The age prediction model was trained and validated on a dataset of approximately 1,500 individuals, while the facial analysis model was trained on more than three million images [10,11].

Four researchers (HA) and the Haut.AI Face Skin Metrics 2.0, an artificial intelligence (AI) application, performed an age analysis based on the first pre-treatment photograph and a wrinkle examination on all photographs. AI also performed an age analysis on the first post-treatment photograph in

order to evaluate the internal consistency of the platform.]Wrinkle severity was assessed by the researchers using the 5-point Merz scale, and treatment outcomes were measured using the Global Aesthetic Improvement Scale (GAIS). Human evaluations were scored on the Merz scale, ranging from 0 to 4, with 4 indicating the most severe condition. The GAIS scores ranged from -3 to +3, with +3 representing the most favorable result. AI wrinkle scores were calculated for the entire face as total face, forehead, and crow's feet scores. AI assessments were scored on a scale of 0–100, with 100 representing the best possible outcome. Additionally, the AI application evaluated the quality of the photos on a scale of 100, based on factors such as image clarity, focus, full-face visibility, lighting, and resolution.

Human and AI wrinkle scores were analyzed by matching the human glabella score with the AI forehead score, human forehead score with the AI forehead score, and human lateral canthal line score with the AI crow's feet score. In the necessary analyses, the AI forehead score was compared to the sum of the human glabella and forehead scores, whereas the AI total face score was compared to the sum of the human glabella, forehead, and lateral canthal lines scores.

Statistical Analysis

Age estimation was analyzed using the mean error (ME), mean absolute error (MAE), and root mean square error (RMSE). Age predictions were compared using the Friedman and Spearman's correlation tests. AI pre- and posttreatment analyses were performed using Wilcoxon signed-rank and Friedman tests.

Facial wrinkle analysis was performed for AI, human, and AI-human comparisons, based on individual and average scores. Posttreatment changes were assessed within and between evaluators. For mutual evaluations, the researchers' scores were reverse-coded to align with the AI scoring system. The wrinkle and GAIS scores were standardized to z-scores to account for scale differences. Inter-rater agreement and consistency were evaluated using the intraclass correlation coefficient (ICC), and the treatment effect was assessed using Spearman's correlation and Wilcoxon signed-rank test. An ICC value of ICC<0.5 indicates poor agreement, 0.5<ICC<0.75 indicates moderate agreement, 0.75<ICC<0.90 indicates good agreement, and ICC>0.9 indicates excellent agreement. A correlation coefficient of $|r|\geq 0.8$ indicates a strong relationship, $0.5\leq|r|<0.8$ indicates a moderate relationship, and $|r|<0.5$ indicates a weak relationship.

Statistical analyses were performed using SPSS 25.0, and a p-value of <0.05 was considered significant.

Results

The study analyzed the photographs of 204 female and 24 male participants. Picture quality was assessed based on focus, face recognition, brightness, and resolution scores, with all photos meeting the required quality standards for analysis (85/100 points and above).

Age Prediction Analysis

The mean age of the participants, predicted ages, and age prediction error criteria are listed in Table 1, and the distribution of the predictions according to the actual age is shown in Figure 1.

A strong correlation was observed between each age prediction and actual age (R_{AI}: 0.859, R_{HA1}:0.861, R_{HA2}: 0.824, R_{HA3}: 0.853, R_{HA4}: 0.818, R_{HM}: 0.891, all p-values <0.001). Significant differences were identified in age predictions among the group comparisons (P<0.001). In the subgroup analysis, differences were observed primarily due to disparities between HA2 and the other groups, as well as between HM and AI/HA1 (all p-values <0.001). AI showed

Table 1. Age analysis.

a) Actual and Predicted Age Means (min-max)							
AA	**AI_B**	**AI_A**	**HR1**	**HR2**	**HR3**	**HR4**	**HM**
41.18±9.03 (23-69)	41.02±8.19 (23-62)	40.57±8.22 (22-61)	40.49±8.45 (18-63)	36.65±9.9 (20-62)	40.62±10.21 (22-72)	40.29±9.11 (23-66)	39.51±8.86 (22.5-64)

b) Age Prediction Error							
	AI_B	**AI_A**	**HR1**	**HR2**	**HR3**	**HR4**	**HM**
Mean Error	-0.07±4.78	-0.54±4.95	-0.69±4.55	-4.53±5.69	-0.56±5.55	-0.89±5.33	-1.67±4.17
Mean Absolute Error	3.77±2.92	4.04±2.92	3.60±2.86	5.90±4.24	4.46±3.33	4.19±3.39	3.55±2.73
Root Mean Squared Error	4.77±5.68	4.98±5.77	4.60±5.35	7.26±8.13	5.57±6.74	5.39±6.63	4.48±5.29

