Check for updates

Accepted Article

INTER-RATER VARIABILITY IN PATCH TEST READINGS AND FINAL INTERPRETATION USING STORE-FORWARD TELEDERMATOLOGY

Molly C. Goodier M.D. *,*†
Joel G. DeKoven M.D.‖
James S. Taylor M.D.††
Denis Sasseville M.D.‡
Joseph F. Fowler, Jr M.D.#
Anthony F. Fransway M.D.§§
Vincent A. DeLeo M.D.‖‖
James G. Marks, Jr M.D.¶¶
Kathryn A. Zug M.D.**
Sara A. Hylwa, M.D. **†*, *†**, *†*
Erin M. Warshaw MD MS **†*, *, *†**

*†Health Partners Institute Dermatology, MN
**†* University of Minnesota Department of Dermatology, Minneapolis, MN
* Minneapolis Veterans Affairs Medical Center, Minneapolis, MN
*†** Park Nicollet Contact Dermatitis Clinic, Minneapolis, MN
** Dartmouth-Hitchcock Medical Center, Lebanon, NH
#University of Louisville, Louisville, KY
†† Cleveland Clinic, Cleveland, OH
‡ Montreal General Hospital, McGill University, Montreal, Canada
§§Associates in Dermatology, Fort Myers, FL
‖‖ Keck School of Medicine, Los Angeles, CA
¶¶ Pennsylvania State University, State College, PA
‖ Sunnybrook Health Sciences Centre, University of Toronto
† Hennepin County Medical Center, Minneapolis, MN

Conflicts of Interest and Source of Funding: None declared. Dr. Sasseville receives royalties from UpToDate (Wolters Kluwer Health). Since the completion of this study, Dr. Warshaw has served as a consultant and principal investigator for Wen by Chaz Dean and a consultant for Noven Pharmaceuticals.

Disclaimer: Support through the use of resources and facilities at the Minneapolis Veterans Affairs Medical Center. The contents do not represent the views of the U.S. Department of Veterans Affairs or the United States Government. Statistical support was provided by Doug Lorenz, PhD.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/cod.13856

This article is protected by copyright. All rights reserved.

Accepted Article

Background: Data regarding teledermatology for patch testing are limited.

Objectives: Compare patch test readings and final interpretation by 2 in-person dermatologists (IPDs) with 8 teledermatologists (TDs).

Methods: Patch tested patients had photographs taken of 70 screening series allergens at 48-hour (48H) and second (2nd) readings. 8 TDs reviewed photos and graded reactions (negative, irritant, doubtful, +, ++, +++) at 48H and 2nd readings; in addition, they coded a final interpretation (allergic, indeterminant, irritant, negative) for each reaction. TDs rated overall image quality and confidence level for each patient and patch test reaction, respectively. Percentage of TD-IPD agreement based on clinical significance (success, indeterminate, and failure) were calculated. Primary outcome was agreement at 2nd reading.

Results: Data were available for 99, 101, and 66 participants at 48H, 2nd reading, and final interpretation, respectively. Pooled failure (+/++/+++ vs negative) at 2nd reading was 13.6% (range 7.9-20.4%). Pooled failure at 48H and final interpretation was 5.4% (range 2.9-6.8%) and 24.6% (range 10.2-36.8%). Confidence in readings was statistically correlated to quality of images and disagreement.

Conclusion: For patch testing, teledermatology has significant limitations including clinically significant pooled failure percentages of 13.6% for 2nd readings and 24.6% for final interpretation.

This article is protected by copyright. All rights reserved.

Accepted Article

INTRODUCTION

Teledermatology is the use of technology to provide dermatology services remotely. In live, interactive teledermatology, the teledermatologist (TD) views live video images and provides clinical recommendations in real time (i.e. with the patient). In store-forward asynchronous teledermatology, the TD typically views digital images (obtained at an earlier time point) accompanied by written history and provides written recommendations to the referring provider. There are several hybrids of these two models. Depending on the specific setting, teledermatology can provide cost savings while curtailing in-person clinic visits.[1-4]

The use of teledermatology, especially store-forward, has been studied for general dermatology.[5-11]. Data are limited on teledermatology for patch testing; nonetheless it has been implemented in some settings.[12, 13] A previous, intra-rater study by our group compared the same IP dermatologist's (IPD) assessment to photographs (at 4 and 8 weeks later) in 101 patch-tested patients; proportions of "success" (no impact on clinical management), "indeterminant" (possible impact on clinical management) and "failure" (definite clinical significance) were 54%, 40% and 6%, respectively.[14] Ivens et al evaluated the inter-rater agreement of 5 patch test experts in assessing photographs of 55 patch test reactions; pairwise agreement was 50-77% for exact match and 80-89% for the 3 groupings of positive, negative, and irritant.[15] Uter et al asked 122 conference attendees (blinded to substance, concentration and vehicle but not to presence of palpable infiltration) to grade 20 photographs of single patch tests; IPD-TD agreement was 63.5% for exact match and 87.8% for "positive" vs. "not positive."[16] After grading these initial photos, participants were asked to review the same photos but with additional context (substance, concentration and vehicle); agreement of unblinded rating with previous, blinded rating for each photo ranged from 62.3% to 91.8%. Substances with significant changes in rating of "allergic" vs "not allergic" included 4 standard allergens (paraphenylendiamine, bufexamac, potassium dichromate, and thiuram mix) as well as 3 irritants (kerosene, detergent, and disinfectant).[17] A separate study of 35 patients with 38 late reactions found that 26.3% (10/38) were missed (and there were 9 false positive reactions) when patient-submitted digital photos were substituted for a IP visit.[18]

This study expands on previous work by comparing 2 IPDs to 8 TDs to further assess agreement and inter-rater agreement.

METHODS

This study received approval from the Institutional Review Board of the Hennepin Healthcare Research Institute. We used previous study[14] photographs. Inclusion criteria included consenting patients over the age of 18 undergoing patch testing to the North American Contact Dermatitis Group (NACDG) 2015-2016 screening series[19] [70 allergens (Allergeaze, SmartPractice, Phoenix, AZ) tested on Finn chambers (SmartPractice, Phoenix, AZ)].

In-Person Encounters

Two board-certified dermatologists with patch test expertise served as IPs (EW, SH). For the 48H and 2nd readings, each patch test site was graded using standard NACDG scoring (six

This article is protected by copyright. All rights reserved.

options):[19] + (erythema, induration, +/- papules); ++ (erythema, induration, papules, vesicles); +++ (intense erythema, induration, coalescing vesicles, bullae, spreading); doubtful (faint erythema with no induration); irritant; or negative. In addition, patients seen by EW (n= 66) had a final interpretation scored (4 options: allergic, unknown, irritant, negative) for each allergen per NACDG criteria.

Photographs

After informed consent, photographs were obtained at the 48-hour (48H) and second (2nd; 96-168 hours) IP readings. Only the screening series (70 allergens)[19] test sites were photographed; patches in supplement series and/or tests with personal products were excluded. Photographs were obtained by one of two research assistants with an Apple iPad Air camera under standard settings (1936 x 2592 pixels, ISO-32, F-stop 2.4, Exposure time 1/40 sec) in 1 of 4 identical exam rooms with overhead room lighting and subject position during both readings. Photographs were taken of the entire test area as well as close-up images of approximately 20 allergens in a photographic field. Images were evaluated at time of acquisition to assess quality and ensure no personal identifying features were present. No post-process editing of the photographs was performed, and all images were stored on a secure hospital server using unique patient study ID number.

