OPEN ACCESS

Impact of real-world confounders on the accuracy of an AI model to support read out of skin prick automated test results

Karolien Roux[1,*], Sven F. Seys[2,3,*], Valérie Hox[4], Adam M. Chaker[5], Peter W. Hellings[1,6], Glynnis De Greve[7], Winde Lemmens[8], Anne-Lise Poirrier[9], Rembert Daems[2,10], Dirk Loeckx[2], Senne Gorris[2,11], Laura Van Gerven[1,6,12]

[1] Department of Otorhinolaryngology, Head and Neck Surgery, UZ Leuven, Leuven, Belgium
[2] Hippo Dx, Aarschot, Belgium
[3] Department of Otolaryngology, Head and Neck Surgery, Medical University of Vienna, Vienna, Austria
[4] Service d'Otorhinolaryngologie, Cliniques Universitaires Saint-Luc, Brussels, Belgium
[5] Department of Otorhinolaryngology and Center of Allergy and Environment (ZAUM), TUM School of Medicine and Health, TUM University Hospital Klinikum rechts der Isar, Technical University of Munich, Munich, Germany
[6] Allergy and Clinical Immunology Research Group, Department of Microbiology, Immunology & Transplantation, KU Leuven, Leuven, Belgium
[7] Department of Otorhinolaryngology-Head and Neck Surgery, ZAS Sint-Augustinus, Antwerp, Belgium
[8] Department of Otorhinolaryngology-Head and Neck Surgery, ZOL, Genk, Belgium
[9] ENT Department, CHU Liège, Liège, Belgium
[10] IDLab, Ghent University-imec, Ghent, Belgium
[11] Department of Otorhinolaryngology, AZ Herentals, Herentals, Belgium
[12] Laboratory of Experimental Otorhinolaryngology, Department of Neurosciences, KU Leuven, Leuven, Belgium

* shared first authorship

Rhinology 64: 5, 0 - 0, 2026
https://doi.org/10.4193/Rhin25.634

Received for publication:
November 3, 2025
Accepted: April 18, 2026

Associate Editor:
Sietze Reitsma

Dear Editor:

Skin prick testing (SPT) is the gold standard for diagnosing allergic sensitization in individuals with a suspected airborne allergy [(1)]. Skin tests are used as first option in 90% of individuals suffering from respiratory allergies and almost two-third of all types of allergies [(2)]. However, its accuracy highly depends on the operator, causing variability during pricking and readout. S.P.A.T. or skin prick automated test standardises the SPT procedure and has been clinically validated [(3)]. The S.P.A.T. device serves 12 simultaneous pricks delivering a fixed amount of test solution with controlled prick force to the patient's forearm. This automation has proven to lower intra-subject variability and bring more consistent test results compared to manual SPT [(4,5)].

Recently, an AI-assisted readout method was developed using a train-test approach including over 10,000 wheals following S.P.A.T. [(6)]. Comparing the measurements by AI with those of the physician revealed an accuracy of 95.4%. Performance evaluation in an independent cohort of 95 patients showed misclassifications that impacted the test interpretation in 0.5% of cases. Given that false positive or negative test results were mostly due to scars, hair, hyperpigmentation or darker skin tone, these real-world confounders were investigated in a post-hoc analysis.

Images of 217 patients (2,604 wheals) of the validation cohort recruited in the previous pivotal study were analysed. The images of the forearms were analysed for skin tone, presence of hair, scars, tattoos and hyperpigmentation. Analysis of skin tone was performed using the individual typology angle (ITA) method [(7)]. Other possible confounders were assessed by visual inspection of the images and applying a qualitative score.

Table S1 provides an overview of the real-world confounders that were assessed. Figures S1-S2 show representative images for each confounder subgroup. Darker skin tone (ITA <0) was observed in 16.8% of patients. Presence of hair, hyperpigmentation, scar tissue or tattoos near the prick location was observed in respectively 30.4%, 3.5%, 1.0% and 1.5% of pricks. In patients with a darker skin tone (ITA < -50, ITA -50 to <-25, -25 to <0) accuracy decreased to respectively 85.4%, 90.3% and 92.0% compared to other ITA categories (96.5-96.7%) (Figure 1A). In patients with tattoo marks (category III), accuracy dropped to 88.3% (compared to 95.7-97.9%) (Figure 1B). For the other confounders, accuracy remained >90% irrespective of the presence of the confounding variable (Figure 1C-E).