Abbreviations: AA: actual age, AI: artificial intelligence, HR: human rater, HM: mean of human predictions, A: after, B: before, min: minimum, max: maximum

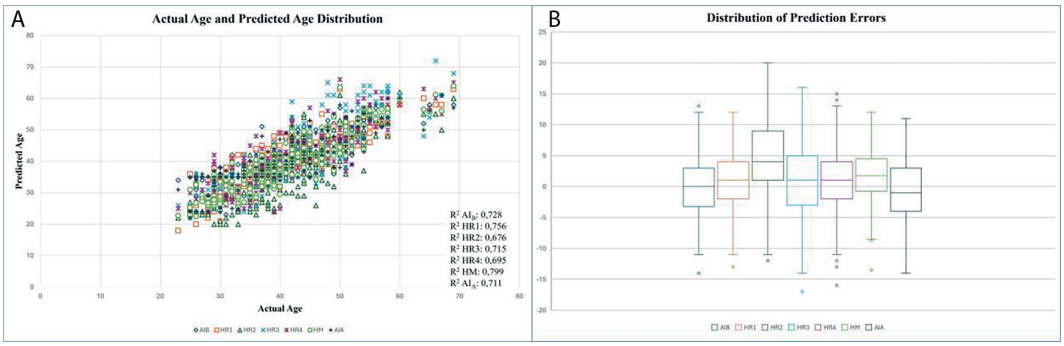

Figure 1. A) Actual Age and Predicted Age Distribution. B) Distribution of Prediction Errors (AI: artificial intelligence, HR: human rater, HM: mean of human predictions, R^2: regression coefficient, A: after, B: before)

Table 2. Pre-treatment wrinkle analysis reliability (ICC scores).

	Human agreement	AI-human agreement	AI-human agreement (mean score)	p
Static glabella	0.913	0.848	0.495	<0.001
Static forehead	0.893	0.915	0.835	<0.001
Static crow's feet (right\left)	0.920	0.910\0.913	0.753\779	<0.001
Dynamic glabella	0.908	0.861	0.526	<0.001
Dynamic forehead	0.891	0.864	0.662	<0.001
Dynamic crow's feet (right\left)	0.930\0.922	0.866\0.831	0.528\0.413	<0.001

Abbreviations: AI: artificial intelligence, ICC: intraclass correlation coefficients, p: p-value

no significant difference between pre- and posttreatment age predictions, with a high correlation (R=0.916, *P*<0.001).

Pre-Treatment Wrinkle Analysis

Excellent agreement was observed in the human pre-treatment static and dynamic wrinkle analyses (all p-values <0.001). The AI-based analysis also showed near-excellent agreement (all p-values <0.001). The agreement between the average human score and AI decreased in the individual score analysis for the static and dynamic lines (Table 2). The agreement between the AI total face score and the sum of the human glabella, forehead, and crow's feet scores as well as the human average score was high (ICC: 0.948, ICC: 0.873, all p-values <0.001, respectively). In the analysis using the sum of the human glabella and forehead line scores, the agreement was excellent, whereas it was good for the average score (ICC: 0.934, ICC: 0.831, all p-values <0.001, respectively).

Posttreatment Wrinkle Analysis

The human posttreatment static and dynamic wrinkle analyses showed near-excellent agreement (all p-values <0.001). In the GAIS analysis, moderate agreement was generally observed among the human raters (all p-values <0.001). The AI-based analysis also showed near-excellent agreement (all p-values <0.001). The agreement between the average human score and AI decreased in the individual score analysis, except for the dynamic forehead lines (Table 3). The agreement between the AI total face score and the sum of the human glabella, forehead, and crow's feet scores was excellent, whereas the agreement with the human average score was moderate (ICC: 0.908, ICC: 0.726, all p-values <0.001, respectively). In the analysis using the sum of the human glabella and forehead line scores, agreement was excellent, whereas it was good for the average score (ICC: 0.928, ICC: 0.778, all p-values <0.001, respectively).