Teledermatology Encounters

Eight members of the NACDG (contact dermatitis experts at tertiary referral patch test clinics) served as TDs. TDs were asked to provide information regarding viewing conditions: laptop/desktop/tablet; natural light/overhead light/both; and windows/no windows. Additionally, TDs provided information regarding years in practice and number of patients seen per month.

TDs were sent encrypted images on a USB drive in two batches. The first batch contained 48H image sets. Scoring sheets and maps identified location and allergens (name, vehicle, concentration) but no additional information (e.g. patient demographics, clinical history, other patch test reactions, presence of induration at patch sites). Each patch test site was graded by TDs using the same standard NACDG scoring method as IPDs. For each positive patch test reaction, TDs were asked to rate their confidence for that specific patch (high/medium/low). TDs were also asked to rate the overall quality of images (high/medium/low) for each subject. TDs could enlarge images at their discretion. After completing the readings in the first batch, TDs returned all completed data sheets and USB drive via mail.

To avoid recall bias, there was a minimum of eight weeks between the first (48H) and second (2nd reading) batches. The second batch contained 2nd reading image sets. To most closely simulate IP 2nd readings, TDs completed grading of the 2nd reading images in the same manner as 48H (standard NACDG scoring, confidence level for each positive reaction, and overall image quality). After completing those assessments, the TD was able to view 48H scores for that patient. Using the prior 48H score while viewing the 2nd images/scores, the TD made a final interpretation (allergic, indeterminant, irritant, negative) for each patch.

Outcome Measures

Primary outcome was agreement of IP ("gold standard") and TD grading of patch test sites for the 2nd reading. Similar to previous methods, clinical significance categories of IP-TD agreement/disagreement were "success", "indeterminate", and "failure".[14] "Success" constituted *complete agreement* (exact match of +/++/+++, doubtful, or irritant grading), *partial agreement*

This article is protected by copyright. All rights reserved.

Accepted Article

(combination of +/++/+++ but not an exact match), and *negative/irritant disagreement* (one evaluator coded as negative and the other coded irritant). "Indeterminate" was comprised of four pairings: *positive/irritant disagreement* (+/++/++ rating vs. irritant*); irritant/doubtful disagreement* (irritant rating vs. doubtful); *negative/doubtful disagreement* (negative rating vs. doubtful); and *positive/doubtful disagreement* (+/++/+++ rating vs. doubtful). "Failure" was defined as complete IP-TD disagreement (+/++/+++ rating vs. negative).

Secondary outcomes included 1) agreement of 48H patch test site grading using the same 3 clinical significance categories as the 2nd reading; 2) agreement of final interpretation for the subset for which IP data was available (n=66) [*success* – exact agreement of allergic or not allergic/irritant; *indeterminant* – allergic/not allergic/irritant with unknown/irritant; *failure* – complete mismatch of allergic with not allergic); 3) correlation between TD confidence in readings (low, moderate, high) and disagreement (defined as any rating other than complete or partial agreement) 4) image quality (high, moderate, low) and disagreement; 5) identification of specific allergens with high proportion (>40%) of disagreement (analysis limited to those with confidence assessments); and 5) association between viewing characteristics – lighting or viewing device – and "success".

Data Analysis

Data were manually entered and stored in Excel (Excel, 2010, Microsoft Corporation, Redmond, WA). Analysis was performed using R software environment (R: A language and environment for statistical computing; R Core Team, R Foundation for Statistical Computing, Vienna, Austria). Frequency of IP-TD agreement/disagreement for the aforementioned categories were calculated for each TD as well as pooled: 48H, 2nd, and final interpretation. Spearman correlation was performed to assess the relationship between IP-TD disagreement and TD confidence in readings or perceived image quality for individual allergens. Graphs were created with IBM SPSS Statistics software (Version 23.0, IBM Corporation, Armonk, NY).

RESULTS

Subject Characteristics

Characteristics of 101 study participants have been reported.[14] Briefly, 72% were female and average age was 50 years. Most (n=94) had NACDG screening allergens applied on the back; other anatomical sites were used for the remainder.

Primary Outcome: IP-TD Agreement for Patch Test Grading at Second Reading

Of 56,536 IPD-TD pairings at the 2nd reading, 53,171 were negative/negative and were excluded from further analysis (Figure 1). Of the 3,365 IPD-TD pairings in which one or both was graded other than "negative," pooled TD percentages based on clinical significance were as follows: success 43.5% (range 33.4-49.4%), indeterminant: 43.0% (range 39.3%-46.6%) and failure:13.6% (range 7.9-20.4%) (Table 1). For the 450 failures, 72.9% were due to TD undercall (IPD +/++/+++ vs TD negative).

Secondary Outcome: IP-TD Agreement for Patch Test Grading at 48 Hour Reading

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

Of 55,416 IPD-TD pairings at 48H, 50,322 were negative/negative and were excluded from further analysis (Figure 1). Of the 5,084 IPD-TD pairings in which one or both was graded other than "negative," pooled percentages of TDs based on clinical significance were as follows: success 27.7% (range 15.5-35.5%), indeterminant 66.8% (range 58.3-78.3%) and failure 5.4% (range 2.9-6.8%) (Table 2). For the 277 failures, 62.8% were due to TD undercall (IPD +/++/+++ vs TD negative).

Secondary Outcome: IP-TD Agreement for Final Interpretation

Of 36,936 IPD-TD pairings for final interpretation, 34,130 were negative/negative and were excluded from further analysis (Figure 1). Of the 2,806 IPD-TD pairings in which one or both was graded other than "negative," pooled percentages of TDs based on clinical significance was as follows: success 59.6%, indeterminant 15.9%, and failure: 24.6% (Table 3). Of the 817 failures, 73.4% were due to TD undercalls (IP allergic vs TD not allergic).

Association of Patch Test Grading Success with Teledermatologists' Experience

TDs had a cumulative average of 32 years of patch testing experience (range 22-46 years). Average number of patch-tested patients per month was 21 (range 6-45). No association was noted between success for 48H and 2nd readings and TD years of experience (Figure 2) or monthly patch test patient volume (Figure 3).

Correlation of Patch Test Grading Success with Teledermatologist Confidence

On average, TDs rated their confidence as low, moderate, or high for 29.7%, 42.3% and 28.1% of patch test readings at 48H and 26.4%, 40.2%, and 33.4% at the 2nd reading, respectively. There was a positive correlation between disagreement (defined as any category other than complete or partial agreement) and low teledermatologist confidence at the 48H and 2nd readings (Figure 4). Spearman correlation of disagreement and low confidence was found to be 0.90 [0.84, 0.94] at 48H and 0.87 [0.80, 0.92] for 2nd reading. A mixed effects logistic regression showed that disagreement significantly differed based on the confidence ratings for both 48H and 2nd readings ($p < 0.0001$).

Image Quality

TD average ratings for image quality as low, moderate or high were 18.1%, 50.0% and 31.9% of patients at 48H and 18.4%, 57.0%, and 24.6% at the 2nd reading, respectively. A positive association was found between disagreement and low image quality at both readings (Figure 5). Spearman correlation of disagreement and moderate/low image quality versus only low image quality was 0.92 [0.87, 0.95] and 0.85 [0.76, 0.90] respectively at 48H. A similar, albeit weaker, correlation was noted at 2nd reading with Spearman correlation of 0.80 [0.69, 0.87] for low/moderate image quality and disagreement versus 0.77 [0.65, 0.85] for low image quality and disagreement.

