Web Application for the Automated Extraction of Diagnosis and Site From Pathology Reports for Keratinocyte Cancers

Bridie S. Thompson, PhD, MPH[1]; Sam Hardy, BCompSci[2]; Nirmala Pandeya, PhD, MSc[1,3]; Jean Claude Dusingize, MBBS, PhD[1]; Adele C. Green, MBBS, PhD[1,4]; Athon Millane, BE[3]; Daniel Bourke, BSc[5]; Ronald Grande, BIT[5]; Cameron D. Bean, GCBus[5]; Catherine M. Olsen, PhD[1,6]; and David C. Whiteman, MBBS, PhD[1,6]

abstract

PURPOSE Keratinocyte cancers are exceedingly common in high-risk populations, but accurate measures of incidence are seldom derived because the burden of manually reviewing pathology reports to extract relevant diagnostic information is excessive. Thus, we sought to develop supervised learning algorithms for classifying basal and squamous cell carcinomas and other diagnoses, as well as disease site, and incorporate these into a Web application capable of processing large numbers of pathology reports.

METHODS Participants in the QSkin study were recruited in 2011 and comprised men and women age 40-69 years at baseline (N = 43,794) who were randomly selected from a population register in Queensland, Australia. Histologic data were manually extracted from free-text pathology reports for participants with histologically confirmed keratinocyte cancers for whom a pathology report was available (n = 25,786 reports). This provided a training data set for the development of algorithms capable of deriving diagnosis and site from free-text pathology reports. We calculated agreement statistics between algorithm-derived classifications and 3 independent validation data sets of manually abstracted pathology reports.

RESULTS The agreement for classifications of basal cell carcinoma ($\kappa = 0.97$ and $\kappa = 0.96$) and squamous cell carcinoma ($\kappa = 0.93$ for both) was almost perfect in 2 validation data sets but was slightly lower for a third ($\kappa = 0.82$ and $\kappa = 0.90$, respectively). Agreement for total counts of specific diagnoses was also high ($\kappa > 0.8$). Similar levels of agreement between algorithm-derived and manually extracted data were observed for classifications of keratoacanthoma and intraepidermal carcinoma.

CONCLUSION Supervised learning methods were used to develop a Web application capable of accurately and rapidly classifying large numbers of pathology reports for keratinocyte cancers and related diagnoses. Such tools may provide the means to accurately measure subtype-specific skin cancer incidence.

JCO Clin Cancer Inform 4:711-723. © 2020 by American Society of Clinical Oncology

INTRODUCTION

Among fair-skinned populations, keratinocyte cancers are more numerous than any other cancer type.[1] Because of volume and limited resources, keratinocyte cancers are either excluded from cancer registration[1,2] or registration is limited to the first incident basal cell carcinoma (BCC) or squamous cell carcinoma (SCC) for each person.[3] Incidence estimates and population trends are typically derived from administrative data sets of treatment information that do not discriminate between subtypes.[1,4] This is a major restriction to the optimal allocation of health resources.

Pathology reports provide information on definitive diagnosis of keratinocyte cancers. Skin cancer pathology is usually reported in a free-text format, and reports often include histologic assessments for multiple lesions. Histology of skin lesions can be complex; a single lesion may show characteristics of more than one diagnosis. Interpretation and data extraction from pathology reports for skin cancers are therefore time consuming and require high-level ability to codify complex clinical information.

Automated encoding of data from free-text pathology reports has been recognized as a useful tool to identify new cancer diagnoses and for cancer registration.[5-7] A variety of machine learning methods have been used to reliably and accurately extract information from free-text pathology reports and from clinical narratives for cancers.[8] At least one study has used natural language processing to identify keratinocyte cancers from pathology reports, although that algorithm did not extract diagnosis or site details.[9]

Globally, Australia experiences the highest incidence rates of skin cancers,[10] and Queensland experiences

ASSOCIATED CONTENT

Appendix

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on May 6, 2020 and published at ascopubs.org/journal/cci on August 5, 2020: DOI https://doi.org/10.1200/CCI.19.00152

CONTEXT

Key Objective

To determine the accuracy of a supervised learning algorithm for the automated extraction of key diagnostic information about keratinocyte cancers from free-text pathology reports.

Knowledge Generated

Validated against manually extracted reports, the algorithm classified basal- and squamous cell carcinomas with nearly perfect accuracy in two dataset (kappa > 0.92) and with very high accuracy in a third dataset of complex reports (kappa 0.82-0.90).

Relevance

In the absence of population-based registration, this supervised learning algorithm can efficiently process large numbers of pathology reports, permitting users to accurately estimate subtype-specific keratinocyte cancer incidence. Such measures are essential for health care planning.

the highest rates of skin cancers within Australia.[11] The QSkin study is a large, population-based, longitudinal study of residents of Queensland, Australia. Large numbers of pathology reports for skin cancers from study participants provided an opportunity to investigate the automated extraction from free-text pathology reports. The ability to automatically process free-text pathology reports on a large scale has the potential for accurately tracking the incidence of keratinocyte skin cancers in various clinical settings, including hospitals and cancer registries. Here, we describe the development and validation of a Web application that uses supervised learning methods to automatically classify BCC, SCC, and related diagnoses from free-text pathology reports.

METHODS

We obtained pathology reports from participants of the QSkin study who had a skin cancer excised between recruitment in 2011 and June 30, 2014.[12] Details of the QSkin study have been described previously.[13] Medically trained staff reviewed each pathology report (n = 25,786 reports) and entered diagnostic information for each lesion into a database (n = 41,356 lesions). This manually extracted data set was considered the gold standard and provided the training data set to develop the supervised learning algorithm. After data cleaning and exclusion of diagnoses with insufficient examples, there were 36,281 lesions in the final data set.