The limited representation of patients with darker skin tones, or hyperpigmentation, scars and tattoos is a weakness of this study and may underestimate the extent of bias. Visuo-perceptive data and scoring are prone to inter-observer variability. It should also

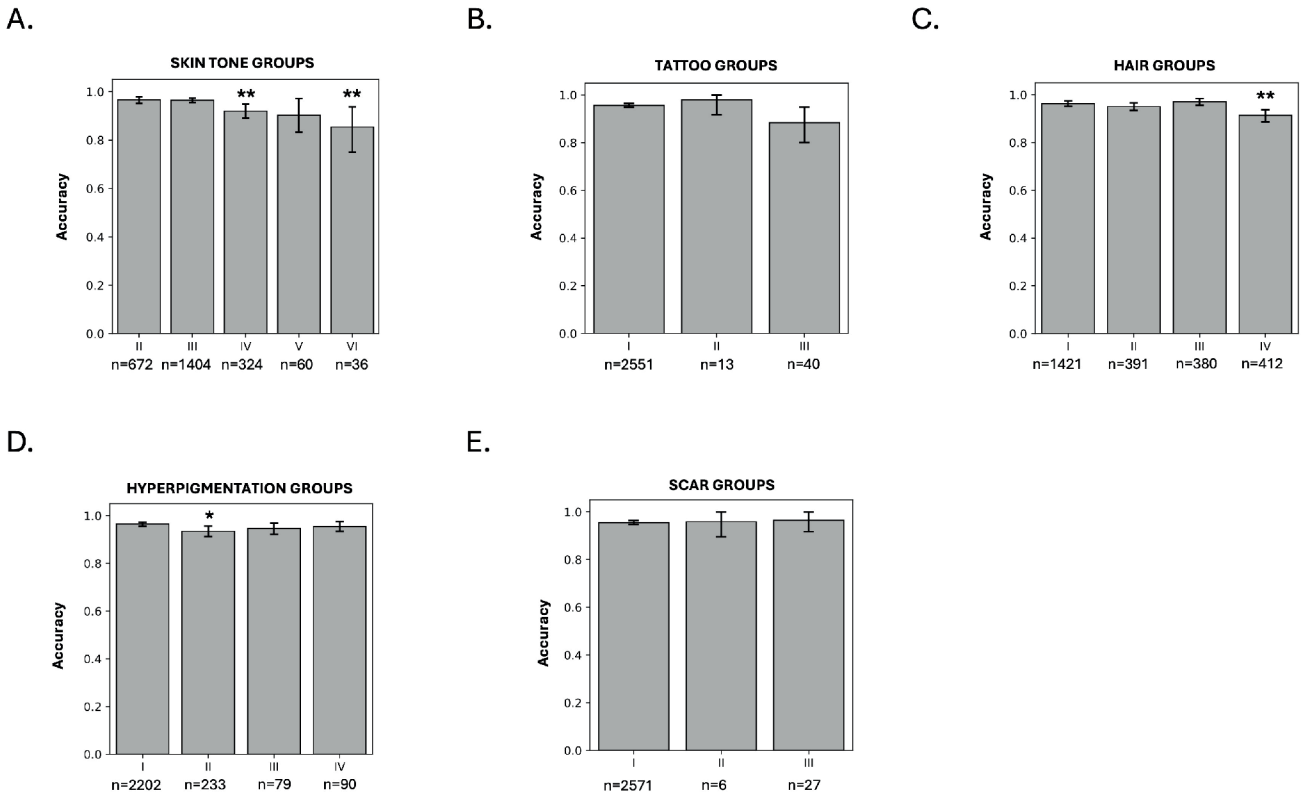

Figure 1. Analysis of accuracy of the AI model in patients stratified by real-world confounders. (A) Skin tone assessed by individual typology angle analysis. I: ITA>50, II: ITA 25 to 50, III: ITA 0 to 25, IV: ITA -25 to 0, V: ITA -50 to -25, VI: ITA <-50. (B) Presence of tattoos assessed by visual inspection: I: no tattoos visible, II: tattoo visible at the side of the arm, III: tattoo visible in close proximity to this prick location. (C) Presence of hair assessed by visual inspection: I: no hair visible, II: minimal hair visible at the side of the arm, III: minimal hair visible in close proximity to this prick location, IV: abundant hair visible in close proximity to this prick location. (D) Presence of hyperpigmentation assessed by visual inspection: I: no hyperpigmentation visible, II: diffuse hyperpigmentation visible, III: Isolated hyperpigmentation visible at the side of the arm, IV: Isolated hyperpigmentation in close proximity to this prick location. Presence of scars assessed by visual inspection: I: no scars visible, II: scar visible at the side of the arm, III: scar visible in close proximity to this prick location. Data are presented as mean with 95% confidence intervals. Between-group comparison by generalised estimating equations model analysis. *: p<0.05; **: p<0.01.

be noted that for now, the AI-assisted readout method can only process images in combination with S.P.A.T. given the need for 32 images captured under various lightning conditions for optimal accuracy of the AI model [8]. Strengths of our study relate to the large number of skin reactions analysed and the systematic evaluation of multiple confounders providing a comprehensive real-world picture.