Viewing Conditions

Most TDs used a desktop (48H 37.5%; 2nd 62.5%) or a laptop (48H 37.5%; 2nd 37.5%) computer. There was no difference in percentage of success for laptop versus desktop for 48H (24% [413/1741] vs 27% [500/1845]) or 2nd reading (45% [968/2142] vs. 41% [503/1223]). There were no statistical associations noted with lighting type (data not shown).

This article is protected by copyright. All rights reserved.

Accepted Article

Specific Allergens

Several allergens were associated with high proportion (>40%) of disagreement (defined as any category other than complete or partial agreement. For 48H, this included carba mix 3% petrolatum(77.8%), iodopropynyl butylcarbamate 0.5% petrolatum(76.2%), oleamidopropyl dimethylamine 0.1% aqueous (74.4%), and cocamidopropyl betaine 1.0% aqueous (67.2%). The bulk of disagreement involved negative versus doubtful reads, which accounted for 44.5%-62.5% of instances for the aforementioned allergens. TD undercalls occurred more frequently than overcalls (data not shown).

At the 2nd reading, allergens associated with high proportion (>40%) of disagreement included iodopropynyl butylcarbamate 0.5% petrolatum (82.1%), methyldibromoglutaronitrile/phenoxyethanol 2.0% petrolatum (78.5%), cinnamic aldehyde 1% petrolatum (73.9%), oleamidopropyl dimethylamine 0.1% aqueous (73.1%), and propylene glycol 100% (69.0%). TD undercalls were more prevalent than overcalls (data not shown).

DISCUSSION:

This study evaluated the agreement of two IPDs and eight TDs for 70 patch test sites in 101 patients. The primary outcome, agreement at the 2nd reading, was considered a failure (definite clinical significance) in 13.4% of patches. Percentage was lower at 48H (5.4%) but higher for final interpretation (29.1%). Disagreement correlated with both lower TD confidence and perceived quality of images, but not TD experience or patient volume.

Second Readings and Final Interpretation

Due to the delayed nature of Type IV hypersensitivity reactions, the 2nd patch test reading is considered essential for accuracy. A final interpretation takes into account the decrescendo/crescendo pattern over time, the nature of the allergen/concentration/vehicle, reactions to other allergens, and is critical for counseling (e.g. allergens with doubtful morphology interpreted as irritant reactions are not typically recommended by the dermatologist for avoidance during patient counseling). For example, a doubtful reaction to formaldehyde 2% aq would likely be interpreted as “irritant/negative” if formaldehyde 1% aq and all other formaldehyde releasing allergens were negative. Conversely the same doubtful reaction would likely be interpreted as “allergic” if multiple other formaldehyde-related allergens were positive. There are other circumstances (not available to the study TDs but available in a clinical setting) such as history of prior reactions, positive tests to personal products containing the allergen, and/or results of duplicate testing.

While many patch test groups classify “positive” and “negative” on morphology alone (+/++/+++, excluding doubtful reactions), this may be changing; Uter et al advocated for the addition of a final interpretation after noting the change in ratings when blinded raters were subsequently unblinded to information on substance/concentration/vehicle.[17] While most epidemiologic patch test data is focused on second reading scores, we strongly advocate for the importance of the additional steps of 1) determining final interpretation and 2) establishing clinical relevance. These are standard NACDG procedures.[19] While the current study did not evaluate agreement on clinical relevance, both the 2nd readings and final interpretation had surprisingly high failure percentages (13.4% and 29.1%, respectively).

This article is protected by copyright. All rights reserved.

Accepted Article

Comparisons to other studies is difficult due to different study designs. Our previous study (using the same image set) assessed intra-rater agreement of the same IPD performing teledermatology readings; failure for the 2nd reading in that setting was 6%; final interpretation was not assessed.[14] Ivens et al evaluated the inter-rater agreement of 5 patch test experts in assessing photographs of 55 patch test reactions; pairwise disagreement was 11-20% for the 3 groupings of positive, negative, and irritant.[15] That outcome is roughly equivalent to our final interpretation outcome; our percentage was higher at 29.1%. In Uter et al[16], the 122 conference attendees included clinicians and researchers; failure for 2nd reading in that setting was 12.2%. This is similar to our 13.4% failure for 2nd reading. In the study of 35 patients with 38 late reactions, 26.3% (10/38) were missed (with an additional 9 false positive reactions) for a potential failure of 40.4% (19/47); their higher percentage may be due to patient-submitted digital photos.[18]

Inter-rater vs. Intra-rater Agreement

Despite efforts to standardize patch test grading[20-22], significant inter-rater variability exists for IPDs, without the added complication of digital images.[20, 23] Svedman et al evaluated the concordance between 11 IPDs in six patients before and after an informational session regarding patch test reading.[20] Prior to any education, agreement between dermatologist and designated gold standard, for allergic reactions was 89.8% while irritant and doubtful reactions showed lower agreement at 21% and 19% respectively. After an informational lecture, allergic reaction agreement was 88.2%, irritant 46%, and doubtful 44%. Inherent subjective differences within grading of patch test reactions can contribute to variation among raters, but the scoring system itself can also play an integral role with respect to rater discordance.[15, 23, 24] Study performed by Ivens et al found that a simplified scoring system consisting of positive, negative, and irritant labels – in place of the International Contact Dermatitis Research Group (ICDRG) – lead to a higher degree of agreement amongst TDs.

Limitations of Teledermatology

Failure in the current study could be due to the lack of tactile information and subtle examination of patch test sites, which have been cited as factors in other general teledermatology studies.[11, 25] Image quality is also important,[9, 26, 27 28] and not surprisingly, in the current study disagreement correlated with TD ratings of image quality. Image quality has also been correlated with lower TD confidence in general teledermatology studies.[29, 30] Here, we also found that low (and moderate) TD confidence was correlated with disagreement. Previous general teledermatology studies have found that confidence in photographic assessment of dermatologic conditions can improve with practice.[25] We found that there was no increase in agreement over the course of the study (data not shown). It is possible that given more time and practice, accuracy of TDs evaluating patch test readings could improve. However, given that all TDs were patch test experts, we feel that this is not a major explanation for failure in this study.

In this study, TD years of experience or monthly patch test volume did not correlate with disagreement. A prior study conducted by Bruze et al found similar findings.[23] That study involved five IPDs grading patch test readings on forty patients to serial dilutions of previously-diagnosed known allergens. Evaluators noted morphological features of each reaction – erythema, infiltration, papules, vesicles – to which a reaction rating was assigned based on either a European or American standard grading system. The dermatologist with the least number of years of practice did as well as the average of peers. Conversely, the study by Uter et al found a

This article is protected by copyright. All rights reserved.

Accepted Article

correlation between misclassification of irritant reactions and experience.[16] Importantly, the latter study incorporated TDs from a variety of professional backgrounds whereas our TDs were members of the NACDG, recognized as patch test experts.

Directional Discordance

In our previous study assessing intra-rater agreement (same IPD serving as TD), we found more TD undercalls in the indeterminate/failure groups (48H-53.6% and 2nd reading-63.4%).[14] Here we also found a higher proportion of TD undercalls (48H failure: 62.8% and 2nd reading-72.9%). This is likely due to the lack of tactile sense and positional clinical exam lights as well as potential subtle color differences in the teledermatology setting.