Supervised machine learning algorithms are developed using training data sets (typically numbering in the thousands of independent records) that contain the variables along with the relevant outcomes. A machine learning algorithm is applied to the training data set and iteratively improved to reduce the error of outcome prediction using optimization techniques.[14] The larger the training data set, the more examples there are with which to develop the algorithm, thereby reducing the degree of error in prediction. The training data set used in this study included the free-text as well as the known outcome for a large number of pathology reports to which we applied supervised learning methods to develop an algorithm to classify diagnosis (BCC, SCC, keratoacanthoma, and intraepidermal carcinoma [IEC]), number of lesions, and site of lesions from free-text pathology reports. Diagnosis and site were modeled as separate multiclass classification problems in which a single label can be assigned to each lesion text. The training data set included all pathology reports for participants (including nonskin lesions, benign skin lesions, and melanoma).

More than a third of the pathology reports in the training data set contained descriptions and diagnoses for multiple skin lesions that had been excised at the same visit; each lesion required identification of a site and diagnosis. These were processed as multilabel classifications, where a model can return multiple labels, given a single text input. Using regex, Python, and Python dictionaries, the report text was split into lesion-specific text. The Web application first processes the free-text within a pathology report to identify and split multiple lesions, and then separate algorithms for diagnosis and site are processed on individual lesions.

Development and Internal Validation

Separate linear support vector machines (LSVMs) were developed for each classification task (ie, diagnosis and site). The data set was split into randomly shuffled train/test splits of 70/30, equating to 25,397 lesions used to derive and train the algorithm and 10,884 lesions used to test the algorithm. Term frequency-inverse document frequency matrix was created using word-based *n*-grams of length 1 or length 2 (short, 1- or 2-word phrases). Words contained within < 10% or > 90% of the reports were ignored, as were common, information-poor stop words (eg, “the,” “a,” “in”). A hyperparameter GridSearch was performed, optimizing for the best F1-Macro score (a function of both precision and recall; Table 1). Each parameter combination was evaluated using 3-fold cross-validation. The best-performing LSVM model was then evaluated against the held-out test data set. The test evaluation tested for

TABLE 1. Calculations Used in the Experiment and Validations

	Predicted Classification			
Actual Classification	Negative	Positive		
Negative	True negative	False positive	All actual negative	
Positive	False negative	True positive	All actual positive	Sensitivity positive = True positive/All actual positive
	All predicted negative	All predicted positive		
		PPV = True positive/All predicted positive		
F1 score[a] = 2 × [(PPV × sensitivity)/(PPV + sensitivity)]				

Abbreviation: PPV, positive predictive value.

[a]F1 score can range between 0 (no accuracy) and 1 (perfect accuracy).

completeness of predicting the classification (sensitivity, or recall), and the resulting evaluation retested for misclassifications (positive predictive value, or precision).

The trained models for each classification problem were then used as the basis for a Web application to upload pathology reports and analyze the free-text. The Web application can parse and analyze reports across a range of formats, commensurate with the different formats used by various laboratories. The output variables are listed in Table 2.

We developed the Web application using Python 3.6 on a machine with Ubuntu Linux (Canonical, London, United Kingdom) that has 16 cores with 8 G of memory. The following libraries were used: Pandas, sklearn, spaCy, various Python 3.6 standard libraries (including regex), and Jupyter notebooks.

External Validation

To assess the real-world performance of the algorithm beyond the historical data set used for training, we compared the classifiers' predictions on 3 independent samples of pathology reports: a random sample of 400 new pathology reports from QSkin participants; 2,345 pathology reports for QSkin participants from pathology laboratories not represented in the training data set; and 42 pathology reports from high-risk transplantation recipients enrolled in the Skin Tumors in Allograft Recipients (STAR) study.[15]

TABLE 2. Data Fields in Output Data From Pathology Classifier Web Application

Data Field	Description
UID	Study-specific person identification number
ReportText	All text as presented within the pathology report
Datecoll	Reported date that the specimen was collected for pathology
LesionID	Identification number for count of individual lesions within a report
LesionText	Text extracted across all sections of the report for the individual lesion
Site	The anatomic site of the lesion
Diagnosis	The algorithm-derived diagnosis for an individual lesion
SiteFace	Face-specific site of lesion where Site is face

The text reports were first reviewed by a medically trained staff member who entered diagnosis and site details into a database; we considered these summary measures to be the gold standard data with which to compare the algorithm-derived measures. Separately and independently of this review, the first author (B.S.T.) uploaded the same reports in their various formats (Excel, comma-separated values, PDF, and Word) into the Web application. The manually extracted data were not always entered in numerical order; therefore, we could not match on lesion, but rather matched on reports including counts of histologic-specific lesions. We calculated standard measures of agreement (kappa score) between manually extracted and algorithm-derived classifications for histology (at least one correct classification for each diagnosis) and for histologic-specific lesion count (0, 1, 2, and ≥ 3 lesions occurring within a single report). Agreement was calculated for each of the 3 independent validation samples.

RESULTS

Development

The algorithm achieved high recall (> 0.9), precision, and F1 scores in the evaluation of the parameter combinations for BCC and SCC within the train/test splits. Agreement measures for diagnosing keratoacanthoma and IEC were slightly less with F1 scores of 0.89 and 0.86, respectively (Table 3). See Appendix Tables A1 and A2 and Appendix Figures A1 and A2 for detailed results.

Validation

We observed high accuracy for classifying histologic subtypes of skin cancer across the 3 validation data sets. Kappa scores for validation data sets 1 and 2 were almost perfect for BCC, SCC, and keratoacanthoma (> 0.9) and were high for IEC (0.89; Table 4). However, approximately 7% of pathology reports from validation data set 2 could not be processed because of formatting irregularities.

Although agreement indices were slightly lower for validation data set 3 (the cohort of organ transplantation recipients with high incidence and multiplicity of skin cancer), kappa scores were high for BCC (0.82), SCC

TABLE 3. Accuracy of Final Algorithm for Diagnosis Classification in Test Split of Training Data Set in Development

	Agreement		
Diagnosis	**F1 Score**	**Recall (sensitivity)**	**Precision (PPV)**
BCC	0.93	0.94	0.93
SCC	0.91	0.92	0.89
Keratoacanthoma	0.89	0.91	0.88
Intraepidermal carcinoma	0.86	0.87	0.85

Abbreviations: BCC, basal cell carcinoma; PPV, positive predictive value; SCC, squamous cell carcinoma.

