Proceedings of ICMMGH 2027 Symposium: Clinical OMICS for Drug Discovery and Personalized Medicine
DOI: 10.54254/2753-8818/2027.MB38000

AI in Skincare Personalisation: A Review of Diagnostic Accuracy and Recommendation Validity

Isabelle Teng

St Paul's Girls' School, London, UK
isabelle.teng08@gmail.com

Abstract. Artificial Intelligence (AI) has become a standard feature of consumer skincare, with major brands including L'Oréal and Neutrogena now offering photograph-based diagnostic tools that translate a single selfie into a personalised product routine. The technology behind these systems have progressed rapidly, from convolutional neural networks (CNN) to newer multimodal systems. However, the basis underpinning this entire market that a meaningful individualised regimen can be produced relies on several unresolved problems and has received far less scrutiny than it should. This review discloses problems that undermine this claim, such as imbalanced training data and an unproven link between prediction and improved outcomes. No trials have been conducted on whether an AI-personalised regimen outperforms a standardised one, and only isolated cases monitor diagnostic output against established clinical scales. These gaps are heightened by a regulatory structure that frames the products as being wellness tools rather than diagnostic devices, thus allowing companies to bypass validation and data-protection standards that would otherwise be applied, as well as introduce a structural conflict of interest in which the party diagnosing a skin condition is the same one selling the treatment. Current systems demonstrate meaningful advances in prediction, but personalisation remains an assumption rather than a demonstrated outcome clinically. Addressing this will require accuracy investigations and controlled trials that the field has not yet undertaken.

Keywords: Skincare Personalisation, Artificial Intelligence (AI), Accuracy, Recommendation Validity

1. Introduction

In the past decade, there has been a clear shift in the cosmetics industry from being a category built on expert recommendations to one increasingly mediated by algorithms. Academic and industry literature has followed closely, with recent work applying deep learning techniques directly to the classification of skin conditions and mapping those conditions onto product recommendations.

Many consumer skin-analysis tools utilise convolutional neural networks (CNN) trained on large datasets of facial images. This can be used in conjunction with labels such as age, wrinkle severity, or pigmentation. One example uses deep learning to classify skin conditions, mapping them directly onto product categories [1]. Another CNN-based system frames itself explicitly as automating a dermatologist's visual assessment [2].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2027.MB38000

A significant proportion of these models are trained to predict what is marketed as 'biological skin age', in which visual features are associated with the chronological age of a patient. This term can lead to a slight misinterpretation as a person's chronological age is not a biological measurement, and a model trained in this way will use population-average aging patterns rather than the individual's actual dermal condition. 'Biological skin age' appears to be a deeper biological insight than it is in practice: an estimate of how old an individual looks, relative to the 'average' face.

Recently, certain systems have been developed which use vision transformers or multimodal architectures as an alternative to CNNs, integrating texture, colour, and spatial information to capture localised pigmentation or micro-texture changes better than a standard CNN. One such system built for extended reality (XR) environments is indicative of this trend towards multimodal capture, combining 2D image data with depth or 3D surface information [3]. Earlier work on facial expression recognition, although not specific to skincare, illustrated how fusing 2D and 3D facial data through deep architectures can outperform single-modality models - a methodological point skincare application has only recently begun to adopt [4]. This is an important distinction, as a model trained on chronological age with a standard CNN is performing something conceptually different and, debatably, less clinically meaningful than a multimodal system attempting to characterise actual tissue properties such as elasticity, hydration-related texture, or vascular pigmentation.

L'Oréal uses tools built on technology from ModiFace, an augmented-reality and AI company which the brand acquired in 2018 - most notable are SkinConsultAI and Skin Genius [5]. A user uploads a bare-faced selfie, and the algorithm scores this image against several aging variables, including under-eye wrinkles, firmness, fine lines, dark spots, and pores. The system then uses the information to suggest a bespoke product routine. The model is trained on thousands of clinical images drawn from L'Oréal's in-house Skin Aging Atlases and refined against selfies taken in varied lighting conditions across multiple ethnic groups, and L'Oréal has since claimed high agreement with dermatologist grading. This brand offers more transparency than most; nevertheless, more difficult questions, such as whether the suggested personalised routine outperforms a standardised one, remain unaddressed.

To bridge this gap, this review examines the validity of AI-driven skincare personalisation across its diagnostic and prescriptive layers. The main objective is to evaluate whether current algorithmic predictions genuinely support effective individualised regimens or merely serve as advanced marketing tools. The primary aims are as follows: (1) to analyse how training data biases compromise diagnostic accuracy across diverse populations; (2) to investigate the empirical disconnect between image-based prediction and true outcome personalisation; and (3) to interrogate the regulatory, ethical, and commercial landscapes. This work provides a structured critique of current technical limitations and outlines key prerequisites which are necessary to elevate beauty AI from prediction to clinically proven personalisation.