Allergens

Several allergens noteworthy for causing doubtful and irritant patch test reactions were associated with high frequency (>40%) of disagreement (defined as any category other than complete or partial disagreement). These included carba mix,[31] oleamidopropyl dimethylamine,[32] iodopropynyl butylcarbamate,[31] and cocamidopropyl betaine, [32] methyldibromoglutaronitrile/phenoxyethanol[33], cinnamic aldehyde,[34] and propylene glycol.[35] All represent "problematic" patch test preparations based on the metrics of positivity ratio (a measure of weak positive reactions)[36] and reaction index (a measure of irritant and doubtful reactions)[37] except cinnamic aldehyde (which is considered "indeterminant").[38] These findings are not surprising; converting a 3-dimensional representation to a 2-dimensional image would not be expected to eliminate the "problematic" issues with these patch test preparations and may compound issues. Knowledge of the specific substance, concentration and vehicle is important in interpreting these reactions; in the study by Uter et al, when substance, concentration and vehicle was revealed to TDs, significant changes in rating of "allergic" vs "not allergic" included 4 standard allergens (paraphenylendiamine, bufexamac, potassium dichromate, and thiuram mix) as well as 3 irritants (kerosene, detergent, and disinfectant).[17]

LIMITATIONS

This study assessed only 70 allergens. Most referral contact dermatitis clinics patch test to more than a screening series including specialty allergens and/or personal/work products.[39, 40] Given a likelihood of greater failure with more patch test assessments, especially those not in "blocks" of 10 patches, failure percentages in patients undergoing comprehensive patch testing may be significantly higher. TDs also commented that the yellow highlighters used to demarcate the allergen "grid" was easily visible in the 48H photographs but not in the 2nd reading images; purple surgical markers for the outer border of the set of 10 patches, however, was clearly visible in both sets of images. Nonetheless, this may have increased frame shift reading errors. TD fatigue in reading multiple sets of patch test sites in a research setting may play a role in errors. Multiple advances in digital imaging, including 3-dimensional images,[41] may lead to better outcomes, however, many teledermatology settings employ images obtained by the patient which could lead to poorer outcomes. Due to dermatitis involving standard patch test sites (upper back), five patients had patches placed on the legs and two patients had patches placed on both the back and arms. Subjectivity in grading of confidence and image quality is a limitation. Sample size of 101 patients for patch test readings and 66 patients for final interpretation outcomes is also a

This article is protected by copyright. All rights reserved.

limitation. Other aspects of teledermatology including patient satisfaction, patient preference, and cost (patient as well as clinic) were not assessed.

CONCLUSIONS

Comparison of patch test readings by 2 IPDs and 8 TDs for 70 allergens in 101 patients found clinically significant discrepancies (failure) in 5.4% and 13.4% of 48H and 2nd readings, respectively. TD undercalling of reactions was prevalent. Failure for final interpretation, the outcome with the greatest potential clinical impact, was 29.1%. Contact dermatitis experts should be aware of the limitations of teledermatology for reading patch test reactions.

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

REFERENCES

1. M GB, Santos A, Cordioli E. Benefits of Teledermatology for Geriatric Patients: Population-Based Cross-Sectional Study. *J Med Internet Res*. Apr 21 2020;22(4):e16700. doi:10.2196/16700
2. Yang X, Barbieri JS, Kovarik CL. Cost Analysis of a Store and Forward Teledermatology Consult System in Philadelphia. *Journal of the American Academy of Dermatology*. 2018;doi:10.1016/j.jaad.2018.09.036
3. Snoswell C, Finnane A, Janda M, Soyer HP, Whitty JA. Cost-effectiveness of Store-and-Forward Teledermatology: A Systematic Review. *JAMA Dermatol*. Jun 1 2016;152(6):702-8. doi:10.1001/jamadermatol.2016.0525
4. Armstrong AW, Dorer DJ, Lugn NE, Kvedar JC. Economic evaluation of interactive teledermatology compared with conventional care. *Telemed J E Health*. Apr 2007;13(2):91-9. doi:10.1089/tmj.2006.0035
5. Warshaw EM, Hillman YJ, Greer NL, et al. Teledermatology for diagnosis and management of skin conditions: A systematic review. *Journal of the American Academy of Dermatology*. 2011;64(4):759-772.e21. doi:10.1016/j.jaad.2010.08.026
6. Moreno-Ramírez D, Raya-Maldonado J, Morales-Conde M, et al. Increasing Frequency of Seborrheic Keratosis Diagnoses as a Favorable Consequence of Teledermatology-Based Skin Cancer Screening: A Cross-sectional Study of 34,553 Patients. *Am J Clin Dermatol*. Oct 2017;18(5):681-685. doi:10.1007/s40257-017-0283-z
7. Finnane A, Dallest K, Janda M, Soyer HP. Teledermatology for the Diagnosis and Management of Skin Cancer: A Systematic Review. *JAMA Dermatology*. 2017;153(3):319-327. doi:10.1001/jamadermatol.2016.4361
8. Chuchu N, Dinnes J, Takwoingi Y, et al. Teledermatology for diagnosing skin cancer in adults. *Cochrane Database Syst Rev*. Dec 4 2018;12(12):Cd013193. doi:10.1002/14651858.cd013193
9. Gemelas J, Capulong D, Lau C, Mata-Diaz S, Raugi GJ. Positive Predictive Value of Melanoma Diagnosis in Store-and-Forward Teledermatology. *Telemed J E Health*. Aug 2019;25(8):701-707. doi:10.1089/tmj.2018.0056
10. Wang M, Gendreau JL, Gemelas J, et al. Diagnosis and Management of Malignant Melanoma in Store-and-Forward Teledermatology. *Telemed J E Health*. Nov 2017;23(11):877-880. doi:10.1089/tmj.2017.0009
11. Du Moulin MF, Bullens-Goessens YI, Henquet CJ, et al. The reliability of diagnosis using store-and-forward teledermatology. *J Telemed Telecare*. 2003;9(5):249-52. doi:10.1258/135763303769211247
12. Cheng HS. Patch Testing Interrupted: Virtual Patch Test Readings During the COVID-19 Pandemic. *Dermatitis*. Jul/Aug 2020;31(4):e35-e36. doi:10.1097/der.0000000000000628
13. Carrascosa JM, Pastor-Nieto MA, Ruiz-González I, et al. Patch Testing During the COVID-19 Pandemic: Recommendations of the AEDV's Spanish Contact Dermatitis and Skin Allergy Research Group (GEIDAC). *Actas Dermosifiliogr*. Oct 2020;111(8):650-654. Recomendaciones del Grupo Español de Investigación en Dermatitis de Contacto y Alergia Cutánea (GEIDAC) de la AEDV en relación con la realización de pruebas epicutáneas durante la pandemia por SARS-CoV-2 (COVID-19). doi:10.1016/j.ad.2020.06.003
14. Grey RK, Hagen LS, Hylwa AS, Warshaw ME. Utility of Store and Forward Teledermatology for Skin Patch Test Readings. *Dermatitis*. 2017;28(2):152-161. doi:10.1097/DER.0000000000000258
15. Ivens U, Serup J, O'Goshi K. Allergy patch test reading from photographic images: disagreement on ICDRG grading but agreement on simplified tripartite reading. *Skin Res Technol*. Feb 2007;13(1):110-3. doi:10.1111/j.1600-0846.2007.00232.x

This article is protected by copyright. All rights reserved.