(0.90), and IEC (0.89). A lower sensitivity was found for BCCs in this data set (83%), largely because the application could not separate 8 BCCs diagnosed in one pathology report.

Across all 3 validation data sets, accuracy of histology-specific lesion counts was slightly lower than for histologic classification. Even so, kappa scores generally remained higher than 0.8 (Appendix Table A3).

Kappa scores for site of lesion were high for validation data set 1 but lower for some sites in validation data set 2 for head and neck (0.89 *v* 0.78, respectively), torso (0.83 *v* 0.69, respectively), and limbs (0.91 *v* 0.74, respectively). Further agreement calculations and agreement for face-specific sites are provided in Appendix Table A4. A gold standard for site of lesion was not available for validation data set 3.

DISCUSSION

We developed a Web application to automatically extract diagnostic information from free-text pathology reports. The application underwent extensive validation and was found to be highly accurate for classifying diagnoses of keratinocyte cancers within a large, prospective study. Its utility among transplantation patients with complex pathology reports was slightly lower. However, it must be noted that the reports in this group frequently described > 10 lesions in a single report. In addition to overall accuracy, sensitivity and positive predictive value for BCC and SCC were particularly high, indicating high ascertainment and few false negatives.

TABLE 4. Accuracy of Classifying at Least One Case of the Diagnosis in Each Report and Agreement Between Algorithm-Derived and Manual Review (gold standard) Sample of Reports in QSkin Study Participants and External Study Participants (STAR study)

	No. of Reports With at Least One Case of the Diagnosis		Agreement (95% CI)			
Source	**Gold Standard**	**AD**	**Sensitivity (%)**	**Specificity (%)**	**PPV (%)**	**κ**
Data set 1: QSkin participants, new pathology reports (n = 348)[a]						
BCC	99	99	98 (92 to 99)	99 (97 to 100)	98 (93 to 100)	0.97 (0.94 to 1.0)
SCC	33	33	94 (80 to 99)	99 (98 to 100)	94 (80 to 99)	0.93 (0.87 to 1.0)
Keratoacanthoma	15	13	87 (60 to 98)	100 (99 to 100)	99 (75 to 100)	0.93 (0.82 to 1.0)
IEC	72	70	90 (81 to 96)	98 (96 to 99)	93 (84 to 98)	0.89 (0.83 to 1.0)
Data set 2: QSkin participants, skin pathology reports from laboratories not included in the training data set (n = 2,159)[a]						
BCC	932	937	98 (97 to 99)	98 (97 to 99)	97 (96 to 98)	0.96 (0.94 to 0.97)
SCC	295	317	98 (95 to 99)	98 (98 to 99)	91 (87 to 94)	0.93 (0.91 to 0.95)
Keratoacanthoma	50	55	96 (87 to 99)	100 (99 to 100)	87 (76 to 94)	0.91 (0.86 to 0.97)
IEC	348	322	88 (84 to 91)	99 (99 to 99)	95 (92 to 97)	0.89 (0.87 to 0.92)
Data set 3: STAR study participants, skin pathology reports (n = 42 reports)[a,b]						
BCC	12	11	83 (55 to 95)	97 (83 to 99)	91 (62 to 98)	0.82 (0.63 to 1.0)
SCC	16	16	94 (72 to 99)	96 (81 to 99)	94 (72 to 99)	0.90 (0.76 to 1.0)
IEC	25	22	88 (70 to 96)	100 (100 to 100)	100 (100 to 100)	0.89 (0.70 to 1.0)

Abbreviations: AD, algorithm derived; BCC, basal cell carcinoma; IEC, intraepidermal carcinoma; PPV, positive predictive value; SCC, squamous cell carcinoma.

[a] The count of reports for histologic diagnoses (BCC, SCC, keratoacanthoma, and IEC) does not sum to the number of reports processed. More than one diagnosis could be counted on a single report, and some reports included other benign skin diagnoses not classified by the algorithm.

[b] Diagnosis of keratoacanthoma and site of lesion were not collected by the study dermatologist.

Agreement between algorithm-derived and manually extracted information on the site of lesion was slightly lower than that observed for type of lesion. This is likely because of inconsistencies in the collection of this data item. Expert reviewers were required to allocate the site of a lesion from an extensive, but not exhaustive, list. As an example, a lesion on the lower neck or upper back region may have been entered as neck, shoulder, or upper back. Similarly, a lesion described as located on the hip could potentially be entered as being on the buttock, torso, or thigh. This inconsistency likely affected the ability of the algorithm to accurately determine site.

To the best of our knowledge, this is the only automated method for extracting diagnostic information from free-text pathology reports for keratinocyte cancers. Eide et al[9] used natural language processing to identify incident cases of keratinocyte cancers from pathology reports appropriate for registration but did not extract pathology data using these methods.

The automated extraction of information from cancer histopathology reports is complex. Free-text reporting by pathologists results in large and complex variety in the language used to describe a diagnosis (or lack of diagnosis).[16,17] The main challenge for the automated algorithm arises from multiple lesions being described in a single pathology report. To overcome this, we developed rules in the application to separately extract information specific for each lesion and then map the components together again. Similar to Currie et al,[18] the Web application generates an alert to flag the small number of reports that failed processing.

Strengths of the study include full manual reviews of > 25,000 pathology reports, yielding a training data set of sufficient quality and size for supervised learning development. However, the application is limited in that it can only assign one diagnosis to a single lesion. For example, “squamous cell carcinoma arising in a keratoacanthoma” was classified as SCC, whereas a medical reviewer would classify this lesion as both SCC and keratoacanthoma. This occurred in approximately 1% of lesions classified by the application. For the purposes of defining skin cancer incidence in a population, we contend that the coding rules developed here are acceptable.

Unlike other attempts to automate the extraction of information from pathology reports,[16,18] we report our detailed methods and used open-source software. Thus, although the findings in this report are specific to the format and language used in pathology reports for keratinocyte cancers in the study population, the preprocessing rules can be easily adapted to suit different text formats and the supervised learning methods could be applied to a different training data set.

