


Artificial Intelligence (AI)-assisted readout method for the evaluation of skin prick automated test results

Received: 3 January 2025

Accepted: 12 September 2025

Published online: 01 October 2025

 Check for updates

Sven F. Seys¹  , Valérie Hox², Adam M. Chaker³ , Glynnis De Greve⁴, Winde Lemmens⁵, Anne-Lise Poirrier⁶, Eline Beckers⁵, Rembert Daems^{1,7}, Zuzana Diamant^{8,9,10} , Carmen Dierickx⁵, Peter W. Hellings^{8,11}, Caroline Huart², Claudia Jerin³, Mark Jorissen^{11,12}, Dirk Loeckx¹ , Hanne Oscé⁴ , Karolien Roux⁹, Mark Thompson¹³ , Sophie Tombu⁶, Saartje Uyttebroek^{11,12}, Andrzej Zarowski⁴, Senne Gorris^{1,14} & Laura Van Gerven^{8,11,12}

The skin prick test (SPT) is the gold standard for diagnosing allergic sensitization to aeroallergens. The Skin Prick Automated Test (SPAT) device has previously demonstrated reduced variability and more consistent test results compared to manual SPT. The current study aims to develop and validate an artificial intelligence (AI) assisted readout method to support physicians in interpreting skin reactions following SPAT. To train the AI algorithm, 7812 wheals (651 patients) are manually labeled. To validate the AI measurement, the longest wheal diameter of 2604 wheals (217 patients) is measured by the treating physician and compared to the AI measurement. In addition, AI-assisted readout is validated on a separate test cohort of 95 patients (1140 wheals). We demonstrate that the AI measurements of the longest wheal diameter exhibit a strong correlation with the physician's measurements. The AI algorithm shows a specificity of 98.4% and sensitivity of 85.0% in determining positive or negative test results in the validation cohort. In the test cohort, physicians adjust 5.8% of AI measurements, leading to a change in the test interpretation for only 0.5% of cases. AI-assisted readout significantly reduces inter- and intra-observer variability and readout time compared to manual physician measurements. Altogether, the AI-assisted readout method demonstrates high accuracy, with minimal misclassification of test results. Adding AI to SPAT further improves standardization across the SPT process, significantly reducing observer variability and time to readout.

Worldwide, 30% of the population suffers from allergic diseases to aeroallergens, with pollen and house dust mite allergies being the most commonly reported^{1,2}. Allergic diseases are driven by sensitization to specific allergens. During this process, specific immunoglobulin E (IgE) is produced by B cells and subsequently binds to Fcε receptors on mast cells, causing crosslinking of the receptors and local histamine release.

The skin prick test (SPT), first introduced in 1959, in part mimics this sequence of events by the application of a small amount of allergen into the skin to provoke a local reaction, thus providing a diagnostic tool to identify allergic sensitization³. Despite a few technological advancements in more recent years, SPT still remains the gold standard for diagnosing allergic diseases, alongside serum-specific IgE

measurement^{4,5}. Although SPT has been proven to be a more sensitive test to detect allergic sensitization than serum specific IgE⁶, it is not without limitations, as its accuracy is highly dependent on the expertise of the operator, the devices and test solutions used. Several studies have highlighted significant variability due to these factors^{7,8}. A joint task force of several American Allergy Associations has called for standardization and consistency in SPT procedures⁹. A survey conducted by the German Society for Allergology highlighted the considerable variability in the execution of SPTs between different hospitals and offices in Germany^{10,11}.

Recently, an automated device, Skin Prick Automated Test (SPAT), has been developed to improve consistency and reduce operator-dependent variability. The SPAT device performs twelve simultaneous epidermal pricks, applying a fixed amount of allergen and controlled prick force on the patient's forearm. After 15 min, images of the forearm are captured to allow digital measurement and interpretation of the wheal responses by the practitioner. Two independent validation studies demonstrated that SPAT has lower intra-subject variability and provides more consistent test results compared to conventional SPT^{12,13}. These advancements have now been implemented in several clinical practices, offering standardization of allergy testing to patients.

To further standardize the full chain of the SPT process, the current study was designed to train and validate a readout method that uses artificial intelligence (AI) to detect and measure the wheals on the SPAT images, thereby supporting practitioners in the readout of the test results.

Results

897 patients were recruited for the training and clinical validation of the AI algorithm (Fig. 2; appendix pp 8). Of these, 29 (3.2%) were excluded because the prick tests were not executed ($n = 6$), no imaging of the prick test was performed ($n = 14$), an error occurred during imaging (2) or during uploading the images ($n = 7$). 868 patients were randomly assigned to the training ($N = 651$, 75%) and the validation ($N = 217$, 25%) cohort, respectively. After validation of the AI algorithm, an additional 96 patients were recruited to prospectively evaluate the fully operational AI-assisted readout method (test cohort: $N = 95$; 1

patient was excluded because no imaging was performed). Patient demographics and clinical characteristics for each of the three cohorts are listed in Table 1.

Clinical validation of the AI algorithm

To validate the AI algorithm, the longest wheal diameter measured by the AI was compared to the gold standard, i.e., the measurement of the longest wheal diameter by the treating physician. A strong correlation between both assessments was observed (Pearson $r = 0.83$; $p < 0.0001$; Fig. 1A). The difference between the AI and the physician measurements (absolute difference median: 0.2 mm, 95% CI: 0.1–0.2; relative difference median: –2.8%, 95% CI: –5.1 to –0.8) is shown in Fig. 1B. In addition to the AI longest diameter, we also calculated the AI area and AI average diameter and assessed their correlation with the physician longest diameter. The correlation coefficient between the AI area and the physician longest diameter was slightly better (Pearson $r = 0.86$; $p < 0.0001$; Fig. 3B appendix pp10), whereas it was slightly worse between the AI average diameter and the physician longest diameter (Pearson $r = 0.76$; $p < 0.0001$; Fig. 3C appendix pp10).