Accepted Article

16. Uter W, Becker D, Schnuch A, Gefeller O, Frosch PJ. The validity of rating patch test reactions based on digital images. *Contact Dermatitis*. 2007;57(5):337-342. doi:10.1111/j.1600-0536.2007.01236.x
17. Uter W, Frosch PJ, Becker D, Schnuch A, Pfahlberg A, Gefeller O. The importance of context information in the diagnostic rating of digital images of patch test reactions. *Br J Dermatol*. Sep 2009;161(3):554-9. doi:10.1111/j.1365-2133.2009.09167.x
18. Yüksel YT, Sedeh FB, Brans R, et al. Value of photo assessment in late patch-test readings-A multicenter study from six European patch-test clinics. *Contact Dermatitis*. Nov 5 2020;doi:10.1111/cod.13736
19. DeKoven JG, Warshaw EM, Zug KA, et al. North American Contact Dermatitis Group Patch Test Results: 2015-2016. *Dermatitis*. Nov/Dec 2018;29(6):297-309. doi:10.1097/der.0000000000000417
20. Svedman C, Isaksson M, Björk J, Mowitz M, Bruze M. 'Calibration' of our patch test reading technique is necessary. *Contact Dermatitis*. Apr 2012;66(4):180-7. doi:10.1111/j.1600-0536.2011.02044.x
21. Farage MA, Maibach HI, Andersen KE, et al. Historical perspective on the use of visual grading scales in evaluating skin irritation and sensitization. *Contact Dermatitis*. Aug 2011;65(2):65-75. doi:10.1111/j.1600-0536.2011.01912.x
22. Davis MD, Yiannias JA. Should macular erythema reactions be counted as positive allergic patch-test reactions? *Dermatitis*. Mar 2006;17(1):12-4. doi:10.2310/6620.2006.05020
23. Bruze M, Isaksson M, Edman B, Björkner B, Fregert S, Möller H. A study on expert reading of patch test reactions: inter-individual accordance. *Contact Dermatitis*. Jun 1995;32(6):331-7. doi:10.1111/j.1600-0536.1995.tb00621.x
24. Baumeister T, Weistenhöfer W, Drexler H, Kütting B. Spoilt for choice – evaluation of two different scoring systems for early hand eczema in teledermatological examinations. *Contact Dermatitis*. 2010;62(4):241-247. doi:10.1111/j.1600-0536.2010.01703.x
25. Giavina Bianchi M, Santos A, Cordioli E. Dermatologists' perceptions on the utility and limitations of teledermatology after examining 55,000 lesions. *J Telemed Telecare*. Aug 14 2019:1357633x19864829. doi:10.1177/1357633x19864829
26. Landow SM, Mateus A, Korgavkar K, Nightingale D, Weinstock MA. Teledermatology: key factors associated with reducing face-to-face dermatology visits. *J Am Acad Dermatol*. Sep 2014;71(3):570-6. doi:10.1016/j.jaad.2014.02.021
27. Paradela-De-La-Morena S, Fernandez-Torres R, Martínez-Gómez W, Fonseca-Capdevila E. Teledermatology: diagnostic reliability in 383 children. *Eur J Dermatol*. Nov-Dec 2015;25(6):563-9. doi:10.1684/ejd.2015.2658
28. Quigley EA, Tokay BA, Jewell ST, Marchetti MA, Halpern AC. Technology and Technique Standards for Camera-Acquired Digital Dermatologic Images: A Systematic Review. *JAMA Dermatol*. Aug 2015;151(8):883-90. doi:10.1001/jamadermatol.2015.33
29. Romero Aguilera G, Cortina de la Calle P, Vera Iglesias E, Sánchez Caminero P, García Arpa M, Garrido Martín JA. Interobserver reliability of store-and-forward teledermatology in a clinical practice setting. *Actas Dermosifiliogr*. Jul-Aug 2014;105(6):605-13. doi:10.1016/j.ad.2013.12.007
30. Krupinski EA, LeSueur B, Ellsworth L, et al. Diagnostic accuracy and image quality using a digital camera for teledermatology. *Telemed J*. Fall 1999;5(3):257-63. doi:10.1089/107830299312005
31. Warshaw EM, Raju S, DeKoven JG, et al. Positive patch test reactions to carba mix and iodopropynyl butylcarbamate: data from the North American Contact Dermatitis Group, 1998-2008. *Dermatitis*. Sep-Oct 2013;24(5):241-5. doi:10.1097/DER.0b013e3182a5a1d4
32. Fowler JF, Jr., Shaughnessy CN, Belsito DV, et al. Cutaneous Delayed-Type Hypersensitivity to Surfactants. *Dermatitis*. Nov-Dec 2015;26(6):268-70. doi:10.1097/der.0000000000000150

This article is protected by copyright. All rights reserved.

Accepted Article

33. Aakhus AE, Warshaw EM. Allergy to methyldibromoglutaronitrile/phenoxyethanol (Euxyl k 400): regulatory issues, epidemiology, clinical characteristics, and management. *Dermatitis*. May 2011;22(3):127-40.
34. Atwater A, Ward J, Liu B, et al. Fragrance and botanical related allergy and associated concomitant reactions: retrospective analysis of North American Contact Dermatitis Group Data 2007-2016. *Dermatitis*. 2020;
35. Warshaw EM, Botto NC, Maibach HI, et al. Positive patch-test reactions to propylene glycol: a retrospective cross-sectional analysis from the North American Contact Dermatitis Group, 1996 to 2006. *Dermatitis*. Jan-Feb 2009;20(1):14-20.
36. Geier J, Uter W, Lessmann H, Schnuch A. The positivity ratio--another parameter to assess the diagnostic quality of a patch test preparation. *Contact Dermatitis*. May 2003;48(5):280-2. doi:10.1034/j.1600-0536.2003.00033.x
37. Brasch J, Henseler T. The reaction index: a parameter to assess the quality of patch test preparations. *Contact Dermatitis*. Sep 1992;27(3):203-4. doi:10.1111/j.1600-0536.1992.tb05267.x
38. Warshaw EM, Nelsen DD, Sasseville D, et al. Positivity ratio and reaction index: patch-test quality-control metrics applied to the north american contact dermatitis group database. *Dermatitis*. Mar-Apr 2010;21(2):91-7.
39. Zhu TH, Suresh R, Warshaw E, et al. The Medical Necessity of Comprehensive Patch Testing. *Dermatitis*. May/Jun 2018;29(3):107-111. doi:10.1097/der.0000000000000362
40. Warshaw EM, Moore JB, Nelson D. Patch-testing practices of American Contact Dermatitis Society members: a cross-sectional survey. *Am J Contact Dermat*. Mar 2003;14(1):5-11. doi:10.2310/6620.2003.38741
41. Chen BR, Poon E, Alam M. Photography in Dermatologic Surgery: Selection of an Appropriate Lighting Solution for a Particular Clinical Application. *Dermatol Surg*. Jan 2018;44(1):106-114. doi:10.1097/dss.0000000000001303

This article is protected by copyright. All rights reserved.