In conclusion, a supervised learning Web application can process large numbers of pathology reports and classify and count diagnoses of keratinocyte cancers described in free-text histopathology reports with a high degree of accuracy. This tool was developed primarily for compiling statistical summary information in settings where such data are not currently able to be recorded as a result of the volume and complexity of data. Similar applications could be implemented into cancer registries and hospitals, which would enable the measurement of histology type–specific keratinocyte cancer incidence rates.

AFFILIATIONS

[1]Department of Population Health, QIMR Berghofer Medical Research Institute, Brisbane Queensland, Australia
[2]Otso, Brisbane, Queensland, Australia
[3]School of Public Health, University of Queensland, Brisbane, Queensland, Australia
[4]Leeds Institute of Medical Research at St James’s, University of Leeds, Leeds, United Kingdom
[5]Max Kelsen, Brisbane, Queensland, Australia
[6]Faculty of Medicine, University of Queensland, Brisbane, Queensland, Australia

CORRESPONDING AUTHOR

David C. Whiteman, MBBS, PhD, Cancer Control Group, QIMR Berghofer Medical Research Institute, 300 Herston Rd, Herston, Queensland 4006, Australia; Twitter: @QIMRBerghofer; e-mail: David.Whiteman@qimrberghofer.edu.au.

SUPPORT

Supported in part by the National Health and Medical Research Council (NHMRC) of Australia (Grant No. 552429) and by a fellowships from the NHMRC (D.C.W.).

AUTHOR CONTRIBUTIONS

Conception and design: Bridie S. Thompson, Sam Hardy, Athon Millane, Daniel Bourke, Cameron D. Bean, Catherine M. Olsen, David C. Whiteman
Financial support: David C. Whiteman
Administrative support: Sam Hardy, Cameron D. Bean, David C. Whiteman
Provision of study materials or patients: David C. Whiteman
Collection and assembly of data: Bridie S. Thompson, Nirmala Pandeya, Jean Claude Dusingize, Adele C. Green, Ronald Grande, Catherine M. Olsen, David C. Whiteman
Data analysis and interpretation: Bridie S. Thompson, Sam Hardy, Nirmala Pandeya, Adele C. Green, Athon Millane, Daniel Bourke, Cameron D. Bean, Catherine M. Olsen, David C. Whiteman
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable of all aspects of the work: All authors

AUTHORS’ DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO’s conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Sam Hardy
Employment: Max Kelsen

Athon Millane
Other Relationship: Max Kelsen

Daniel Bourke
Employment: Max Kelsen

Ronald Grande
Employment: Max Kelsen

Cameron D. Bean
Speakers' Bureau: London Speakers Bureau (I)

David C. Whiteman
Employment: Fullerton Health Care (I)

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

We acknowledge the work of Marcia Davis, MD, who was responsible for reporting skin cancer diagnoses from pathology reports for participants of the STAR study.

REFERENCES

1. Rogers HW, Weinstock MA, Feldman SR, et al: Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the U.S. population, 2012. JAMA Dermatol 151:1081-1086, 2015
2. Staples MP, Elwood M, Burton RC, et al: Non-melanoma skin cancer in Australia: The 2002 national survey and trends since 1985. Med J Aust 184:6-10, 2006
3. National Cancer Intelligence Network: Non-melanoma skin cancer in England, Scotland, Northern Ireland, and Ireland: NCIN data briefing 2013. http://www.ncin.org.uk/publications/data_briefings/non_melanoma_skin_cancer_in_england_scotland_northern_ireland_and_ireland
4. Fransen M, Karahalios A, Sharma N, et al: Non-melanoma skin cancer in Australia. Med J Aust 197:565-568, 2012
5. Hanauer DA, Miela G, Chinnaiyan AM, et al: The registry case finding engine: An automated tool to identify cancer cases from unstructured, free-text pathology reports and clinical notes. J Am Coll Surg 205:690-697, 2007
6. Jouhet V, Defossez G, Burgun A, et al: Automated classification of free-text pathology reports for registration of incident cases of cancer. Methods Inf Med 51:242-251, 2012
7. Glaser AP, Jordan BJ, Cohen J, et al: Automated extraction of grade, stage, and quality information from transurethral resection of bladder tumor pathology reports using natural language processing. JCO Clin Cancer Inform 2:1-8, 2018
8. Spasić I, Livsey J, Keane JA, et al: Text mining of cancer-related information: Review of current status and future directions. Int J Med Inform 83:605-623, 2014
9. Eide MJ, Tuthill JM, Krajenta RJ, et al: Validation of claims data algorithms to identify nonmelanoma skin cancer. J Invest Dermatol 132:2005-2009, 2012
10. Bray F, Ferlay J, Soerjomataram I, et al: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68:394-424, 2018
11. Australian Institute of Health and Welfare: Skin cancer in Australia. Cat. no. CAN 96. Canberra, Australia, AIHW. 2016
12. Thompson BS, Olsen CM, Subramaniam P, et al: Medicare claims data reliably identify treatments for basal cell carcinoma and squamous cell carcinoma: A prospective cohort study. Aust N Z J Public Health 40:154-158, 2016
13. Olsen CM, Green AC, Neale RE, et al: Cohort profile: The QSkin Sun and Health Study. Int J Epidemiol 41:929-929i, 2012
14. Sidey-Gibbons JAM, Sidey-Gibbons CJ: Machine learning in medicine: A practical introduction. BMC Med Res Methodol 19:64, 2019
15. Iannacone MR, Sinnya S, Pandeya N, et al. Prevalence of skin cancer and related skin tumors in high-risk kidney and liver transplant recipients in Queensland, Australia. *J Invest Dermatol* 136:1382-1386, 2016
16. Buckley JM, Coopey SB, Sharko J, et al: The feasibility of using natural language processing to extract clinical information from breast pathology reports. J Pathol Inform 3:23, 2012
17. Nguyen AN, Moore J, O'Dwyer J, et al: Assessing the utility of automatic cancer registry notifications data extraction from free-text pathology reports. AMIA Annu Symp Proc 2015:953-962, 2015
18. Currie AM, Fricke T, Gawne A, et al: Automated extraction of free-text from pathology reports. AMIA Annu Symp Proc 2006:899, 2006