Next, accuracy was assessed based on the number of positive (≥ 4.5 mm) and negative (< 4.5 mm) test results. Taking into account all 2604 wheals (allergens and controls), specificity of 98.4%, sensitivity of 85.0%, positive predictive value (PPV) of 93.3% and negative predictive value (NPV) of 96.2% were observed (Table 2). When only considering allergen wheals, sensitivity was lower (77.4%), whereas specificity (98.6%), PPV (90.7%) and NPV (96.1%) all remained above 90%.

False positive and false negative test results were observed in 1.2% and 3.4% of cases, respectively, resulting in an accuracy of 95.4%. Analysis of false positive and false negative test results per site can be found in Table 2 (appendix pp 7). Major causes for inaccuracy were scars, hyperpigmentation, dark skin tone and hairy forearms (Fig. 2). False positive and false negative test results were analyzed by an independent physician, who concluded that the AI algorithm was correct in 66.0% and incorrect in 24.0% of these cases. In 10.0% of cases, it was not possible to determine whether the AI or the treating physician was correct. Adding these cases in which the AI algorithm was deemed correct increased the accuracy to 98.4% (95.4% + (4.6% \times 0.66)).

Table 1 | Patient characteristics

	All patients	Training cohort	Validation cohort	Test cohort
Number of patients, n	963	651	217	95
Female–male–other (%)	60.1–39.8–0.1	60.5–39.4–0.2	60.1–39.9–0.0	57.9–42.1–0.0
Age, mean \pm SD	40.2 \pm 14.7	40.8 \pm 14.4	39.3 \pm 15.5	38.2 \pm 14.2
Smoking status (never–ex–current), (%)	77.7–8.8–13.5	78.1–9.6–12.4	79.5–8.3–12.2	74.5–6.4–19.1
Self-reported allergy, (%)	53.9	57.4	48.8	42.2
Allergy medication, (%)	33.8	36.3	28.0	30.9
Asthma medication, (%)	11.5	12.8	6.7	13.7
VAS total nose & eye symptoms	5.0 (2.0–7.0)	5.0 (2.0–7.0)	5.0 (2.0–7.0)	5.0 (1.0–7.0)
VAS runny nose	4.0 (1.0–7.0)	4.0 (1.0–7.0)	4.0 (1.0–7.0)	4.0 (1.0–6.5)
VAS blocked nose	5.0 (1.0–8.0)	5.0 (1.0–8.0)	5.0 (1.0–8.0)	5.0 (1.0–7.0)
VAS itchy eyes	2.0 (0.0–6.0)	3.0 (0.0–6.0)	2.0 (0.0–6.0)	2.0 (1.0–6.0)
VAS itchy nose	2.0 (0.0–5.0)	2.0 (0.0–6.0)	2.0 (0.0–5.0)	2.0 (0.0–5.0)
VAS sneezing	4.0 (1.0–7.0)	4.0 (1.0–7.0)	4.0 (1.0–6.0)	4.0 (1.0–6.5)
VAS post-nasal drip	3.0 (0.0–7.0)	3.0 (0.0–7.0)	3.0 (0.0–7.0)	4.0 (0.0–7.0)
VAS impaired smell	1.0 (0.0–5.0)	2.0 (0.0–5.0)	1.0 (0.0–4.0)	1.0 (0.0–5.0)
VAS facial pain	3.0 (0.0–7.0)	3.0 (0.0–7.0)	3.0 (0.0–6.8)	2.0 (0.0–6.3)
VAS coughing	2.0 (0.0–5.0)	2.0 (0.0–5.0)	2.0 (0.0–4.0)	2.0 (0.0–5.0)
VAS dyspnea	1.0 (0.0–5.0)	1.0 (0.0–5.0)	1.0 (0.0–5.0)	1.5 (0.0–4.3)
VAS wheezing	0.0 (0.0–3.0)	0.0 (0.0–3.0)	0.0 (0.0–3.0)	0.0 (0.0–2.5)

Data are represented as median and interquartile range (for non-parametric data) unless otherwise specified. SD: standard deviation.

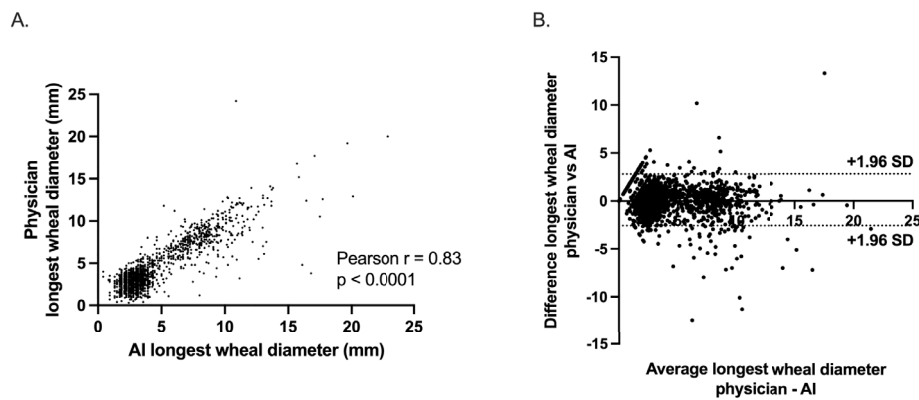


Fig. 1 | Agreement between the measurement of the longest wheal diameter by AI versus the treating physician. **A** Correlation between measurement of the longest wheal diameter by the AI algorithm versus the measurement by the treating physician ($n = 217$ patients and 2604 wheals). The correlation was analyzed by a

two-sided Pearson test. **B** Bland-Altman plot showing the absolute difference between the measurement of the longest wheal diameter by the AI algorithm versus the treating physician, in relation to the average of the longest diameter measured by the AI algorithm and the treating physician ($n = 217$ patients and 2604 wheals).