The SPAT AI-assisted readout method showed high accuracy (>90%) in most patients irrespective of hair, hyperpigmentation or scar tissue, reinforcing the robustness of the AI model in real-world conditions. Darker skin tone and the presence of tattoo marks were observed in some patients and lowered performance. A brightness slider has recently been added in the web viewer, allowing better visualisation of images with darker skin tone and subsequent evaluation of the AI-assisted readout. Further evaluation of this method in more diverse patient populations is needed to confirm external validity. This will lead to more fair and inclusive AI models to support allergy diagnostics so more patients can benefit from these emerging technologies.

Acknowledgements

We would like to thank all physicians and study coordinators who contributed to this study.

Authorship contribution

VH, AC, GDG, WL, ALP, and LVG were responsible for recruitment. KR, SFS, DL, RD, and LVG were responsible for data analysis. RD was responsible for the development of the AI algorithm. The study was conceived, designed, set-up, analyzed, and interpreted by KR, SFS, DL, SG and LVG. SFS, DL and LVG have verified the underlying data. All authors accept responsibility for the decision to submit for publication.

Conflict of interest

SFS, RD, DL are employees of Hippocreates BV. SFS, RD, DL, SG and LVG hold shares of Hippocreates BV. SFS serves/ed as associate editor for Respiratory Medicine (ongoing) and Heliyon Immunology (2022-2024). AMC reports grants, speaker honoraria, consultancy or advisory fees and/or research support and other, all via Technical University of Munich from Allergopharma, ALK Abello, Astra Zeneca, Bencard/Allergen Therapeutics, GSK, Novartis, Hippo Dx, LETI, Roche, Zeller, Sanofi, Regeneron, Thermo Fisher, European Institute of Technology (EIT Health) and Federal Ministry of Research and Education Germany. Other authors have nothing to disclose related to this study.

Funding

The study was funded by Hippocreates BV. Hippocreates BV was supported by a grant from VLAIO (HBC.2021.1170). LVG was supported by the Research Foundation Flanders (FWO) Senior Clinical Investigator Fellowship (18B2222N).