APPENDIX

TABLE A1. Test Results for Accuracy of Algorithm Prediction for Diagnosis

Label name[a]	Histologic Name	PPV	Sensitivity	F1 Score[b]	Test Data Count
diagnosis_11	BCC	0.93	0.94	0.93	2,343
diagnosis_12	SCC	0.89	0.92	0.91	809
diagnosis_13	Melanoma	0.83	0.88	0.85	163
diagnosis_21	Keratoacanthoma	0.88	0.91	0.89	193
diagnosis_22	IEC	0.85	0.87	0.86	1,581
diagnosis_24	Solar keratosis	0.77	0.74	0.75	976
diagnosis_44	Lentigo maligna	0.58	0.65	0.61	40
diagnosis_61	BCC re-excision	0.78	0.72	0.75	96
diagnosis_62	SCC re-excision	0.62	0.52	0.57	54
diagnosis_63	IEC re-excision	0.61	0.40	0.48	43
diagnosis_64	Melanoma re-excision	0.84	0.92	0.88	107
diagnosis_65	Squamoproliferative lesions	0.61	0.49	0.54	35
diagnosis_88	Nonmalignant	0.91	0.90	0.91	2,310
Micro average		0.88	0.88	0.88	8,750
Macro average		0.78	0.76	0.76	
Weighted average		0.88	0.88	0.88	

Abbreviations: BCC, basal cell carcinoma; IEC, intraepidermal carcinoma; PPV, positive predictive value; SCC, squamous cell carcinoma.

[a] Label names are machine derived, were assigned in the development process, and identify discrete classification categories that mapped to histological diagnoses.

[b] F1 score is a measure of accuracy and represents the harmonic mean of PPV and sensitivity.

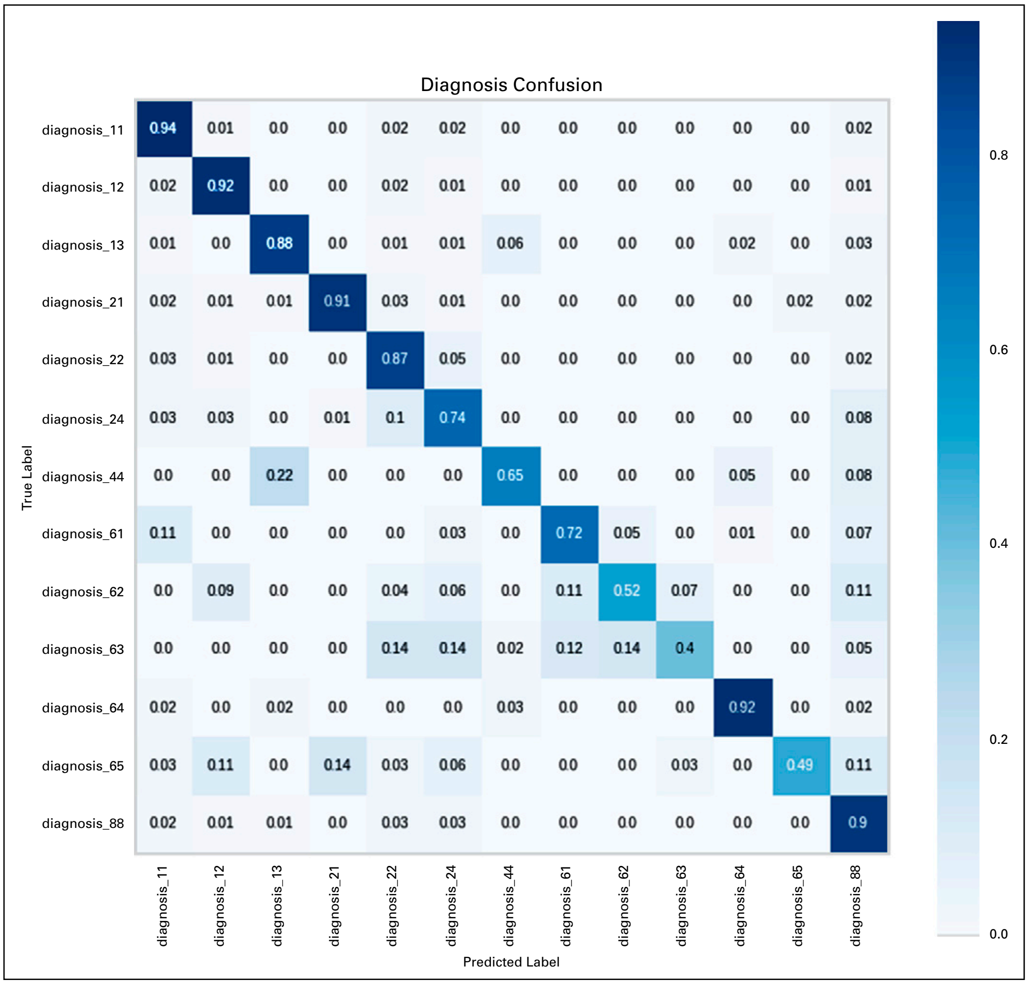


FIG A1. Test results for agreement (F1 score) and discordance of diagnoses between the predicted labels (algorithm derived classification) and true labels (actual diagnosis). Histologic names for labels are detailed in Table A1.