Table 2 | Analysis of accuracy of SPAT AI model using the measurement of the longest wheal diameter

AI vs physician	Control wheals (N = 434)	Allergen wheals (N = 2170)	All wheals (N = 2604)
Specificity	96.9%	98.6%	98.4%
Sensitivity	96.7%	77.4%	85.0%
PPV	96.7%	90.7%	93.3%
NPV	96.9%	96.1%	96.2%

The longest wheal diameter was measured by the treating physician and by the AI model. A cutoff value of 4.5 mm was applied to determine test positivity.

Evaluation of the performance of the SPAT AI-assisted readout method

Since the AI-assisted readout is planned to serve as a reliable method to support the measurement by the physician at the outpatient clinic, a separate test cohort of 95 patients was analyzed. Out of a total of 1140 wheals, 5.8% ($n = 63$ wheals) of the AI measurements were adjusted (median change of adjusted wheals and interquartile range: -1.4 mm (-2.1 mm to -0.8 mm)) by the treating physician (Fig. 3A). The adjustment led to a change in the interpretation of the test result (switch positive-negative) in only 0.5% ($n = 5$) of the cases (Fig. 3A).

Analysis of inter-observer and intra-observer variability

To underscore the need for standardization in the readout of skin prick test results, we evaluated the real-world variability between and within three physicians. The wheals ($n = 1200$) of a random selection of 100 patients from the validation cohort were evaluated by three physicians (inter-observer variability). The longest wheal diameter showed a median coefficient of variation (CoV) of 19.8% (IQR: 11.7–29.8%) (Fig. 4A). The CoV was highest for glycerol-control wheals (24.2%, IQR: 14.8–40.6%) followed by allergen wheals (20.3%, IQR: 12.0–30.2%) and lowest in histamine wheals (12.3%, IQR: 7.8–18.1%). Interestingly, CoV was significantly reduced when the same three physicians evaluated these wheals with AI-assisted readout for each of the wheal categories ($p < 0.0001$ and $p = 0.002$; Fig. 4A).

In addition, the physicians also measured the wheals of a random sample of 30 out of the 100 patients ($n = 720$ wheals) a second time to evaluate the intra-observer variability. Considerable variation of the longest wheal diameters was observed for each of the physicians: CoV = 12.9% (IQR: 5.5–23.9%) for physician 1, 9.4% (IQR: 3.4–22.3%) for

physician 2 and 6.5% (IQR: 3.3–11.4%) for physician 3 (Fig. 4B). Again, CoV was significantly reduced when the physicians evaluated the wheals with AI-assisted readout ($p < 0.0001$; Fig. 4B).

Finally, we evaluated how long it takes for the physician to perform the readout of the skin reaction manually or with the SPAT AI-assisted readout. We observed that physicians completed the readout 3.7 times faster with the AI-assisted readout (median: 23.9, IQR: 11.3–58.5 s) compared to the manual readout (median: 88.5 s, IQR: 57.3–154.5 s) ($p < 0.0001$ and $p = 0.0008$; Fig. 3B).

Safety

No adverse events have been reported during the study.

Discussion

Applying the SPAT device in a total of 963 patients (11,556 wheals), we demonstrated the clinical validity of an AI-assisted readout method for accurate and time-saving readout of SPT results by health care professionals. The high accuracy of the AI algorithm was confirmed by comparison of the AI measurements with those performed by the treating physician. The higher consistency was confirmed by showing that the inter-observer and intra-observer variability of physician measurements supported by the AI-assisted readout method were significantly lower compared to the manual measurements by the physicians. Furthermore, the time to readout was reduced 3.7-fold when physicians were supported by AI compared to the manual measurement. Adding the AI-assisted readout method to the SPAT device makes it the first of its kind solution for standardization of the entire SPT process.

As previously discussed, the variability introduced by the operator performing the prick test, can be reduced by the use of the SPAT device^{12,13}. In addition, we now show that the readout of the skin reactions is prone to a similar level of variability. We detected a median inter-observer variability of 20%. This is in line with a previous study demonstrating inter-observer variability in scoring and interpreting epidermal skin tests⁸ and underscores the need for standardization in SPT readout.