Accepted Article

Table 1: Primary Outcome: Agreement between In-Person Dermatologist and Teledermatologist Grading at Second Reading*

Type of Agreement	5-8 Day Reading								
	TD1	**TD2**	**TD3**	**TD4**	**TD5**	**TD6**	**TD7**	**TD8**	**TOTAL**
	N = 396	**N = 394**	**N = 401**	**N = 412**	**N = 541**	**N = 389**	**N = 433**	**N = 399**	**N=3365**
Success (No Clinical Significance)									
Complete Agreement (exact match of +/++/+++, doubtful, irritant values)	129 (32.6%)	135 (34.3%)	74 (18.4%)	76 (18.4%)	197 (36.4%)	98 (25.2%)	166 (38.3%)	140 (35.1%)	
Partial Agreement (+/++/+++ on both evaluations but not exact match)	48 (12.1%)	51 (12.9%)	57 (14.2%)	61 (14.8%)	53 (9.8%)	52 (13.4%)	42 (9.7%)	45 (11.3%)	
Negative/Irritant Disagreement	5 (1.3%)	5 (1.3%)	3 (0.7%)	6 (1.5%)	14 (2.6%)	5 (1.3%)	6 (1.4%)	3 (0.8%)	
Total	**182 (46.0%)**	**191 (48.5%)**	**134 (33.4%)**	**143 (34.7%)**	**264 (48.8%)**	**155 (39.8%)**	**214 (49.4%)**	**188 (47.1%)**	**1471 (43.7%)**
Indeterminate (Possible Clinical Significance)									
Positive/Irritant Disagreement (+/++/+++ vs Irritant)	10 (2.5%)	6 (1.5%)	0 (0.0%)	2 (0.5%)	1 (0.2%)	3 (0.7%)	1 (0.2%)	0 (0.0%)	
Negative/Doubtful Disagreement	128 (32.3%)	123 (31.2%)	141 (35.2%)	142 (34.5%)	172 (31.8%)	130 (33.4%)	135 (31.2%)	129 (32.3%)	
Doubtful/Irritant Disagreement	4 (1.0%)	3 (0.8%)	0 (0.0%)	0 (0.0%)	4 (0.7%)	1 (0.3%)	4 (0.9%)	0 (0.0%)	
Doubtful/Positive Disagreement (+/++/+++ vs Doubtful)	28 (7.1%)	40 (10.2%)	46 (11.5%)	41 (10.0%)	53 (9.8%)	23 (5.9%)	30 (6.9%)	44 (11.0%)	
Total	**170 (42.9%)**	**172 (43.7%)**	**187 (46.6%)**	**185 (44.9%)**	**230 (42.5%)**	**157 (40.4%)**	**170 (39.3%)**	**173 (43.4%)**	**1444 (42.9%)**
Failure (Definite Clinical Significance)									
Complete Disagreement (+/++/+++ vs Negative)	**44 (11.1%)**	**31 (7.9%)**	**80 (20.0%)**	**84 (20.4%)**	**47 (8.7%)**	**77 (19.8%)**	**49 (11.3%)**	**38 (9.5%)**	**450 (13.4%)**

*Excludes negative/negative IP/TD agreement

This article is protected by copyright. All rights reserved.

Table 2: Secondary Outcome: Agreement between In-Person Dermatologist and Teledermatologist Grading at 48 Hours*

	48 Hour Reading								
Type of Agreement	**TD1 N = 607**	**TD2 N = 645**	**TD3 N = 554**	**TD4 N = 602**	**TD5 N = 847**	**TD6 N = 578**	**TD7 N = 618**	**TD8 N = 633**	**TOTAL N=5084**
"Success" (No Clinical Significance)									
Complete Agreement (exact match of +/++/+++, doubtful, irritant)	126 (20.8%)	152 (23.6%)	57 (10.3%)	58 (9.6%)	239 (28.2%)	97 (16.8%)	160 (25.9%)	174 (27.5%)	
Partial Agreement (+/++/+++ on both evaluations but not exact match)	30 (4.9%)	34 (5.3%)	25 (4.5%)	30 (4.9%)	39 (4.6%)	42 (7.3%)	26 (4.2%)	32 (5.1%)	
Negative/Irritant Disagreement	27 (4.4%)	4 (0.6%)	4 (0.7%)	46 (7.6%)	23 (2.7%)	7 (1.2%)	1 (0.2%)	5 (0.8%)	
Total	**183 (30.1%)**	**190 (29.5%)**	**86 (15.5%)**	**134 (22.3%)**	**301 (35.5%)**	**146 (25.3%)**	**187 (30.3%)**	**211 (33.3%)**	**1438 (28.3%)**
"Indeterminate" (Possible Clinical Significance)									
Positive/Irritant Disagreement (+/++/+++ vs irritant)	1 (0.2%)	1 (0.2%)	0 (0.0%)	17 (2.8%)	0 (0.0%)	1 (0.2%)	0 (0.0%)	0 (0.0%)	
Negative/Doubtful Disagreement	350 (57.7%)	388 (60.2%)	395 (71.3%)	347 (57.6%)	384 (45.3%)	329 (56.9%)	346 (56.0%)	347 (54.8%)	
Doubtful/Irritant Disagreement	14 (2.3%)	4 (0.6%)	0 (0.0%)	43 (7.1%)	19 (2.2%)	8 (1.4%)	2 (0.3%)	7 (1.1%)	
Doubtful/Positive Disagreement (+/++/+++ vs doubtful)	33 (5.4%)	43 (6.7%)	39 (7.0%)	20 (3.3%)	91 (10.7%)	58 (10.0%)	45 (7.3%)	37 (5.8%)	
Total	**398 (66.6%)**	**436 (67.6%)**	**434 (78.3%)**	**427 (70.9%)**	**494 (58.3%)**	**396 (68.5%)**	**393 (63.6%)**	**391 (61.8%)**	**3369 (66.3%)**
"Failure" (Definite Clinical Significance)									
Complete Disagreement (+/++/+++ vs Negative)	**26 (4.3%)**	**19 (2.9%)**	**34 (6.1%)**	**41 (6.8%)**	**52 (6.1%)**	**36 (6.2%)**	**38 (6.1%)**	**31 (4.9%)**	**277 (5.4%)**

* Excludes negative/negative IP/TD agreement

Accepted Article

This article is protected by copyright. All rights reserved.

Table 3: Secondary Outcome: Agreement between In-Person Dermatologist and Teledermatologist for Final Interpretation*

Final Interpretation		Instances
IP	TD	N (%)
Success		
Allergic	Allergic	924
Not Allergic/Irritant*	Not Allergic/Irritant	747
	Total	**1671 (59.6%)**
Indeterminant		
Allergic	Indeterminant	147
Not Allergic/Irritant*	Indeterminant	171
Allergic	Irritant	71
Irritant	Allergic	57
	Total	**446 (15.9%)**
Failure		
Allergic	Not Allergic	600
Not Allergic	Allergic	89
	Total	**689 (24.6%)**

*IP and TD negative/negative agreement excluded; this includes negative/irritant and irritant/negative

Accepted Article

This article is protected by copyright. All rights reserved.

TABLES/ FIGURES:

Table 1: Primary Outcome: Agreement between In-Person Dermatologist and Teledermatologist Grading at Second Reading
Table 2: Secondary Outcome: Agreement between In-Person Dermatologist and Teledermatologist Grading at 48 Hours
Table 3: Secondary Outcome: Agreement between In-Person Dermatologist and Teledermatologist for Final Interpretation

Figure 1: Data points and outcomes for first read, second read, and final interpretation.