Analysis of Treatment Outcomes

In the human evaluation of treatment effects, a significant improvement was observed in the average wrinkle scores across all photos (all p-values <0.001). There was a negative correlation between the GAIS score for the static facial wrinkles and treatment difference score (all p-values <0.001). Similarly, a negative correlation was found between the GAIS scores of the dynamic lines and the difference score (all p-values <0.001). When analyzing the treatment effect for each rater, an improvement in posttreatment scores was observed for all researchers (all p-values <0.001). Wrinkle score differences were negatively correlated with the GAIS scores (all p-values <0.001) (Table 4).

The AI analysis of treatment effects showed significant improvements in scores for static expression (first photo),

Table 3. Posttreatment wrinkle analysis reliability (ICC scores).

	Human agreement	Human GAIS agreement	AI-human agreement	AI-human agreement (mean score)	p
Static glabella	0.916	0.691	0.883	0.611	<0.001
Static forehead	0.899	0.797	0.913	0.854	<0.001
Static crow's feet (right\left)	0.893	0.668	0.882\0.883	0.694\0.711	<0.001
Dynamic glabella	0.895	0.722	0.891	0.742	<0.001
Dynamic forehead	0.823	0.655	0.873	0.884	<0.001
Dynamic crow's feet (right\left)	0.889\0.885	0.755\0.693	0.826\0.830	0.477\0.475	<0.001

Abbreviations: GAIS: Global Aesthetic Improvement Score, AI: artificial intelligence, ICC: intraclass correlation coefficients, p: p-value.

Table 4. Treatment effect and GAIS correlation.

Photo No.	Anatomical Region	HR1	HR2	HR3	HR4	HM	p
1	Glabella (r)	-0.738	-0.547	-0.771	-0.989	-0.819	<0.001
	Forehead (r)	-0.796	-0.727	-0.914	-0.984	-0.928	<0.001
	Crow's feet (r)	-0.487	-0.687	-0.936	-0.945	-0.896	<0.001
2	Glabella (r)	-0.831	-0.340	-0.943	-0.999	-0.907	<0.001
3	Forehead (r)	-0.574	-0.409	-0.948	-0.973	-0.911	<0.001
4	Crow's feet right (r)	-0.620	-0.442	-0.963	-0.958	-0.855	<0.001
5	Crow's feet left (r)	-0.686	-0.399	-0.963	-0.968	-0.847	<0.001

Abbreviations: HA: human rater, HM: human mean score, r: correlation coefficient, p: p-value.

frown lines and horizontal forehead lines (second and third photos), and crow's feet wrinkles (fourth and fifth photos) (all p-values <0.001).

The AI-human agreement for difference scores was analyzed using individual and averaged researcher scores as well as GAIS, and the results are summarized in Table 5. The significant correlation values ranged from weak to moderate. No correlation was found between AI and HA2 for static glabella lines (first photo) or between AI and HA1/HA2 for dynamic lateral canthal lines (fourth photo). In the fifth photo, no correlation was observed between the AI and any evaluation.

A weak linear relationship was observed between treatment differences in the static and dynamic photographs (Figures 2 and 3).

Discussion

This study offers valuable insights into the application of AI in cosmetic assessments, particularly for age estimation and wrinkle analysis.

These findings demonstrate that AI provides more accurate age predictions with lower error rates than human raters do. Furthermore, a significant difference was observed between one researcher and the others. Previous studies have shown that AI performs well in age estimation, even surpassing human accuracy [12,13]. The proximity of the average age estimates of the three researchers to the actual age appeared to mitigate the substantial deviation observed in the fourth researcher's estimates. This finding underscores the potential utility of AI as a supportive tool for correcting individual variability in patient assessment.

Although no statistical difference was found in the AI's posttreatment age estimations, and the predictions showed a high correlation, there was a tendency for an increase in the error rate. This suggests that even small changes induced in a short time by abobotulinum toxin can be detected by AI. In various studies, procedures such as facelifts, rhinoplasty, and blepharoplasty have resulted in a perceived reduction in age [14-16]. Although the present study did not include a procedure capable of creating a significant difference such as surgical treatment, long-term follow-up of minimally invasive procedures could demonstrate treatment success by showing that the increase in the estimated age remains lower than the actual age. This highlights the significance of AI analysis for the evaluation of treatment outcomes. Additionally, AI's initial age estimation deviations and the degree of posttreatment correction could serve as valuable guides for physicians in patient assessment and decision-making regarding treatment approaches [17].