The first attempts to automate the readout of SPT results date from the early nineties. Initially, scanning methods of wheals marked with a pen and transferred with adhesive tapes on white paper were described^{14,15}. Later, 3D cameras or infrared cameras were used to detect and quantify the skin reactions^{16–18}. Morales et al. described a device based on a 3D camera, digitizing the SPT results. They found a moderate correlation between the measurement by the device and the

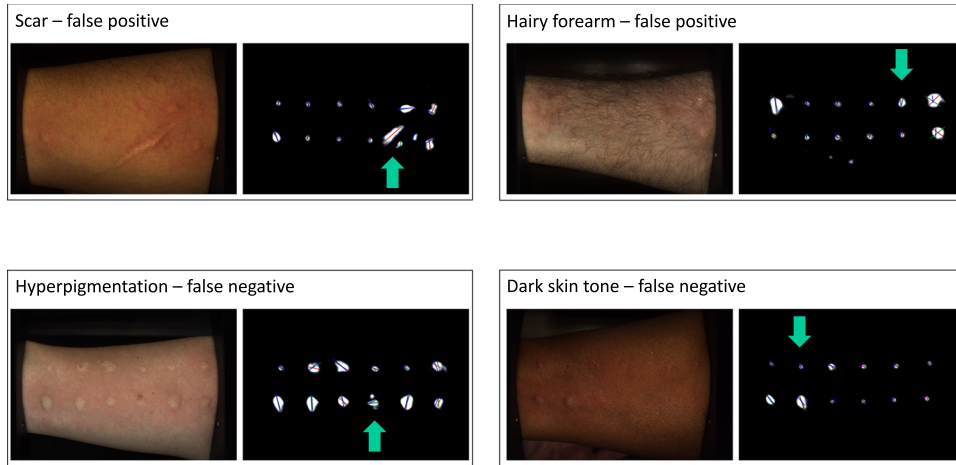


Fig. 2 | Representative images of difficult cases for the AI algorithm measurement. Errors due to the presence of scar tissue, hairy forearm, hyperpigmentation, dark skin tone are marked by a green arrow.

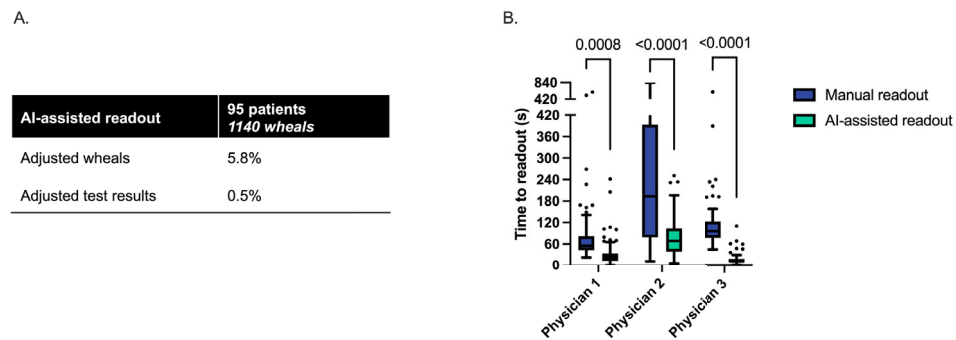


Fig. 3 | Performance of the AI-assisted readout method. **A** Treating physicians were asked to evaluate the measurement by the AI algorithm and adjust the measurement in case of need. The proportion of wheals that required adjustment, as well as the proportion of test results that shifted from positive to negative or vice versa, were analyzed in the test cohort ($n = 95$ patients and 1140 wheals). **B** The time

to perform readout of the test by SPAT, with or without AI-assisted readout, is represented ($n = 100$ patients randomly selected from the validation cohort). Data are presented as Tukey box and whiskers, including median with interquartile range. Repeated measures ANOVA and Sidak's multiple comparison test were performed to compare AI-assisted readout to manual readout.

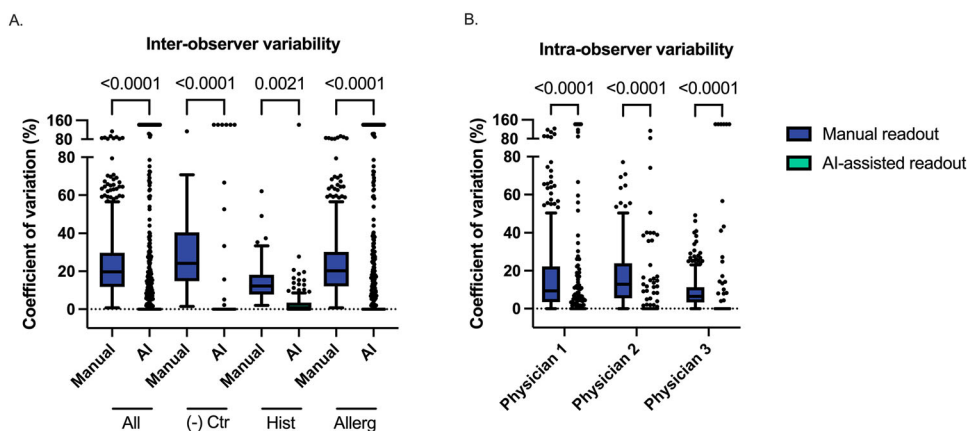


Fig. 4 | Variability of skin prick automated test measurements by multiple observers. **A** Inter-observer variability was analyzed by comparing the coefficient of variation of the longest wheal diameter measured by three physicians with or without support by the AI-assisted readout ($n = 100$ patients randomly selected from the validation cohort). All: all wheal analyzed, (-) ctr: negative control wheals, Hist: histamine wheals, Allerg: allergen wheals. **B** Intra-observer variability was

analyzed by comparing the coefficient of variation of the longest wheal diameter measured twice by three physicians with or without support from the AI-assisted readout ($n = 30$ patients randomly selected from the validation cohort). Data are presented as Tukey box and whiskers, including median with interquartile range. Repeated measures ANOVA and Sidak's multiple comparison test were performed to compare AI-assisted readout to manual readout.

manual procedure, with an intraclass coefficient of 0.68 and 0.49, for the first and second reading of the longest wheal diameter, respectively¹⁷. A follow-up study applying this method to intradermal testing also showed low to moderate agreement (0.4 and 0.65) for the first and second readings of these tests, comparing the device and the manual procedure¹⁹. In both studies, tests were performed with control solutions only. AI-based algorithms may result in better outcomes since Neumann et al. reported a 93.6% accuracy of a convolutional neural network analyzing skin reactions after imaging combining infrared and visible light cameras¹⁸. However, in this study, the imaging technique based on thermography posed challenges related to the blood flow and hair on the forearms.