2. Problems with the training data

A model's output is only as credible as the population it was trained on - this is where the clearest problem in consumer skincare AI lies. Even comparatively transparent companies such as L'Oréal do not report accuracy broken down by the groups it claims to serve, and the disparity between what training data is typically used for and what needs to be covered to support accurate personalisation for all users must be examined.

DOI: 10.54254/2753-8818/2027.MB38000

Dermatology AI has a well-documented history of underperforming on darker skin tones, largely attributed to the fact that dermatological image datasets are disproportionally drawn from lighter Fitzpatrick skin types (I-III) [6, 7]. This imbalance has been noted in cosmetogenomics and broader dermatology AI literature as a structural issue rather than an incidental gap [8]. As a result, the practical consequence in terms of skin age and pigmentation prediction is that accuracy is not uniform across all users: a model may perform better for the demographic on which it was trained, but substantially worse for others. Regardless of this, the model will present its output with the same apparent confidence. This inconsistency in underlying accuracy has been raised as a specific driver of misleading recommendations in the field of beauty AI [9], as a personalisation layer built on an unevenly accurate diagnosis inherits this unevenness, although it will not disclose this. Even brands that publish some such dataset detail rarely report accuracy broken down by skin tone, rather than simply describing the training set as 'diverse'.

A less discussed but equally important issue is that facial datasets do not just skew by skin tone but also incorporate a bias towards beauty norms. What is seen as 'aged' texture, 'enlarged' pores, or 'uneven' pigmentation is not a neutral biological classification; rather, it reflects the labelling conventions and cultural assumptions encoded in the dataset by the scientists annotating it. These features can be inadvertently tagged by a model as deviations from a 'healthy' baseline simply because they look different to the majority in the training data. This risks an AI system pathologising normal variation rather than assessing actual skin health.

Both issues are heightened by the fact that consumer-facing apps almost never publish their training data demographics or report validation accuracy by skin tone. Thus, it is impossible for a user to know for certain whether their 'personalised' result came from a model well-suited towards their skin type or one trained largely on a different population. This transparency gap seems to be a reasonably resolvable problem, fixed simply by disclosure. Unless companies are willing to publish such information, users are left basing the accuracy of their results on trust alone.

A single photograph is also a poor instrument for capturing a biologically dynamic system, regardless of how sophisticated the model analysing it may be. Skin conditions fluctuate with several factors including diet, sleep, hydration status, and hormonal cycles. A model treating a single image as a stable biological truth ignores all transient variations and risks mistaking a temporary state (eg. dehydrated skin after a poor night's sleep) for a stable trait requiring a permanent product solution. Genuine personalisation in a clinically meaningful sense would require tracking an individual's response to a particular active over time and adjusting the regimen based on the observed response, rather than using a one-off prediction. Systems that collect additional information rather than relying on a photo alone attempt to address this issue. The Optune system by Shiseido [10] delves deeper than a diagnosis based off of a single photo: a companion app captures a facial image along with self-reported sleep and menstrual cycle data, combining this with environmental inputs such as temperature and humidity. The app then utilises an algorithm which selects a combination of serums and moisturisers which are dispensed by a dedicated home device. This is an example progressing further towards the ideal of longitudinal tracking, as the system is explicitly designed to adjust its output on a regular basis rather than generating a single static recommendation.

3. Prediction vs personalisation

Even with the assumption of a reasonably accurate diagnostic layer, there is another fundamental question needing to be addressed: is there sufficient evidence to suggest that an AI-personalised regimen produces better outcomes than a standardised, dermatologically-approved routine?

DOI: 10.54254/2753-8818/2027.MB38000

Currently, this question remains open and has been rarely tested. Some system-description papers account detailed descriptions of the diagnostic pipeline (how a recommendation engine works, what data it ingests, what categories of product it outputs) but fail to justify the claim that personalisation substantially improves outcomes, preferring to leave this as an implicit assumption. The main idea is that greater specificity of input should translate to greater benefit of output - that is, more details being recorded must lead to a better regimen. This assumption, while logical, has not been tested empirically and therefore should not be taken as fact.

There seems to be a rather straightforward solution to this oversight, albeit an expensive trial to run - a randomised, controlled trial comparing an AI-personalised regimen against a standardised routine, using pre-registered objective outcome measures instead of self-reported satisfaction. No such trial appears to have been documented in literature, nor in public material published by brands' marketing strategies. Considering this, companies may therefore be compounding the uncertainty of an unvalidated recommendation step on top of a diagnostic step of already-varying accuracy. The relevant question is therefore not whether an algorithm can generate a plausible-sounding regimen from a photo but whether that regimen outperforms a standardised routine or a dermatologist's visual assessment.