Wrinkles are often perceived as a hallmark of aging and a key concern in skin aesthetics. Wrinkle evaluation is

Table 5. Correlation of treatment differences.

			HR1	HR2	HR3	HR4	HM	GAIS1	GAIS2	GAIS3	GAIS4	GAISm
AI	Static Lines	Lines score face* (r)	0.453	0.382	0.414	0.419	0.495	0.409	0.330	0.436	0.435	0.509
		p-value	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
		Lines score forehead+ (r)	0.253	0.115	0.202	0.182	0.279	0.320	0.184	0.244	0.196	0.333
		p-value	**0.000**	0.116	**0.005**	**0.012**	**0.000**	**0.000**	**0.011**	**0.001**	**0.007**	**0.000**
		Lines score forehead++ (r)	0.625	0.476	0.360	0.460	0.611	0.659	0.446	0.414	0.482	0.629
		p-value	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
		Lines score forehead+++ (r)	0.596	0.425	0.380	0.440	0.574	0.615	0.373	0.441	0.460	0.600
		p-value	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
		Lines score crow's feet right (r)	0.316	0.232	0.300	0.311	0.358	0.186	0.170	0.268	0.311	0.314
		p-value	**0.000**	**0.001**	**0.000**	**0.000**	**0.000**	**0.010**	**0.019**	**0.000**	**0.000**	**0.000**
		Lines score crow's feet left (r)	0.335	0.345	0.381	0.317	0.417	0.228	0.314	0.359	0.317	0.422
		p-value	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.002**	**0.000**	**0.000**	**0.000**	**0.000**
		Lines score crow's feet mean (r)	0.392	0.351	0.407	0.376	0.467	0.250	0.291	0.378	0.376	0.442
		p-value	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.001**	**0.000**	**0.000**	**0.000**	**0.000**
	Dynamic Lines	Lines score forehead+ (r)	0.223	0.342	0.253	0.275	0.317	0.254	0.157	0.271	0.277	0.322
		p-value	**0.001**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.018**	**0.000**	**0.000**	**0.000**
		Lines score forehead++ (r)	0.513	0.406	0.294	0.347	0.459	0.473	0.398	0.302	0.357	0.503
		p-value	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
		Lines score crow's feet right (r)	0.079	0.086	0.156	0.155	0.144	0.065	0.134	0.192	0.155	0.178
		p-value	0.267	0.225	**0.028**	**0.028**	**0.042**	0.359	0.059	**0.006**	**0.029**	**0.012**
		Lines score crow's feet left (r)	0.140	0.038	0.157	0.140	0.161	0.290	0.074	0.145	0.139	0.203
		p-value	0.071	0.626	0.042	0.070	0.037	**0.000**	0.343	0.060	0.073	**0.008**

*: human glabella, forehead, crow's feet sum +: human glabella ++: human forehead +++: human glabella, forehead sum, HR: human rater, GAIS: Global Aesthetic Improvement Score, M/m: average.