There are a few other manuscripts also reporting on AI-based algorithms for segmentation of SPT images^{18,20,21}. Accuracy of these segmentation models ranged from 85.9% to 99.9% with rather low or moderate sensitivity (56.2–70.2%). Gomes et al. used smartphone images as input, but the segmentation model relied on a fully convolutional network instead of U-net, which was applied by the other groups and reported to preserve fine details in images, thereby being the preferred option for medical and small object segmentation²⁰. A detailed and unbiased comparison of different deep learning architectures would be worth investigation and is reserved for future work. Such a previous study indicated that U-net remains an excellent option²². It should also be noted that none of these studies so far provided wheal diameters as outcomes, allowing physicians to interpret the test results and validate the approach against current standard clinical practice.

In our study, the longest wheal diameter measured by the AI algorithm correlated well with the measurement by the physician (Pearson $r = 0.83$). More importantly, the accuracy of the AI measurement to detect positive or negative skin reactions was high, with a specificity of 98.4% and a sensitivity of 85.0%. The remaining false positive and negative AI measurements were reviewed afterwards by an independent physician, showing that the AI was correct in 66.0% of the cases. In cases where the AI measurement produced false positive or false negative results, misclassification was primarily due to the presence of scars, excessive hair on the forearms, hyperpigmentation or a darker skin tone. Our findings were further validated in an independent cohort in which we showed that the AI-assisted readout method provided a correct qualification of the SPAT results in 99.5% of cases.

The high performance of the AI-assisted readout may be explained by the fact that we apply a patented imaging technique to visualize the wheals on the patient's forearm using a 2D camera and altering angles of illumination in combination with a U-net segmentation model and an additional rule-based algorithm for measurements of the longest wheal diameter. As such, we overcome many of the challenges others may have faced in the past. The strengths of our study also lie in the extensive number of real-life cases used to train and validate the AI algorithm, including both control and allergen-induced wheals. Also, clinical validation was performed in two independent cohorts. A limitation of the study is that the AI-assisted readout is dependent on imaging of the skin reactions through the SPAT device only. In addition, external validity needs to be confirmed since forearms with scars, hyperpigmentation, darker skin tone, excessive hair, and tattoos are limited in the current dataset. To address the potentially lower performance of the AI algorithm in these cases, the dataset of future training cohorts will be enriched with such cases. Also, external validation with cases from other hospitals in the real-world setting will further contribute to the demonstration of the robustness of the AI-assisted readout method. Recently, fairness-aware training strategies have been introduced to mitigate and detect bias during training²³. These approaches might further reduce the bias in the model and, especially if bias is detected in the real-world setting, can be incorporated in future iterations.

In conclusion, the AI-assisted readout method for the interpretation of SPT has demonstrated a notable accuracy, reduction in inter- and intra-observer variability and in the time required for readout. These innovations will contribute to a more efficient and accurate approach to the diagnosis of allergy with benefits for both routine clinical practice and research.

The enhanced precision and expedited results facilitated by the SPAT device and AI technology, not only streamline workflow but also potentially improve patient outcomes by enabling quicker and more reliable identification of allergic skin reactions. This innovative approach represents a promising step forward in the field of allergy diagnostics.

Methods

Study design

Five Belgian hospitals (UZ Leuven, UCL Saint-Luc, CHU Liege, ZOL Genk, GZA Sint-Augustinus) and one German hospital (MRI-TUM) participated in an international, multi-center, prospective study to determine the accuracy of an AI-assisted readout method that could detect and measure the wheals on the images of an allergy test conducted by the SPAT device. The study was subdivided into three phases: 1. training of the AI algorithm and software development, 2. clinical validation of the AI algorithm (AI versus physician), and 3. operationalization of the software and clinical validation of the AI-assisted readout method (AI + physician versus physician alone). 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).

Study population

Adults (aged 18–87 years) consulting the outpatient clinic for an allergy diagnosis to aeroallergens were recruited between September 2023 and September 2024. All study participants provided written informed consent before inclusion in the study.

Applied exclusion criteria were in line with the commonly used exclusion criteria for conventional SPT⁴ and detailed in the appendix (pp 1).

SPAT procedure

Automated skin prick tests were carried out by the SPAT medical device (Hippo-Dx, Aarschot, Belgium), operated by a trained nurse or study co-worker of the ear-nose-throat (ENT) department in the participating hospitals. A new lancet (Yilmaz Medikal, Gaziantep, Turkey) was used per individual prick. A panel of common aeroallergens was tested along with histamine and glycerol-saline as positive and negative controls, respectively. The allergen panel was not standardized across the participating hospitals because the type of allergen was not included as a variable in the AI software. An overview of the allergen panel used in the different hospitals is listed in the appendix (Table 1; pp 6). The prick location on the forearm has previously been shown not to impact the outcome of SPAT²⁴.

Subjects were asked to hold their forearm against the armrest of the SPAT device after the nurse or study co-worker initiated the test on the touch screen, and the automated pricking procedure was started as described earlier^{12,13}. In brief, 12 individually pressure-controlled pricks were applied simultaneously on the patient's forearm, followed by a clockwise rotation. Fifteen minutes after the prick process, the patients were asked to position the arm against the armrest of the device in such a way that the prick locations were aligned with the skin reactions on the forearm. Then, 35 digital images of the tested forearm were made by the SPAT device. These recordings were processed through an existing AI algorithm that generates a composite image.