Most AI systems have also not been reviewed against a dermatologist's evaluation. A model's diagnostic output should be compared with an established clinical instrument such as SCINEXA for skin aging or the Fitzpatrick and Glogau scales for pigmentation and photoaging severity. One notable partial exception to this is Neutrogena's Skin360 [11]. This analyses a selfie across several attributes, including hydration, smoothness, tone evenness, wrinkles, and dark spots, to produce an overall numeric score alongside a recommended skincare routine. Neutrogena's imaging scientists sampled the AI-generated scores against the Griffiths scale, a standard dermatological grading system, and found meaningful correlation between the two [8]; this is one of few examples of a company affirming its AI tool against established clinical standards, a practice that should arguably be implemented across this industry. However, it should be noted that this tests whether the AI's diagnostic is accurate, but there are no records of controlled comparisons between an AI-personalised regimen and one that is standardised. Clinical assessment using validated scales, although perhaps imperfect, is currently the closest available standard, and claims of diagnostic accuracy can be seen as unanchored within this comparison.

A small number of companies have started to combine facial imaging with skin microbiome swabs or single-nucleotide polymorphism (SNP) genotyping, an approach reviewed systematically under the heading of 'cosmetogenomics' [8]. The theoretical appeal is relatively straightforward: genetic variants and microbial community composition plausibly influences how an individual's skin responds to specific active ingredients, potentially grounding a recommendation in mechanism rather than surface appearance alone. The success in current SNP panels and microbiome markers in predicting differential response to specific actives (eg. retinoids, niacinamide) and whether this additional data collection leads to a substantial return in outcome must be reviewed.

While the logic in the use of AI in individualised skincare appears at face value to be straightforward, the reality of the situation is that diagnostic models with demographically uneven accuracy are feeding recommendation systems with no demonstrated advantage over standardised alternatives, usually based off static images that do not properly represent the microbiome of the skin.

DOI: 10.54254/2753-8818/2027.MB38000

4. The regulatory and ethical gap

Most consumer skincare AI apps are marketed under a 'wellness' or lifestyle framing rather than as medical devices. This is not an incidental position, but allows these products to avoid clinical validation, accuracy reporting and regulatory reviews that bodies (eg. FDA in the US, MHRA in the UK) require from tools making diagnostic claims in a medical context. These companies tell the user their 'biological skin age' is above their chronological age and recommend a serum rather than diagnosing a skin condition requiring treatment - there is no real difference between these two claims, but this nuance allows for a different regulatory bar depending on the marketing angle.

This regulatory gap intersects with a further ethical issue concerning the collection of facial images and supplemental genomic, microbiome, or lifestyle data by companies operating outside strict medical-data regulation. In most jurisdictions, biometric and genetic data are subject to stricter consent and privacy protections as they uniquely identify an individual and cannot be changed if compromised. Menstrual and sleep data occupy a similarly sensitive category in many privacy frameworks yet are collected through a consumer skincare app rather than a regulated health product. Whether these skincare apps currently have consent processes and data-handling practices meeting the standard expected of such data in a clinical setting is difficult to determine from an outsider's perspective, but it is a relevant and increasingly urgent question given the rapid growth of this data collection.

There is also a structural conflict of interest which raises some concern. In conventional clinical practice, the party selling a treatment is typically separate from the one diagnosing the condition, to avoid any commercial incentives shaping the assessment. This, therefore, in principle, preserves the validity of the diagnosis. In the consumer skincare AI model, both roles are executed by the same company. One such example of this is Neutrogena [11], where the routines recommended by the AI are made of products solely drawn from the company's own line, raising the issue of problems concerning divided interests. This fact does not necessarily indicate that any diagnosis is dishonest, but the system lacks the structural safeguard against biased diagnoses seen in clinical medicine.

5. Consumer satisfaction

The absence of an agreed outcome measure is another underlying problem which affects most of this field. Even where companies do publish some sort of validation, it is typically self-reported user satisfaction. This measure of outcome is vulnerable to placebo and expectation effects: a regimen framed as 'personalised specifically for you' may be perceived as more effective than an identical formula presented as a generic routine, independent of any actual difference in the products. This is a well-established effect in broader consumer health and would need to be explicitly controlled in any randomised control trials comparing personalised regimens against standard ones; For instance, by not disclosing to participants which group of the study they are in.

More objective outcome measures exist but are underused in this industry, including transepidermal water loss as a measure of skin barrier function and standardised photographic scoring using validated instruments (eg. SCINEXA, Fitzpatrick, Glogau, or Griffiths scale). Until studies in this field adopt such outcome measures rather than relying on consumer self-reporting, any claims of diagnostic accuracy or recommendation efficacy will remain difficult to substantiate with confidence.