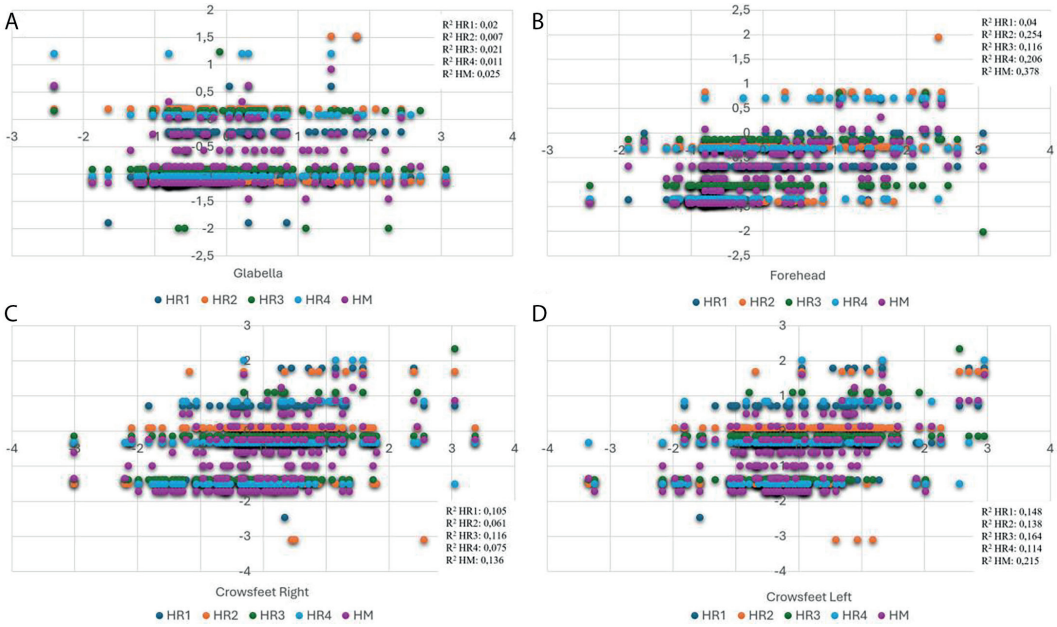

Figure 2. Scatter Plot of Treatment Differences in Static Facial Lines A) Glabellar wrinkles B) Forehead wrinkles C) Right lateral canthal wrinkles D) Left lateral canthal wrinkles (HR: human rater, HM: mean of human predictions).

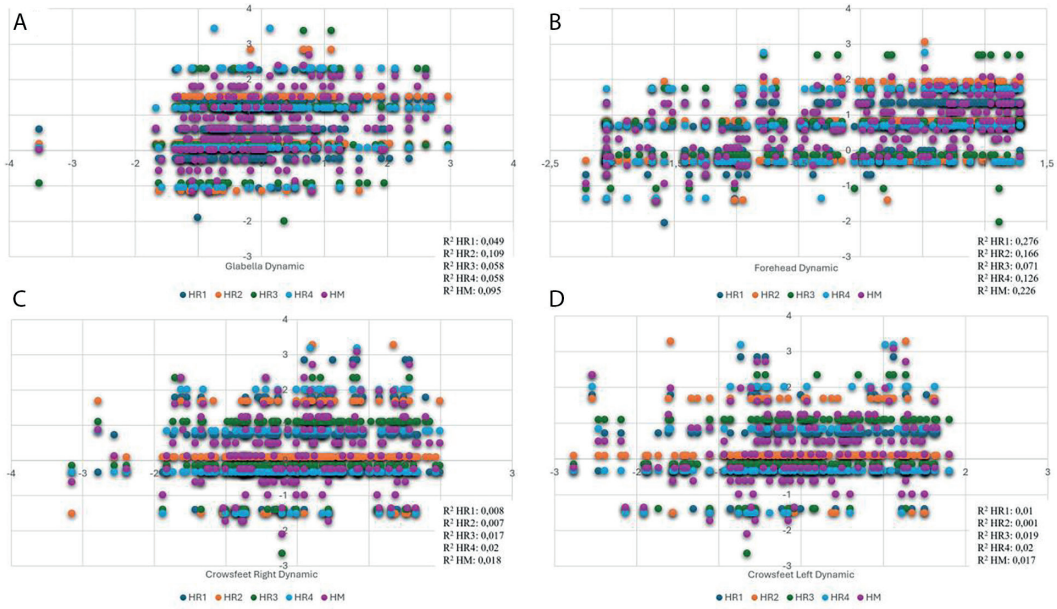

Figure 3. Scatter Plot of Treatment Differences in Dynamic Facial Lines A) Glabellar wrinkles B) Forehead wrinkles C) Right lateral canthal wrinkles D) Left lateral canthal wrinkles (HR: human rater, HM: mean of human predictions).

challenging because of significant variations in their length, depth, and pattern [18]. In the second part of the study, which focused on wrinkle analysis, consistency was observed both among researchers and between AI and researchers. It is well established that AI can detect wrinkles using various algorithms [19]. Studies conducted across different ethnic groups have demonstrated that AI can identify subtle differences and produce results that are consistent with human evaluations [20-22]. In the pre-treatment wrinkle assessment, a decrease in consistency was noted in the average scores of glabellar and lateral canthal lines between AI and researchers compared to individual scores. This decline was only observed in the lateral canthal region in the posttreatment evaluation. This suggests that small variations among individuals in different anatomical regions can lead to significant differences and that wrinkle analysis may be more challenging for humans in certain complex areas. From this perspective, an objective measurement tool that aligns

with human evaluations and reduces variations is essential for dermatological assessments [23,24]. Another advantage of such a tool is its ability to perform numerous analyses in a short time, reducing the burden of tedious and time-consuming evaluations for humans [10].