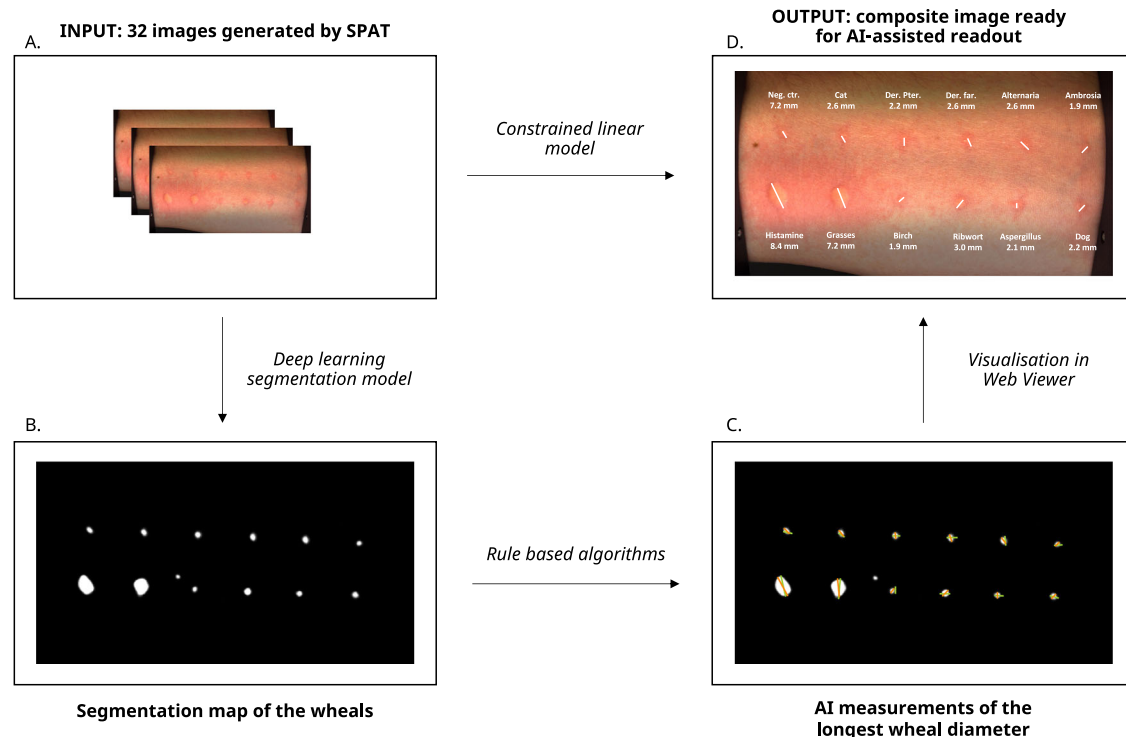


Fig. 5 | Main processing steps of artificial intelligence (AI)-assisted readout software. Green line: measurement by the treating physician, orange line: AI measurement.

The longest wheal diameters were measured by use of the digitally processed images in the SPAT Web Viewer on a local network desktop by a staff member of each of the participating ENT departments (Fig. 1; appendix pp 8). The SPAT result is considered positive if the longest wheal diameter ≥ 4.5 mm, as previously published^{12,25}.

Training of the AI algorithm and software development

The 35 captured test images consisted of a subset of 32 images with its respective lighting source, and 3 control images without lighting or with full lighting. As input for the AI algorithm, we used the subset of 32 images to determine the presence of wheals (Fig. 5A). All images acquired capture the entire section of interest of the forearm, i.e. 12 wheals were visible on each image.

The AI-assisted readout software consisted of two consecutive parts. First, the 32 images are fed into a deep learning segmentation model (details in appendix pp 1), that predicted as output a segmentation map of the forearm, where wheal regions are segmented (Fig. 5B).

The deep learning segmentation model was trained on 7812 wheals of 651 patients (75% of patients randomly assigned from the 868 included patients). To this end, each wheal was labeled manually by drawing the circumference line on the images. This task was executed by three labelers properly trained. The same images of 301 patients were labeled by at least 2 labelers, with a concordance between these labelers of 87%. The labeling itself was performed by use of a custom-built web application, where the labelers scrolled through all 35 different images and labeled wheal regions using polygons, i.e., clicking along the edge of a wheal defines the points of a closed polygon shape.

Subsequently, this segmentation map was interpreted by traditional white box algorithms, specifically designed for this purpose. In brief, three consecutive steps were followed: identification of distinct connected regions, matching the detected regions with the known prick locations of the device, and measurement of relevant wheal sizes in each region (details in appendix pp 1). To estimate the longest wheal diameter, the Feret maximum diameter algorithm was used²⁶, which finds the maximum distance between any two points belonging to the

region (Fig. 5C). The estimates of the longest wheal diameters for each wheal were then visualized on the SPAT composite image in the Web Viewer (Fig. 5D).

Clinical validation of the AI algorithm

Images of a separate cohort of 217 patients (25% of patients randomly assigned from the 868 included patients), corresponding to 2604 wheals, were used to validate the performance of the AI algorithm. The AI measurements of the longest wheal diameter were compared with the measurements by the treating physician. In order to gain further insights into the performance of the AI algorithm, a post-hoc analysis of all false positive (AI measurement ≥ 4.5 mm–physician measurement < 4.5 mm) and false negative (AI measurement < 4.5 mm–physician measurement ≥ 4.5 mm) SPAT results in the validation cohort was performed by one independent physician (Laura Van Gerven).