Figure 2: IP-TD success rate as a function of TD experience

Figure 3: IP-TD success rate as a function of estimated monthly TD patch testing patient volume.

Figure 4: IP-TD Disagreement (Indeterminate or Failure) as a function of Low TD confidence with reading for all 70 allergens.

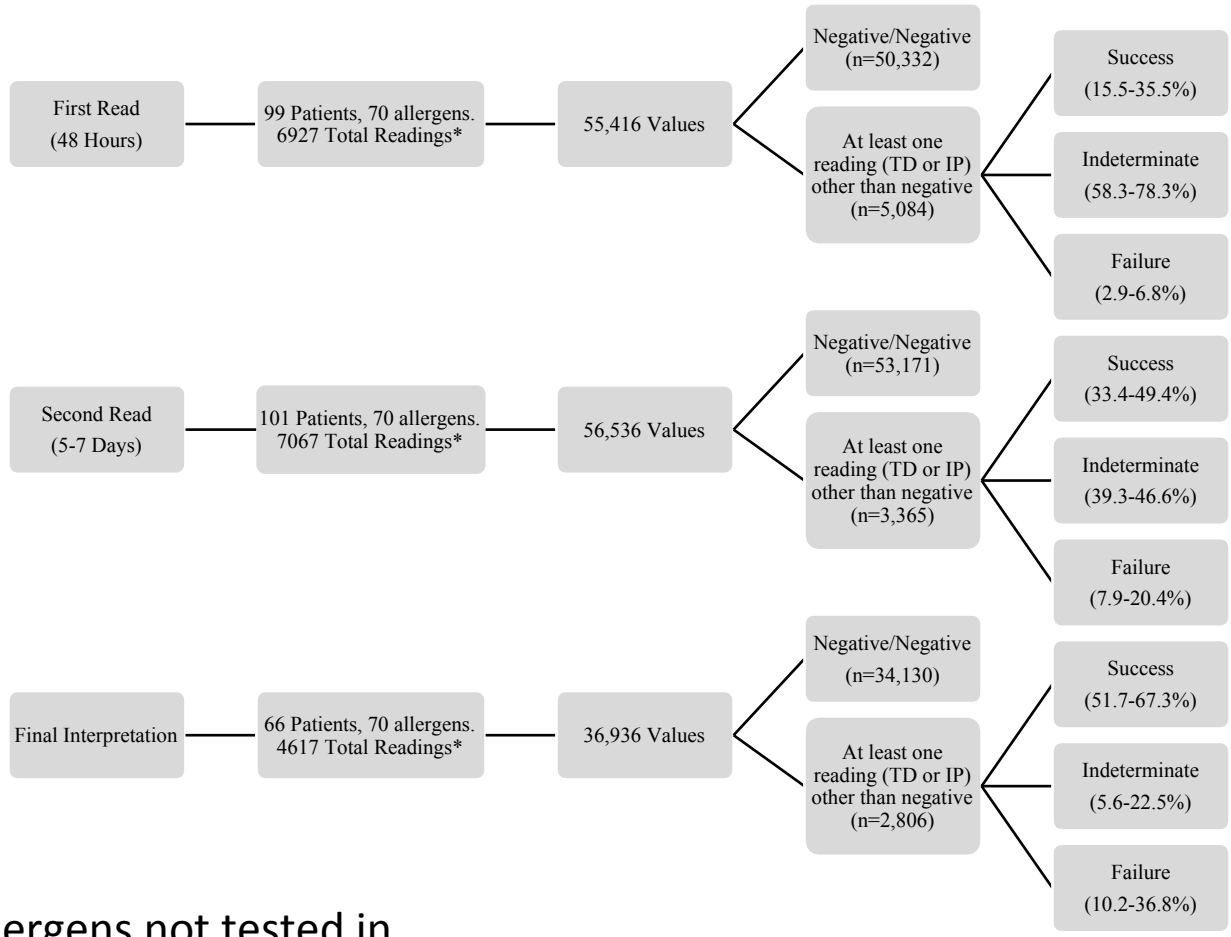
Figure 5: IP-TD Disagreement (Indeterminate or Failure) as a function of Low Image Quality with reading for all 70 allergens.

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

Figure 1: Flow Chart

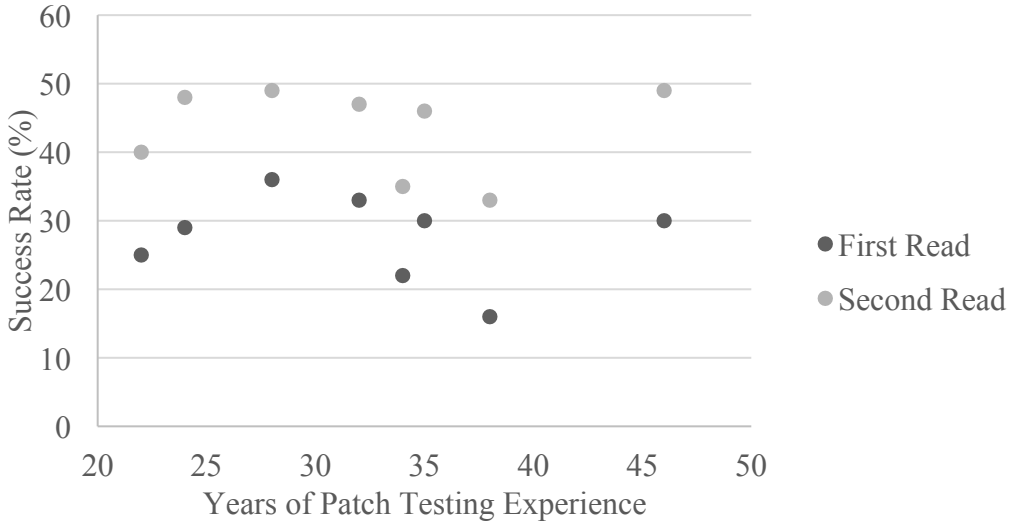


*3 allergens not tested in one patient

This article is protected by copyright. All rights reserved.