In the final part of the study, analyzing the treatment effects, both humans and AI observed improvements. However, when evaluating differences in treatment outcomes, a strong correlation between AI and researchers was not found for wrinkle reduction or GAIS. Although there was consistency in wrinkle assessments before and after treatment, this consistency did not extend to the analysis of treatment-related changes. This contrasts with studies suggesting that AI can simulate treatment outcomes or that treatment effects correlate with patient satisfaction [25,26]. Compared to a study in which glabellar wrinkle assessments showed only moderate agreement, our findings showed higher agreement; however, this was not reflected in the analysis of differences [27]. In the study by Yoon and Shin, the treatment outcomes of 34 patients who underwent bipolar radiofrequency and high-intensity focused ultrasound on the face were comparatively analyzed between an AI-based system and two human evaluators. The AI-derived treatment difference score for wrinkle changes immediately posttreatment and at two months was analyzed using binary logistic regression to determine its ability to identify patients who showed improvement according to the GAIS. For human evaluators, an improvement of one point or greater by at least one rater was considered an effective treatment. Based on this analysis, the authors concluded that the AI difference score was consistent with the human assessments. The odds ratios (OR) for posttreatment improvement were 1.79 (P=0.048) and 2.28 at the 2-month follow-up (P=0.013). The area under the curve (AUC) was reported as 0.73 and 0.86, respectively [28].

When the correlation between AI scores and human evaluation was examined, the correlation coefficients ranged from 0.36 to 0.62 (P<0.05), which is similar to the results of our study. In our study, wrinkle outcomes were assessed not only by GAIS but also through a dedicated wrinkle score, with correlation analysis yielding similarly consistent results. Although the binary logistic regression analysis in their study showed a significant relationship between AI and human evaluations in identifying group membership based on the GAIS, this analysis was not performed in our study.

However, it should be noted that including all cases where any rater noted improvement as part of the effective treatment group may have inflated the number of patients considered improved. Yoon and Shin's study did not report individual anatomical subregion analyses or wrinkle scores per region. Moreover, in scenarios where the treatment modality itself significantly determines the outcome, the number of patients in the effective treatment group may be disproportionately high, potentially diminishing the impact of the AI-derived difference score in regression models. In light of these findings, and considering the comparable levels of correlation, the alignment between AI and human raters regarding treatment effects remains controversial.

From the perspective of the treatment difference analysis, the lack of concordance between AI and human evaluations can be interpreted in two ways. First, AI may be capable of detecting subtle changes that the human eye cannot easily discern, particularly in wrinkle assessment. Second, the 5-point ordinal scale used by human raters is inherently less sensitive than the 100-point scale employed by the AI platform, potentially limiting the ability to capture nuanced improvements.

When considering GAIS, a key yet thought-provoking finding is the weak correlation between AI-generated treatment difference scores and GAIS scores assigned by human raters. Given that aesthetic improvement and facial harmony are ultimately judged by human perception, increasing reliance on mathematical modeling by AI raises concerns. The dominance of AI in aesthetic evaluations may drive a trend toward homogenization in beauty standards. Rather than following the fast-paced and industry-driven development of cosmetic AI tools, it may be prudent to actively develop these models in the field of cosmetic dermatology. In an era where individualized aesthetic approaches that take into account ethnic and cultural diversity are emphasized over rigid ideals such as the golden ratio, a cautious approach to reintroducing mathematical models as universal standards seems advisable.

These findings highlight the need to incorporate different methodologies in the development of AI platforms. Although each evaluation method provided valuable insights on its own, the lack of consistency suggests the necessity of a gold standard scale. To improve the scale, an appropriate initial step would be to replicate each analysis across ethnic groups. Subsequently, using rating scales with broader ranges that allow for more sensitive human assessments, integrating biophysical measurements with photographic evaluations into machine learning models, and incorporating posttreatment change scoring as a separate criterion in AI assessments, along with its correlation with human evaluations, may help to address potential practical limitations in this field. Long-term, multicultural, and international studies supported by expert consensus during the training of AI models could contribute significantly to the development of more robust and generalizable systems. However, determining the gold standard remains an open question.

The limitations of this study include its retrospective design, the use of a platform not specifically developed for this

research, which may have introduced inherent limitations, reliance on photographic evaluations, and the lack of full compatibility among the assessment scales.

Conclusion

In conclusion, although consistency was observed between the AI and human evaluations, it was not found in the assessment of treatment effects. The reliability of treatment effect evaluations can be improved through additional gold standard methods and new scales.