DOI: 10.54254/2753-8818/2027.MB38000

6. Conclusion

AI-driven skincare diagnostics show a very high potential in the industry, and models based on CNNs and transformers can plausibly extract meaningful insights from facial images. Major brands have already implemented this technical development for consumer use. However, the implicit assumptions leading from predictions to personalisation are currently supported by remarkably little evidence in both commercial examples and academic literature. Training data biases mean diagnostic accuracy is unevenly distributed across skin tones and likely incorporate beauty norms as biological fact. The belief that AI-personalised skincare outperforms standardised alternatives is unvalidated, and diagnostic accuracy has been tested against clinician assessment in only isolated cases. Factors that would increase genuine personalisation, including longitudinal response tracking and validated genomic or microbiome integration, remain largely aspirational, even in more sophisticated commercial systems. The regulatory framing of these tools as 'wellness' products, in addition to the business incentive of the companies acting as a conflict of interest, means there is little motivation to address these unanswered questions.

This review is itself limited in scope, as it draws on publicly available material and the absence of information undisclosed by companies concerning training data demographics or validation methodology can only be noted. The systems discussed are a non-exhaustive sample of major consumer brands, and smaller, more clinically orientated competitors may address some of the issues identified more rigorously than the companies discussed in this paper. The cosmetogenomics and microbiome-integration approaches are covered at the level of concept rather than in technical depth, reflecting how little documentation exists in this subfield.

Future work in this area would benefit most directly from RCTs to validate claims that an AI-personalised regimen is substantially more beneficial than a standardised routine, using more objective outcome measures instead of self-reported satisfaction. Independent reviews of diagnostic accuracy by skin type, conducted outside of commercial incentives of the brands whose products are being evaluated, would be equally valuable. In the long term, the ambition of longitudinal tracking for personalisation will depend on whether genomic and microbiome data can be proven to meaningfully predict differential response to specific actives; until such evidence exists, the claim remains that AI can individualise skincare, but the industry is selling personalisation on the strength of prediction alone.

References

[1] Nithyadevi, S., Rose, B., Senthil Kumar, S. S., Vijayalakshmi, A., Kingsly Jabakumar, A., & Velmurugan, V. (2025). Facial skincare product recommendation using deep learning techniques. *Communications on Applied Nonlinear Analysis, 32*, 267–278.

[2] Yadav, A., & Singh, K. (2025). Smart dermatology: Revolutionizing skincare with AI-driven CNN-based detection and product recommendation system. *Journal of Scientific Innovation and Advanced Research (JSIAR), 1*(2), 230–235.

[3] Rajegowda, G., Spyridis, Y., Villarini, B., & Argyriou, V. (2024). An AI-assisted skincare routine recommendation system in XR. https://doi.org/10.1007/978-981-99-9018-4_28

[4] Li, H., Sun, J., Xu, Z., & Chen, L. (2017). Multimodal 2D+3D facial expression recognition with deep fusion convolutional neural network. *IEEE Transactions on Multimedia, 19*(12), 2816–2831.

[5] L'Oréal and ModiFace: An artificial intelligence-powered skin diagnostic. (n.d.). https://www.loreal.com/en/news/research-innovation/loreal-and-modiface-an-artificial-intelligencepowered-skin-diagnostic/

[6] Olsen, D., Nguyen, C., Hall, M., Carletti, M., & Weis, S. E. (2026). The efficacy of the Fitzpatrick scale in clinical practice. *HCA Healthcare Journal of Medicine, 7*(1), 5–11. https://doi.org/10.36518/2689-0216.2207

DOI: 10.54254/2753-8818/2027.MB38000

[7] Ehsan, E., Truong-Balderas, K., Ituarte, B. E., Clark, K., & Wei, E. X. (2025). Skin of color representation in cosmetic dermatology literature, 2018-2023. *Archives of Dermatological Research, 317*(1), 458. https://doi.org/10.1007/s00403-024-03767-7

[8] Haykal, D., Flament, F., Amar, D., Cartier, H., Kourosh, A. S., Lee, D. H., & Rowland-Payne, C. (2025). Cosmetogenomics unveiled: A systematic review of AI, genomics, and the future of personalized skincare. *Frontiers in Artificial Intelligence, 8*, 1660356. https://doi.org/10.3389/frai.2025.1660356

[9] Bharani, T. (2026, May 17). Algorithms to aesthetics: Understanding the misleading recommendation of AI in beauty. SSRN. https://ssrn.com/abstract=7281098

[10] Shiseido Company. (2019, July 1). Shiseido launches full-scale IoT*1 skincare service brand "optune" | news release | Shiseido Company. *Corp.Shiseido.Com*. https://corp.shiseido.com/en/news/detail.html?n=00000000002712

[11] Meintjes, M. (2025, September 5). Haut.AI and Neutrogena refresh AI-driven skin analysis tool for science-backed inclusivity. *CNS Media*. https://www.personalcareinsights.com/news/hautai-neutrogena-skin360-refresh.html