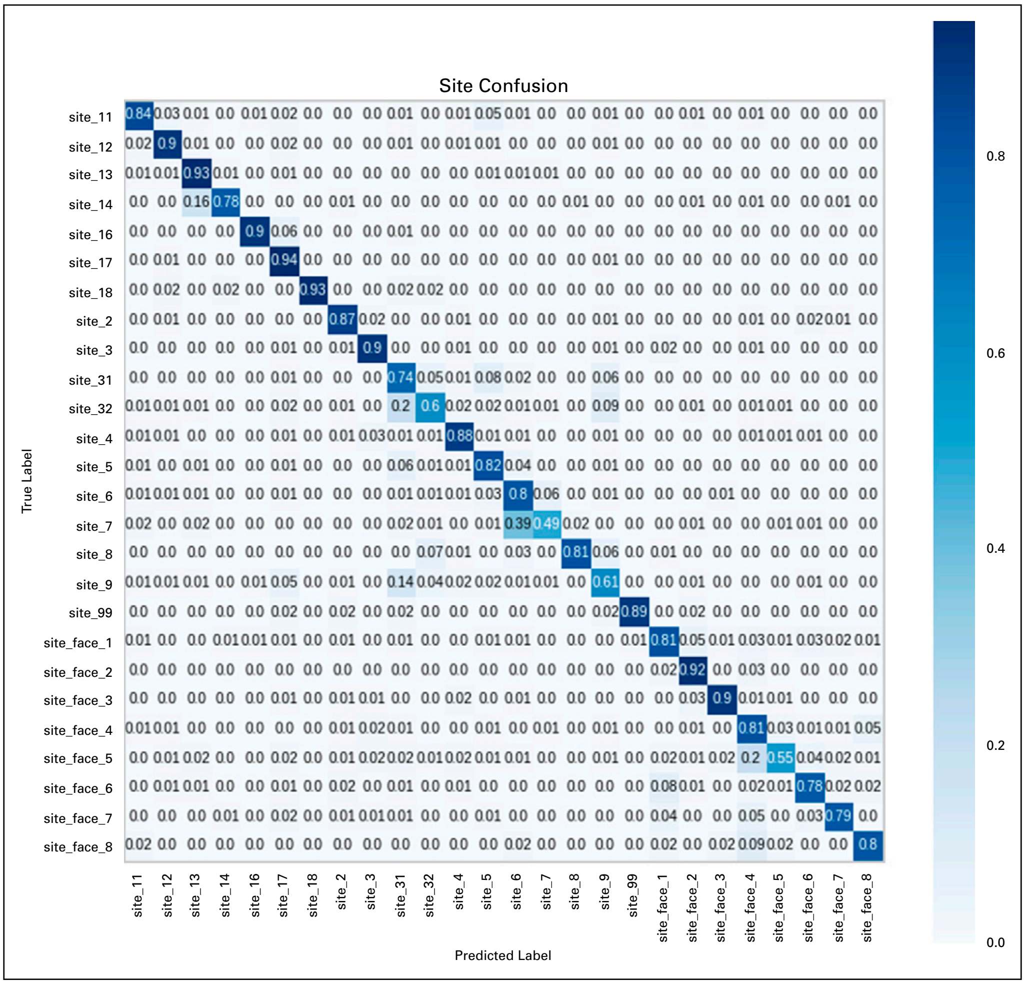
TABLE A2. Test Results for Accuracy of Algorithm Prediction for Site

Label Name[a]	Anatomic Site	PPV	Sensitivity	F1 Score[b]	Test Data Count
site_face_1	Skin of orbit/eyelid	0.72	0.81	0.76	199
site_face_2	Nose	0.91	0.92	0.92	481
site_face_3	Lips	0.86	0.90	0.88	117
site_face_4	Cheeks	0.82	0.81	0.81	495
site_face_5	Chin/jaw	0.69	0.55	0.61	125
site_face_6	Forehead	0.86	0.78	0.82	357
site_face_7	Temple	0.87	0.79	0.83	221
site_face_8	Face (not specified)	0.54	0.80	0.65	46
site_2	Scalp	0.87	0.87	0.87	216
site_3	Ears	0.86	0.90	0.88	264
site_4	Neck	0.88	0.88	0.88	362
site_5	Shoulders	0.81	0.82	0.81	484
site_6	Upper chest/sternoclavicular	0.73	0.80	0.76	338
site_7	Breast	0.66	0.49	0.56	125
site_8	Abdomen	0.88	0.81	0.84	69
site_9	Back (not specified)	0.59	0.61	0.60	209
site_11	Upper arm	0.86	0.84	0.85	397
site_12	Forearm, elbow, or wrist	0.93	0.90	0.92	780
site_13	Back of hand	0.87	0.93	0.90	345
site_14	Palmar skin, fingers	0.80	0.78	0.79	86
site_16	Thigh	0.91	0.90	0.91	201
site_17	Lower leg, ankle, knee	0.93	0.94	0.93	1,017
site_18	Top of feet	0.88	0.93	0.90	55
site_31	Upper back	0.69	0.74	0.71	417
site_32	Lower back	0.66	0.60	0.63	174
site_99	Nonskin	0.93	0.89	0.91	44
Micro average		0.84	0.84	0.84	7,624
Macro average		0.81	0.81	0.81	
Weighted average		0.84	0.84	0.84	

Abbreviation: PPV, positive predictive value.

[a] Label names are machine derived, were assigned in the development process, and identify discrete classification categories that mapped to anatomical site.

[b] F1 score is a measure of accuracy and represents the harmonic mean of PPV and sensitivity.