References

1. Ong AA, Sherris DA. Neurotoxins. *Facial Plast Surg*. 2019;35(3): 230-8. DOI: 10.1055/s-0039-1688844. PMID: 31189195
2. Lemperle G, Holmes RE, Cohen SR, Lemperle SM. A classification of facial wrinkles. *Plast Reconstr Surg*. 2001;108(6):1735-50; discussion 51-2. DOI: 10.1097/00006534-200111000-00048. PMID: 11711957
3. Carruthers A, Carruthers J. A validated facial grading scale: the future of facial ageing measurement tools? *J Cosmet Laser Ther*. 2010;12(5):235-41. DOI:10.3109/14764172.2010.514920. PMID: 20825260
4. Lim T, Kerscher M, Ogilvie A, et al. Novel Validated Five-point Photonumeric Scales for Assessment of Static and Dynamic Forehead Lines. *Plast Reconstr Surg Glob Open*. 2023;11(9):e5287. DOI: 10.1097/gox.0000000000005287. PMID: 37744770
5. Frank K, Day D, Few J, et al. AI assistance in aesthetic medicine-A consensus on objective medical standards. *J Cosmet Dermatol*. 2024. DOI: 10.1111/jocd.16481. PMID: 39091136
6. Elder A, Cappelli MO, Ring C, Saedi N. Artificial intelligence in cosmetic dermatology: An update on current trends. *Clin Dermatol*. 2024;42(3):216-20. DOI :10.1016/j.clindermatol.2023.12.015. PMID: 38181887
7. Kania B, Montecinos K, Goldberg DJ. Artificial intelligence in cosmetic dermatology. *J Cosmet Dermatol*. 2024;23(10):3305-11. DOI: 10.1111/jocd.16538. PMID: 39188183
8. Elder A, Ring C, Heitmiller K, Gabriel Z, Saedi N. The role of artificial intelligence in cosmetic dermatology-Current, upcoming, and future trends. *J Cosmet Dermatol*. 2021;20(1):48-52. DOI: 10.1111/jocd.13797. PMID: 33151612
9. Ahmadi N, Niazmand M, Ghasemi A, Mohaghegh S, Motamedian SR. Applications of Machine Learning in Facial Cosmetic Surgeries: A Scoping Review. *Aesthetic plastic surgery*. 2023;47(4):1377-93. DOI: 10.1007/S00266-023-03379-Y. PMID: 37277660
10. Georgievskaya A. Artificial Intelligence Confirming Treatment Success: The Role of Gender- and Age-Specific Scales in Performance Evaluation. *Plast Reconstr Surg*. 2022;150(4 Suppl):34s-40s. DOI: 10.1097/prs.0000000000009671. PMID: 36170434
11. Georgievskaya A, Tlyachev T, Kiselev K, et al. Predicting human chronological age via AI analysis of dorsal hand versus facial images: A study in a cohort of Indian females. *Experimental Dermatology*. 2024;33(3):e15045. DOI: 10.1111/exd.15045. PMID: 38509744
12. Zhang BH, Chen K, Lu SM, et al. Turning Back the Clock: Artificial Intelligence Recognition of Age Reduction after Face-Lift Surgery Correlates with Patient Satisfaction. *Plastic and reconstructive surgery*. 2021;148(1):45-54. DOI: 10.1097/PRS.0000000000008020. PMID: 34181603
13. Du H, Liang H, Peng B, Qi Z, Jin X. Age Reduction After Face-Lift Surgery in Chinese Population: An Outcome Study Using Artificial Intelligence and Objective Observer-Based Assessment. *Aesthetic Plast Surg*. 2024. DOI:10.1007/s00266-024-04258-w. PMID: 39085528
14. Gibstein AR, Chen K, Nakfoor B, et al. Facelift Surgery Turns Back the Clock: Artificial Intelligence and Patient Satisfaction Quantitate Value of Procedure Type and Specific Techniques. *Aesthet Surg J*. 2021;41(9):987-99. DOI: 10.1093/asj/sjaa238. PMID: 33217756
15. Goodyear K, Saffari PS, Esfandiari M, Baugh S, Rootman DB, Karlin JN. Estimating apparent age using artificial intelligence: Quantifying the effect of blepharoplasty. *Journal of plastic, reconstructive & aesthetic surgery : JPRAS*. 2023;85:336-43. DOI: 10.1016/J.BJPS.2023.07.017. PMID: 37543022
16. Elliott ZT, Bheemreddy A, Fiorella M, et al. Artificial intelligence for objectively measuring years regained after facial rejuvenation surgery. *Am J Otolaryngol*. 2023;44(2):103775. DOI: 10.1016/j.amjoto.2022.103775. PMID: 36706713
17. Bobrov E, Georgievskaya A, Kiselev K, et al. PhotoAgeClock: deep learning algorithms for development of non-invasive visual biomarkers of aging. *Aging (Albany NY)*. 2018;10(11):3249-59. DOI: 10.18632/aging.101629. PMID: 30414596
18. Fujino S, Iwanaga T. Real-time wrinkle evaluation method using Visual Illusion-based image feature enhancement System. *Skin Res Technol*. 2023;29(1):e13206. DOI: 10.1111/srt.13206. PMID: 36382793
19. Osman OF, Elbashir RMI, Abbass IE, Kendrick C, Goyal M, Yap MH. Automated assessment of facial wrinkling: A case study on the effect of smoking. *2017 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2017*. 2017;2017-January: 1081-6. DOI: 10.1109/SMC.2017.8122755. PMID: -
20. Flament F, Jacquet L, Ye C, et al. Artificial Intelligence analysis of over half a million European and Chinese women reveals striking differences in the facial skin ageing process. *Journal of the European Academy of Dermatology and Venereology : JEADV*. 2022;36(7):1136-42. DOI: 10.1111/JDV.18073. PMID: 35279898
21. Flament F, Zhang Y, Jiang R, et al. Objective and automatic grading system of facial signs from selfie pictures of South African women: Characterization of changes with age and sun-exposures. *Skin research and technology : official journal of International Society for Bioengineering and the Skin (ISBS) [and] International Society for Digital Imaging of Skin (ISDIS) [and] International Society for Skin Imaging (ISSI)*. 2022;28(4):596-603. DOI: 10.1111/SRT.13153. PMID: 35490368
22. Park H, Park SR, Lee S, et al. Development and application of artificial intelligence-based facial skin image diagnosis system: Changes in facial skin characteristics with ageing in Korean women. *Int J Cosmet Sci*. 2024;46(2):199-208. DOI: 10.1111/ics.12924. PMID: 37881146
23. Flament F, Velleman D, Yamashita E, et al. Japanese experiment of a complete and objective automatic grading system of facial signs from selfie pictures: Validation with dermatologists and characterization of changes due to age and sun exposures. *Skin research and technology : official journal of International Society for Bioengineering and the Skin (ISBS) [and] International Society*