References

1. Heinzerling L, Mari A, Bergmann KC, et al. The skin prick test – European standards. Clin Trans Allergy. 2013 Feb 1;3:3.
2. Cardona V, Demoly P, Dreborg S, et al. Current practice of allergy diagnosis and the potential impact of regulation in Europe. Allergy. 2018 Feb;73(2):323–7.
3. Seys SF, Gherasim A, Odul F, et al. Validation of the Skin Prick Automated Test (SPAT) cut-off value in birch pollen and house dust mite allergic rhinitis patients. Allergy. 2025 Dec;80(12):3302-3309.
4. Gorris S, Uyttebroek S, Backaert W, et al. Reduced intra-subject variability of an automated skin prick test device compared to a manual test. Allergy. 2023 May;78(5):1366–8.
5. Seys SF, Roux K, Claes C, et al. Skin Prick Automated Test device offers more reliable allergy test results compared to a manual skin prick test. Rhinology. 2024;62(2):216–22.
6. Seys SF, Hox V, Chaker AM, et al. Artificial Intelligence (AI)-assisted readout method for the evaluation of skin prick automated test results. Nat Commun. 2025 Oct 1;16(1):8637.
7. Krishnapriya KS, King MC, Bowyer KW. Analysis of manual and automated skin tone assignments for face recognition applications [Internet]. arXiv; 2021 [cited 2025 Oct 21]. Available from: http://arxiv.org/abs/2104.14685
8. Daems R, Seys S, Hox V, et al. Improved Allergy wheal detection for the skin prick automated test device. In: Bellazzi R, Juarez Herrero JM, Sacchi L, Zupan B, editors. Artificial Intelligence in Medicine. Cham: Springer Nature Switzerland; 2025. p. 116–20.

Sven Seys, PhD
Hippo Dx
Betekomsesteenweg 69E
3200 Aarschot
Belgium

E-mail: sven.seys@hippo-dx.com

This manuscript contains online supplementary material

SUPPLEMENTARY MATERIAL

Materials and Methods

Ethics

The study (CIV-23-04-042754) complied with all relevant ethical regulations, was approved by the institutional review boards (Belgium: a Belgian recognized Ethics Committee assigned by the Belgian competent authority; Germany: Ethikkommission der Technischen Universität München) and registered online at www.clinicaltrials.gov (NCT05918354).

Accuracy evaluation of the AI algorithm

Images of 217 patients (2,604 wheals) of the validation cohort recruited in the previous pivotal study (NCT05918354) were analysed. The AI measurements of the longest wheal diameter were compared with the measurements by the treating physician. A confusion matrix (2x2 table) was created with binary values (true: wheal ≥ 4.5mm or false: wheal < 4.5mm) of the AI and physician measurements per wheal for each of the pre-defined patient subgroups (based on the confounder categories, see Table S1). Accuracy was determined for each the confusion matrices.

Quantitative & qualitative scoring of the images

The images of the forearms were analysed for skin tone, presence of hair, scar tissue, tattoos and hyperpigmentation. Analysis of skin tone was performed quantitatively using the individual typology angle (ITA) method per patient [7]. Other possible confounders were assessed by visual inspection of the images per prick location and applying a qualitative score by 2 observers separately. In case of mismatch between the 2 observers, these images were evaluated together after which consensus was reached.

Statistical analysis

To analyze differences in accuracy between groups, we fitted a generalized estimating equations (GEE) model with a binomial distribution and logit link function. This approach accounts for the non-independence of repeated observations within clusters (i.e. the patients). Group was included as a categorical predictor, with group I (no visible signs of the confounder) serving as the reference category, except for skin tone where the largest group (group III) was selected as reference category to account for bias based on skin tone. Patients with a tattoo (group II-III) were excluded from the skin tone analysis to avoid bias because of the tattoo colour in the ITA analysis. Accuracy with 95% confidence intervals were reported in the graphs. All analyses were conducted in Python using the statsmodels package.

Table S1. Applied measurements for the confounder analysis of the AI-assisted readout method.