Operationalization of the software and clinical validation of the AI-assisted readout method

Following validation of the algorithm, the AI-assisted readout software was operationalized in the SPAT Web Viewer. In this phase of the study, 96 additional patients were recruited and tested using the SPAT device. Images were available for 95 patients (test cohort), and the AI measurements of the longest wheal diameter were evaluated by the treating physician. This physician was asked to adjust the longest wheal diameter measured by the AI if they disagreed with this measurement and to validate the test result afterwards. The time to perform the readout in the Web Viewer was recorded in seconds.

Inter- and intra-observer variability of the wheal measurement

Images of a random selection of 100 patients from the validation cohort were analyzed independently by three physicians (SU, EB, HO) to determine the inter-observer variability. Images of random selection of 30 patients out of the latter 100 patients were analyzed a second time by the same three physicians to determine the intra-observer variability.

Time to readout measurement

The time to perform the manual readout as well as AI-assisted readout was recorded for the measurements of the randomly selected 100 patients. The time to readout is defined by the time that is needed to perform the wheal measurements (manual readout) or to verify and adjust the wheal measurement (AI-assisted readout). The time starts when the patient file is opened in the web viewer and ends when the physician validates the test results by clicking on a dedicated tick box.

Allergy questionnaire

A questionnaire was filled out by each patient to obtain information on demographics, lifestyle, allergy, rhinitis & asthma symptoms and treatment (appendix pp 2–5). Rhinitis and asthma symptoms were scored on a 10 cm visual analog scale from 0 ('not at all') to 10 ('extremely').

Statistics

Descriptive statistics were applied to describe the patient's demographic and clinical characteristics. A Shapiro–Wilk test was performed to assess normality. Correlation was assessed by the Pearson coefficient. Accuracy of the SPAT AI-assisted readout software was assessed by analysis of specificity, sensitivity, PPV, and NPV. Repeated measures ANOVA and Sidak's multiple comparison test were performed to compare the AI-assisted readout to the manual readout. A p value ≤ 0.05 was considered significant. Data were analyzed with Graphpad Prism 10 (Boston, USA).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

A minimum dataset is made available in the online supplement in a de-identified/anonymized format. The images of the skin tests are not publicly available due to the presence of information that could compromise research participant privacy. Researchers may request access by contacting the corresponding author, with requests evaluated based on legitimate research purposes. Access requires completion of formal data use agreements with Hippocreates BV, and is limited to non-commercial academic research purposes. Requests will be answered within one month.

Code availability

The AI-assisted readout method (Hippocreates BV, Belgium) used in this study is a medical device regulated and supervised by the Federal Agency for Medicines and Health Products (FAMHP) in Belgium, and is commercially available from the manufacturer Hippo-Dx or its distributors. This system integrates many components, each of which contributes to its performance, safety, and usability by clinic staff without ML expertise in the real-world, as validated in the FAMHP pivotal trial. Source codes are made available upon request to the editors, reviewers and for research purposes upon a request made to the corresponding author (sven.seys@hippo-dx.com). Requests will be answered within one month. At this point, we are not sharing the code publicly in order not to compromise commercialization of our system.

References

- Sánchez-Borges, M. et al. The importance of allergic disease in public health: an iCAALL statement. *World Allergy Organ. J.* **11**, 8 (2018).
- Asher, M. I. et al. Worldwide time trends in the prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and eczema in childhood: ISAAC Phases One and Three repeat multicountry cross-sectional surveys. *Lancet* **368**, 733–743 (2006).
- Ebruster, H. [The prick test, a recent cutaneous test for the diagnosis of allergic disorders]. *Wien. Klin. Wochenschr.* **71**, 551–554 (1959).
- Heinzerling, L. et al. The skin prick test – European standards. *Clin. Transl. Allergy* **3**, 3 (2013).
- Bernstein, I. L. et al. Allergy diagnostic testing: an updated practice parameter. *Ann. Allergy Asthma Immunol.* **100**, S1–148 (2008).
- Gureczny, T. et al. Allergy screening with extract-based skin prick tests demonstrates higher sensitivity over in vitro molecular allergy testing. *Clin. Transl. Allergy* **13**, e12220 (2023).
- Carr, W. W. et al. Comparison of test devices for skin prick testing. *J. Allergy Clin. Immunol.* **116**, 341–346 (2005).
- McCann, W. A. & Ownby, D. R. The reproducibility of the allergy skin test scoring and interpretation by board-certified/board-eligible allergists. *Ann. Allergy Asthma Immunol.* **89**, 368–371 (2002).
- Fatfeh, S., Rekkerth, D. J. & Hadley, J. A. Skin prick/puncture testing in North America: a call for standards and consistency. *Allergy Asthma Clin. Immunol.* **10**, 44 (2014).
- Klimek, L. et al. Physician's perspectives on skin prick testing and allergy diagnostics in Germany. *Allergo J. Int.* **33**, 159–164 (2024).
- Klimek, L. et al. Bedeutung von Haut-Pricktests und Allergiediagnostik in Deutschland aus Sicht der Allergologen. *Allergo J.* **33**, 32–37 (2024).
- Gorris, S. et al. Reduced intra-subject variability of an automated skin prick test device compared to a manual test. *Allergy* **78**, 1366–1368 (2023).
- Seys, S. F. et al. Skin prick automated test device offers more reliable allergy test results compared to a manual skin prick test. *Rhinology* **62**, 216–222 (2024).
- Pijnenborg, H., Nilsson, L. & Dreborg, S. Estimation of skin prick test reactions with a scanning program. *Allergy* **51**, 782–788 (1996).
- Wöhrl, S., Vigl, K., Binder, M., Stingl, G. & Prinz, M. Automated measurement of skin prick tests: an advance towards exact calculation of wheal size. *Exp. Dermatol.* **15**, 119–124 (2006).
- Pineda, J. et al. Robust automated reading of the skin prick test via 3D imaging and parametric surface fitting. *PLoS One* **14**, e0223623 (2019).
- Morales-Palacios, M. P. et al. Reliability of a novel electro-medical device for wheal size measurement in allergy skin testing: an exploratory clinical trial. *Allergy* **78**, 299–301 (2023).
- Neumann, E. et al. Thermography based skin allergic reaction recognition by convolutional neural networks. *Sci. Rep.* **12**, 2648 (2022).
- Morales-Palacios, M. P. et al. Evaluation of a novel automated allergy intradermal skin test reader: a diagnostic accuracy study. *Clin. Exp. Allergy* **54**, 1006–1009 (2024).
- Gomes, R. H. M., Perger, E. L. P., Vasques, L. H., Gagete, E. & Simões, R. P. Deep learning method applied to autonomous image diagnosis for prick test. *Life* **14**, 1256 (2024).
- Lee, Y. H. et al. Allergy wheal and erythema segmentation using attention U-Net. *J. Imaging Inform. Med.* **38**, 467–475 (2025).
- Kugelman, J. et al. A comparison of deep learning U-Net architectures for posterior segment OCT retinal layer segmentation. *Sci. Rep.* **12**, 14888 (2022).
- Li, X., Cui, Z., Wu, Y., Gu, L. & Harada, T. Estimating and improving fairness with adversarial learning. Preprint at <https://doi.org/10.48550/arXiv.2103.04243> (2021).
- Seys, S. F. et al. Evaluation of skin prick location on the forearm using a novel skin prick automated test device. *Front. Allergy* **4**, 1289031 (2023).
- Seys, S. 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*. <https://doi.org/10.1111/all.16645> (2025). Epub ahead of print.