for Digital Imaging of Skin (ISDIS) [and] International Society for Skin Imaging (ISSI). 2021;27(4):544-53. DOI: 10.1111/SRT.12982. PMID: 33368725

24. Sattler S, Frank K, Kerscher M, et al. Objective Facial Assessment With Artificial Intelligence: Introducing the Facial Aesthetic Index and Facial Youthfulness Index. *J Drugs Dermatol*. 2024;23(1):e52-e4. DOI: 10.36849/jdd.7080. PMID: 38206157
25. Landau M, Goldust M. Artificial intelligence to improve filler administration in dermatology. *J Cosmet Dermatol*. 2024;23(9):3045-6. DOI: 10.1111/jocd.16472. PMID: 39015042
26. Flament F, Maudet A, Ye C, et al. Comparing the self-perceived effects of a facial anti-aging product to those automatically detected from selfie images of Chinese women of different ages and cities. *Skin research and technology : official journal of International Society for Bioengineering and the Skin (ISBS) [and] International Society for Digital Imaging of Skin (ISDIS) [and] International Society for Skin Imaging (ISSI)*. 2021;27(5):880-90. DOI: 10.1111/SRT.13037. PMID: 33822402
27. Yoelin S, Green JB, Dhawan SS, et al. The Use of a Novel Artificial Intelligence Platform for the Evaluation of Rhytids. *Aesthetic surgery journal*. 2022;42(11):NP688-NP94. DOI: 10.1093/ASJ/SJAC200. PMID: 35869540
28. Yoon W, Shin HK. Efficacy of artificial intelligence-based skin analysis for calculating wrinkle improvement and skin firmness after simultaneous radiofrequency and high-intensity focused ultrasound therapy: a retrospective clinical study. *Arch Aesthetic Plast Surg*. 2025;31(2):46-54. DOI: 10.14730/aaps.2025.01340.