	site_11	site_12	site_13	site_14	site_16	site_17	site_18	site_2	site_3	site_31	site_32	site_4	site_5	site_6	site_7	site_8	site_9	site_99	site_face_1	site_face_2	site_face_3	site_face_4	site_face_5	site_face_6	site_face_7	site_face_8
site_11	0.84	0.03	0.01	0.0	0.01	0.02	0.0	0.0	0.0	0.01	0.0	0.01	0.05	0.01	0.0	0.0	0.01	0.0	0.0	0.01	0.0	0.01	0.0	0.0	0.0	0.0
site_12	0.02	0.9	0.01	0.0	0.0	0.02	0.0	0.0	0.0	0.01	0.0	0.01	0.01	0.0	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
site_13	0.01	0.01	0.93	0.01	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.01	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
site_14	0.0	0.0	0.16	0.78	0.0	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.0	0.0	0.0	0.01	0.0	0.01	0.0	0.0	0.01	0.0
site_16	0.0	0.0	0.0	0.0	0.9	0.06	0.0	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
site_17	0.0	0.01	0.0	0.0	0.0	0.94	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
site_18	0.0	0.02	0.0	0.02	0.0	0.0	0.93	0.0	0.0	0.02	0.02	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
site_2	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.87	0.02	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.01	0.0	0.02	0.01	0.0
site_3	0.0	0.0	0.0	0.0	0.0	0.01	0.0	0.01	0.9	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.01	0.0	0.02	0.0	0.0	0.01	0.0	0.0	0.0	0.0
site_31	0.0	0.0	0.0	0.0	0.0	0.01	0.0	0.0	0.0	0.74	0.05	0.01	0.08	0.02	0.0	0.0	0.06	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
site_32	0.01	0.01	0.01	0.0	0.0	0.02	0.0	0.01	0.0	0.2	0.6	0.02	0.02	0.01	0.01	0.0	0.09	0.0	0.0	0.01	0.0	0.01	0.01	0.0	0.0	0.0
site_4	0.01	0.01	0.0	0.0	0.0	0.01	0.0	0.01	0.03	0.01	0.01	0.88	0.01	0.01	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.01	0.01	0.01	0.0	0.0
site_5	0.01	0.0	0.01	0.0	0.0	0.01	0.0	0.0	0.0	0.06	0.01	0.01	0.82	0.04	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
site_6	0.01	0.01	0.01	0.0	0.0	0.01	0.0	0.0	0.0	0.01	0.01	0.01	0.03	0.8	0.06	0.0	0.01	0.0	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0
site_7	0.02	0.0	0.02	0.0	0.0	0.0	0.0	0.0	0.0	0.02	0.01	0.0	0.01	0.39	0.49	0.02	0.0	0.0	0.0	0.01	0.0	0.0	0.01	0.01	0.0	0.0
site_8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.07	0.01	0.0	0.03	0.0	0.81	0.06	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0
site_9	0.01	0.01	0.01	0.0	0.01	0.05	0.0	0.01	0.0	0.14	0.04	0.02	0.02	0.01	0.01	0.0	0.61	0.0	0.0	0.01	0.0	0.0	0.0	0.01	0.0	0.0
site_99	0.0	0.0	0.0	0.0	0.0	0.02	0.0	0.02	0.0	0.02	0.0	0.0	0.0	0.0	0.0	0.0	0.02	0.89	0.0	0.02	0.0	0.0	0.0	0.0	0.0	0.0
site_face_1	0.01	0.0	0.0	0.01	0.01	0.01	0.0	0.01	0.0	0.01	0.0	0.0	0.01	0.01	0.0	0.0	0.0	0.01	0.81	0.05	0.01	0.03	0.01	0.03	0.02	0.01
site_face_2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.02	0.92	0.0	0.03	0.0	0.0	0.0	0.0
site_face_3	0.0	0.0	0.0	0.0	0.0	0.01	0.0	0.01	0.01	0.0	0.0	0.02	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.03	0.9	0.01	0.01	0.0	0.0	0.0
site_face_4	0.01	0.01	0.0	0.0	0.0	0.0	0.0	0.01	0.02	0.01	0.0	0.0	0.01	0.0	0.01	0.0	0.01	0.0	0.0	0.01	0.0	0.81	0.03	0.01	0.01	0.05
site_face_5	0.0	0.01	0.02	0.0	0.0	0.02	0.0	0.01	0.02	0.02	0.01	0.02	0.01	0.01	0.0	0.0	0.01	0.0	0.02	0.01	0.02	0.2	0.55	0.04	0.02	0.01
site_face_6	0.0	0.01	0.01	0.0	0.0	0.01	0.0	0.02	0.0	0.01	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.08	0.01	0.0	0.02	0.01	0.78	0.02	0.02
site_face_7	0.0	0.0	0.0	0.01	0.0	0.02	0.0	0.01	0.01	0.01	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.04	0.0	0.0	0.05	0.0	0.03	0.79	0.0
site_face_8	0.02	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.02	0.0	0.0	0.0	0.0	0.02	0.0	0.02	0.09	0.02	0.0	0.0	0.8

FIG A2. Test results for agreement (F1 score) and discordance of site between the predicted labels (algorithm-predicted site) and true labels (actual site). Anatomic site names for labels are detailed in Table A2.

TABLE A3. Count of Each Diagnosis for Each Person and Agreement Between Algorithm-Derived Extraction and Manual Review (gold standard) From 3 Validation Sources

	Diagnosis Count by Manually Reviewed Gold Standard					
Algorithm-Derived Diagnosis Count	**0**	**1**	**2**	**≥ 3**	**Total**	**Weighted κ (95% CI)**
Data set 1: Random sample of 400 new reports for QSkin participants						
BCC						0.94 (0.90 to 0.98)
0	247	2	0	0	249	
1	2	73	1	2	78	
2	0	2	15	0	17	
≥ 3	0	0	0	4	4	
Total	249	77	16	6	348	
SCC						0.92 (0.84 to 1.0)
0	313	2	0	—	315	
1	1	29	0	—	30	
2	1	0	2	—	3	
Total	315	31	2	—	348	
Keratoacanthoma						0.93 (0.83 to 1.0)
0	333	2	0	—	335	
1	0	12	0	—	12	
2	0	0	1	—	1	
Total	333	14	1	—	348	
Intraepidermal carcinoma						0.88 (0.81 to 0.95)
0	271	6	0	1	278	
1	4	52	0	0	56	
2	1	1	9	1	12	
≥ 3	0	0	0	2	2	
Total	276	59	9	4	348	
Data set 2: Reports from other laboratories for QSkin participants						
BCC						0.85 (0.83 to 0.86)
0	1,177	20	0	0	1,197	
1	21	577	2	0	600	
2	2	129	97	1	229	
≥ 3	2	26	22	58	108	
Total	1,202	752	121	59	2,134	
SCC						0.80 (0.77 to 0.83)
0	1,810	7	0	0	1,817	
1	28	199	2	0	229	
2	1	58	11	0	70	
≥ 3	0	14	3	1	18	
Total	1,839	278	16	1	2,134	
Keratoacanthoma						0.84 (0.78 to 0.91)
0	2,077	2	0	0	2,079	
1	7	39	1	0	47	