Confounders	Measurement	Number of pricks	Percentage
Skin tone	*ITA (Individual Typology Angle) score*		
I	>50	0	0.0%
II	25 and 50	672	26.9%
III	0 and 25	1404	56.3%
IV	-25 and 0	324	13.0%
V	-50 and -25	60	2.4%
VI	< -50	36	1.4%
Hair	*Qualitative score*		
I	No hair visible	1421	54.6%
II	Hair visible at the side of the arm (not in close proximity to this prick location)	391	15.0%
III	Minimal hair visible in close proximity to this prick location	380	14.6%
IV	Abundant hair visible in close proximity to this prick location	412	15.8%
Hyperpigmentation	*Qualitative score*		
I	No hyperpigmentation visible	2202	84.6%
II	Diffuse hyperpigmentation visible (and no isolated hyperpigmentation in close proximity to this prick location)	233	8.9%
III	Isolated hyperpigmentation visible at the side of the arm	79	3.0%
IV	Isolated hyperpigmentation in close proximity to this prick location	90	3.5%
Scar	*Qualitative score*		
I	No scars visible	2571	98.7%
II	Scar visible at the side of the arm (not in close proximity to this prick location)	6	0.2%
III	Scar visible in close proximity to this prick location	27	1.0%
Tattoo	*Qualitative score*		
I	No tattoos visible	2551	98.0%
II	Tattoo visible at the side of the arm (not in close proximity to this prick location)	13	0.5%
III	Tattoo visible in close proximity to this prick location	40	1.5%

Skin tone was assessed by individual typology angle (ITA) analysis per patient. The presence of hair, hyperpigmentation, scars and tattoos was assessed qualitatively per prick location.

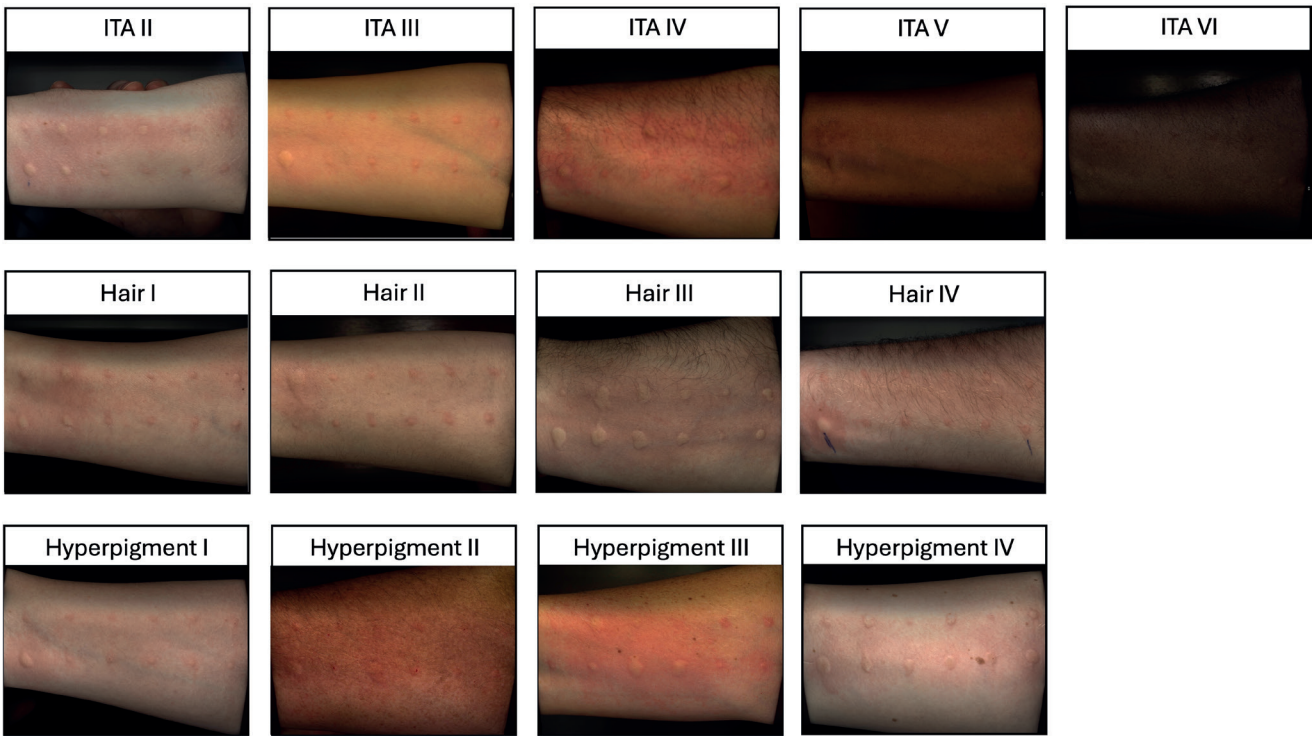

Figure S1. Representative images of patient's forearms per skin tone, hair and hyperpigmentation class. Skin tone was assessed by individual typology angles (ITA) analysis per patient. The presence of hair or hyperpigmentation was assessed qualitatively per prick location.