26. Pabst, W. & Gregorova, E. Characterization of particles and particle systems. https://old.vscht.cz/sil/keramika/Characterization_of_particles/CPPS%20_English%20version_.pdf (ICT Prague, 2007).

Acknowledgements

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

Author contributions

V.H., A.C., G.D.G., W.L., A.L.P., E.B., C.D., P.W.H., C.H., C.J., M.J., H.O., S.T., S.U., A.Z. and L.V.G. were responsible for recruitment. S.F.S., D.L., R.D., E.B., H.O., S.U., K.R. and L.V.G. were responsible for data analysis. R.D. was responsible for the development of the AI algorithm. The study was conceived, designed, set up, analyzed, and interpreted by S.F.S., M.J., D.L., S.G. and L.V.G. S.F.S., D.L. and L.V.G. have verified the underlying data. All authors accept responsibility for the decision to submit for publication.

Competing interests

MJT received consulting fees for statistical advice for the study. 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). ZD received speaker or consultant honoraria and/or served on advisory boards (past 36 months) at: Antabio, Arcede, Biosion, Foresee Pharmaceuticals, Galenus Health, GlaxoSmithKline, Hippo-Dx, Pleuran, QPS-Netherlands, Sanofi-Genzyme-Regeneron. During the last 3 years of her assignment as Research Director, Respiratory and Allergy at QPS-Netherlands, the company received a European grant from ERA4TB and funding from Foresee Pharmaceuticals for early clinical studies. ZD serves/ed as associate editor for Allergy (2018–2023) and Respiratory Medicine (ongoing) and acted as Chair of the Asthma Section at EAACI (2017–2019) and Expert Panel at EUFOREA (2020–2024). 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.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-64334-w>.

Correspondence and requests for materials should be addressed to Sven F. Seys.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025, corrected publication 2026

¹Hippo Dx, Aarschot, Belgium. ²Service d'Otorhinolaryngologie, Cliniques Universitaires Saint-Luc, Brussels, Belgium. ³Department of Otorhinolaryngology and Head and Neck Surgery, TUM School of Medicine and Health, TUM University Hospital Klinikum rechts der Isar, Technical University of Munich, Munich, Germany. ⁴Department of Otorhinolaryngology-Head and Neck Surgery, ZAS Sint-Augustinus, Antwerp, Belgium. ⁵Department of Otorhinolaryngology-Head and Neck Surgery, ZOL, Genk, Belgium. ⁶Service d'Otorhinolaryngologie, CHU Liège, Liège, Belgium. ⁷D2LAB, Ghent University, Ghent, Belgium. ⁸Allergy and Clinical Immunology Research Group, Department of Microbiology, Immunology & Transplantation, KU Leuven, Leuven, Belgium. ⁹Univ Groningen, Univ Med Ctr Groningen, Dept Clin Pharm & Pharmacol, Groningen, Netherlands. ¹⁰First Faculty of Medicine, Department of Respiratory Medicine; Charles University and Thomayer Hospital, Prague, Czech Republic. ¹¹Department of Otorhinolaryngology-Head and Neck Surgery, UZ Leuven, Leuven, Belgium. ¹²Laboratory of Experimental Otorhinolaryngology, Department of Neurosciences, KU Leuven, Leuven, Belgium. ¹³Zurich University of Applied Sciences, Zurich, Switzerland. ¹⁴Department of Otorhinolaryngology, AZ Herentals, Herentals, Belgium.

✉ e-mail: sven.seys@hippo-dx.com