Accepted Article

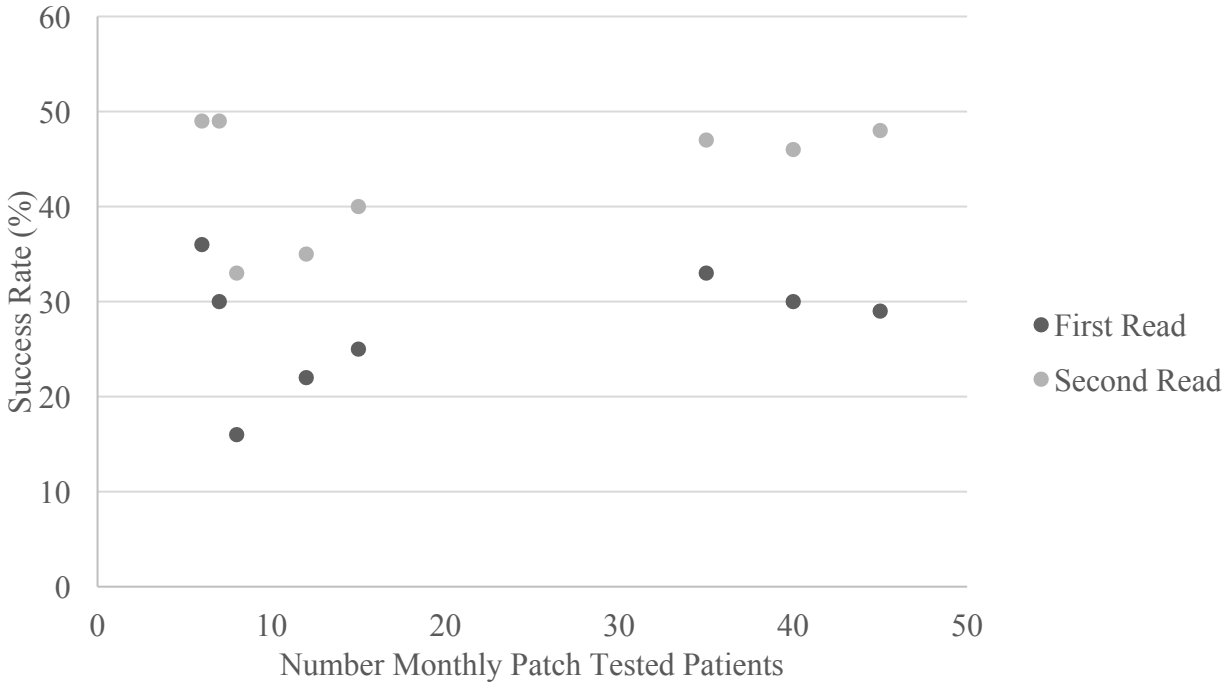
Figure 2: IPD-TD Success as Function of TD Experience



This article is protected by copyright. All rights reserved.

Accepted Article

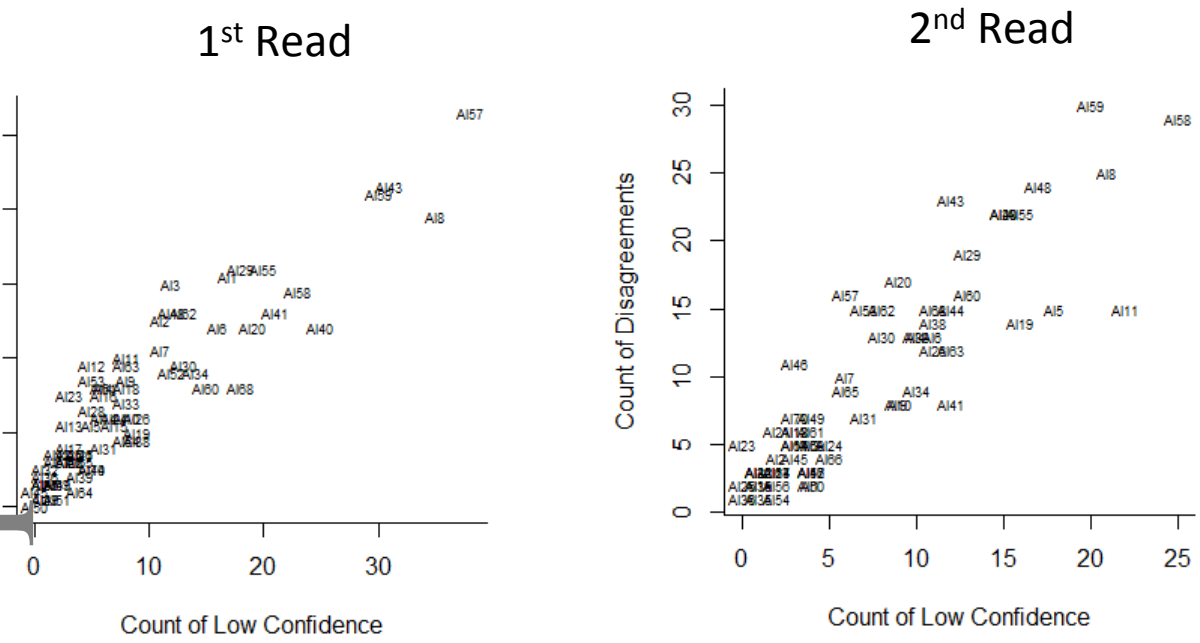
Figure 3: IPD-TD Success as Function of TD Patient Volume



This article is protected by copyright. All rights reserved.

Accepted Article

Figure 4: IPD-TD Disagreement (Indeterminate or Failure) as Function of Low TD Confidence for all 70 Allergens.*

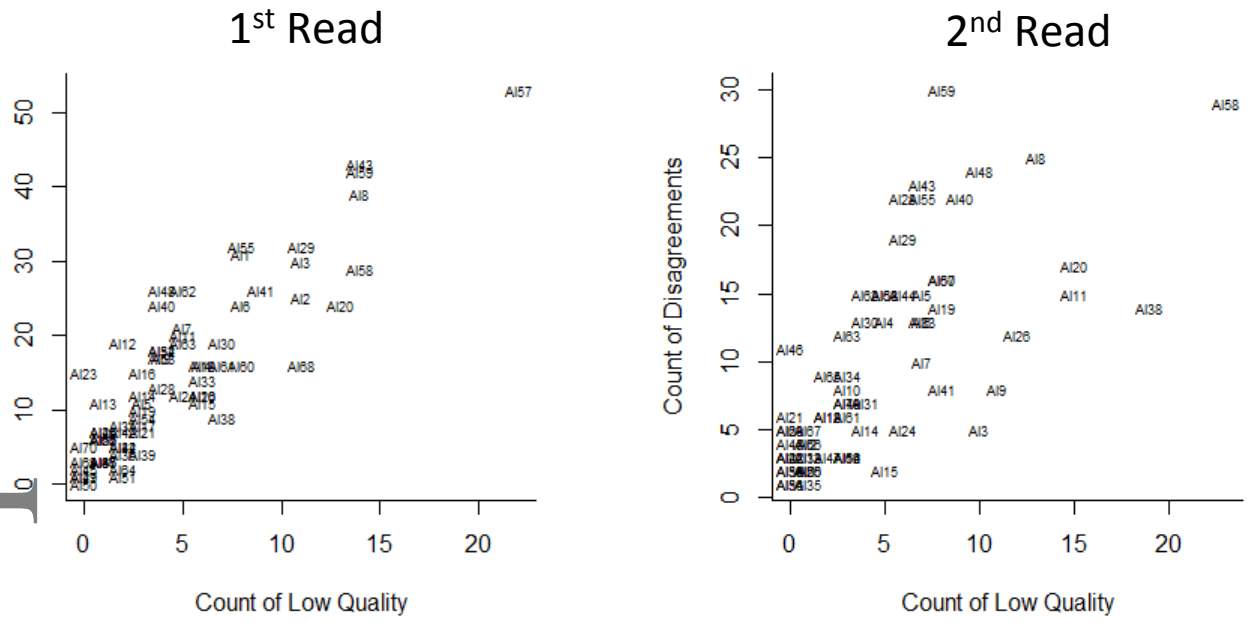


*These analyses do not reflect disagreements in which the TD rating was negative while the IP rating was something other than negative as TD confidence was not provided for negative reads.

This article is protected by copyright. All rights reserved.

Accepted Article

Figure 5: IPD-TD Disagreement (Indeterminate or Failure) as a Function of Low Image Quality for all 70 Allergens.



This article is protected by copyright. All rights reserved.