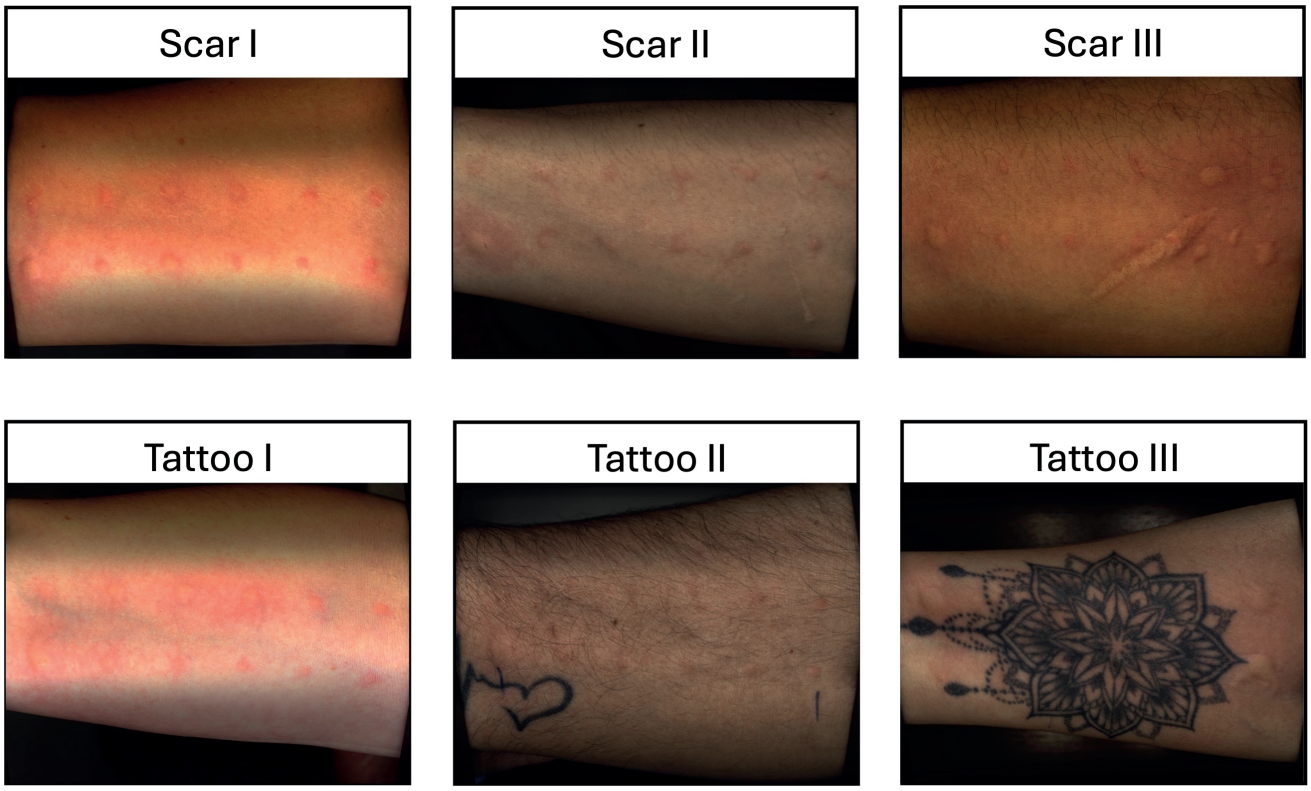

Figure S2. Representative images of patient's forearms per scar and tattoo class. The presence of scars or tattoos was assessed qualitatively per prick location.