(Continued on following page)

TABLE A3. Count of Each Diagnosis for Each Person and Agreement Between Algorithm-Derived Extraction and Manual Review (gold standard) From 3 Validation Sources (Continued)

	Diagnosis Count by Manually Reviewed Gold Standard					
Algorithm-Derived Diagnosis Count	**0**	**1**	**2**	**≥ 3**	**Total**	**Weighted κ (95% CI)**
2	0	6	1	0	7	
≥ 3	0	1	0	0	1	
Total	2,084	48	2	0	2,134	
Intraepidermal carcinoma						0.83 (0.81 to 0.86)
0	1,770	42	0	0	1,812	
1	16	223	3	0	242	
2	0	45	20	1	66	
≥ 3	0	2	5	7	14	
Total	1,786	312	28	8	2,134	
Data set 3: Reports from participants from the STAR study						
BCC						0.83 (0.66 to 1.0)
0	29	2	0	0	31	
1	1	7	0	0	8	
2	0	1	1	0	2	
≥ 3	0	0	0	1	1	
Total	30	10	1	1	42	
SCC						0.91 (0.78 to 1.0)
0	25	1	0	—	26	
1	1	14	0	—	15	
2	0	0	1	—	1	
Total	26	15	1	—	42	
Intraepidermal carcinoma						0.88 (0.77 to 0.99)
0	17	3	0	0	20	
1	0	11	0	1	12	
2	0	1	1	0	2	
≥ 3	0	0	0	8	8	
Total	17	15	1	9	42	

Abbreviations: BCC, basal cell carcinoma; SCC, squamous cell carcinoma.

TABLE A4. Accuracy of Classifying at Least One Keratinocyte Cancer at Each Site in a Report and Agreement Between Algorithm-Derived Extraction and Manual Review (gold standard) Sample of Reports

	No. of Reports With at Least One Lesion Occurring in That Site		Agreement (95% CI)			
Source and Site	**Gold Standard**	**AD**	**Sensitivity (%)**	**Specificity (%)**	**PPV (%)**	**κ**
Data set 1: QSkin participants, new pathology reports (n = 349)						
Site (limited to confirmed BCC, SCC, keratoacanthoma, or IEC; n = 199)[a]						
Head and neck	92	97	97 (91 to 97)	93 (87 to 97)	92 (86 to 97)	0.89 (0.83 to 0.95)
Torso	39	44	92 (79 to 98)	95 (90 to 98)	82 (67 to 92)	0.83 (0.74 to 0.93)
Limbs	89	88	94 (87 to 98)	96 (91 to 99)	96 (89 to 99)	0.91 (0.85 to 0.97)
Face-specific site (n = 63)						
Eye orbit and lid	9	7	78 (40 to 97)	100 (94 to 100)	100 (59 to 100)	0.86 (0.66 to 1.0)
Nose	18	17	94 (79 to 100)	100 (92 to 100)	100 (81 to 100)	0.96 (0.88 to 1.0)
Lips	2	2	100 (16 to 100)	100 (94 to 100)	100 (16 to 100)	1.0 (1.0 to 1.0)
Cheeks	8	11	100 (66 to 100)	95 (85 to 99)	73 (43 to 95)	0.82 (0.61 to 1.0)
Jaw and chin	3	2	33 (1 to 91)	98 (91 to 100)	50 (1 to 99)	0.38 (< 0.0 to 0.93)
Forehead	14	15	100 (68 to 100)	98 (93 to 100)	93 (77 to 100)	0.96 (0.87 to 1.0)
Temple	8	6	75 (35 to 97)	100 (94 to 100)	100 (54 to 100)	0.84 (0.62 to 1.0)
Face (not specified)	4	4	75 (19 to 99)	98 (90 to 100)	95 (19 to 99)	0.73 (0.38 to 1.0)
Data set 2: QSkin participants, skin pathology reports from laboratories not included in the training data set (n = 2,159 reports)						
Site (limited to confirmed BCC, SCC, keratoacanthoma, or IEC; n = 1,472)[a]						
Head and neck	636	605	85 (82 to 88)	92 (90 to 94)	89 (87 to 92)	0.78 (0.75 to 0.81)
Torso	407	510	86 (82 to 90)	85 (83 to 87)	69 (64 to 73)	0.69 (0.65 to 0.73)
Limbs	549	544	84 (81 to 86)	91 (89 to 93)	86 (83 to 89)	0.74 (0.71 to 0.78)
Face-specific site (n = 473)						
Eye orbit and lid	52	54	67 (54 to 78)	100 (93 to 97)	65 (51 to 76)	0.62 (0.50 to 0.73)
Nose	147	156	88 (82 to 93)	92 (89 to 95)	83 (77 to 88)	0.79 (0.73 to 0.85)
Lips	18	18	89 (67 to 97)	100 (98 to 100)	89 (67 to 70)	0.88 (0.77 to 1.0)
Cheeks	123	124	79 (71 to 85)	92 (89 to 95)	78 (70 to 85)	0.71 (0.64 to 0.78)
Jaw and chin	29	17	48 (31 to 66)	99 (98 to 100)	82 (59 to 94)	0.59 (0.42 to 0.76
Forehead	95	84	74 (64 to 81)	96 (94 to 98)	83 (74 to 90)	0.73 (0.65 to 0.81)
Temple	36	32	72 (56 to 84)	99 (97 to 99)	81 (65 to 91)	0.75 (0.63 to 0.87)
Face (not specified)	3	2	67 (21 to 94)	100 (100 to 100)	100 (100 to 100)	0.80 (0.41 to 1.0)

Abbreviations: AD, algorithm derived; BCC, basal cell carcinoma; IEC, intraepidermal carcinoma; PPV, positive predictive value; SCC, squamous cell carcinoma.

[a] The count of reports for site does not sum to the number of reports processed because > 1 site could be counted on a single report.