Development and Validation of an Algorithm to Identify Patients with Advanced Cutaneous Squamous Cell Carcinoma from Pathology Reports

JID Open

Celeste Eggermont[1], Marlies Wakkee[1], Annette Bruggink[2], Quirinus Voorham[2], Kay Schreuder[3], Marieke Louwman[3], Antien Mooyaart[4] and Loes Hollestein[1,3]

To facilitate nationwide epidemiological research on advanced cutaneous squamous cell carcinoma (cSCC), that is, locally advanced, recurrent, or metastatic cSCC, we sought to develop and validate a rule-based algorithm that identifies advanced cSCC from pathology reports. The algorithm was based on both hierarchical histopathological codes and free text from pathology reports recorded in the National Pathology Registry. Medical files from the Erasmus Medical Center of 186 patients with stage III/IV/recurrent cSCC and 184 patients with stage I/II cSCC were selected and served as the gold standard to assess the performance of the algorithm. The rule-based algorithm showed a sensitivity of 91.9% (95% confidence interval = 88.0–95.9), a specificity of 96.7% (95% confidence interval = 94–2-99.3), and a positive predictive value of 78.5% (95% confidence interval = 74.2–82.8) for all advanced cSCC combined. The sensitivity was lower per subgroup: locally advanced (52.3–86.2%), recurrent cSCC (23.3%), and metastatic cSCC (70.0%). The specificity per subgroup was above 97%, and the positive predictive value was above 78%, with the exception of metastatic cSCC, which had a positive predictive value of 62%. This algorithm can be used to identify advanced patients with cSCC from pathology reports and will facilitate large-scale epidemiological studies of advanced cSCC in the Netherlands and internationally after external validation.

Journal of Investigative Dermatology (2023) **143,** 98–104; doi:10.1016/j.jid.2022.07.008

INTRODUCTION

Cutaneous squamous cell carcinoma (cSCC) is one of the most common cancers in humans and is still increasing (Lomas et al., 2012; Tokez et al., 2020). Despite the high incidence rates, cSCC is excluded from many national cancer registries, including the United States (Wehner, 2020). Even if data on the primary tumor are registered, no country collects data on follow-up (Adalsteinsson et al., 2021; Guorgis et al., 2020; Stang et al., 2019; Tokez et al., 2022; Venables et al., 2019). The rationale is that given the high incidence rates and relatively low occurrence of metastatic cSCC, manually reviewing all pathology reports and patient files for disease progression is not feasible. However, owing to the high overall cSCC incidence rates, the absolute number of patients with advanced cSCC is also significant. These advanced patients are at risk of death (Tokez et al., 2022), but no national data are currently being collected. However, a good estimate of the probability and risk factors for disease recurrence and the likelihood of local and systemic progression would help to improve treatment decisions and surveillance recommendations. Automated identification of advanced cSCC (i.e., locally advanced, recurrent, or metastatic cSCC) could therefore represent a feasible, cost-effective solution for cancer registries to target this subgroup of patients with cSCC.

The use of automated extraction from free-text pathology reports to select patients for cancer registries has been previously reported in the literature (Glaser et al., 2018; Hanauer et al., 2007; Jouhet et al., 2012; Nguyen et al., 2015). Two studies used natural language processing (i.e., the application of computational techniques that aid computers in comprehending, interpreting, and manipulating human language) to automatically identify keratinocyte cancers but did not concentrate on advanced cSCC (Eide et al., 2012; Thompson et al., 2020).

We aimed to develop and validate an algorithm to identify patients with advanced cSCC from pathology reports. Automatic identification of patients with advanced cSCC using this algorithm will facilitate research on advanced cSCC at a population-based level.

[1]*Department of Dermatology, Erasmus MC Cancer Institute, University Medical Center, Rotterdam, The Netherlands;* [2]*Nationwide Network and Registry of Histo- and Cytopathology (PALGA), Houten, The Netherlands;* [3]*Department of Research and Development, Netherlands Comprehensive Cancer Organization (IKNL), Utrecht, The Netherlands; and* [4]*Department of Pathology, Erasmus MC Cancer Institute, University Medical Center, Rotterdam, The Netherlands*

Correspondence: Loes Hollestein, Department of Dermatology, Erasmus Medical Center, PO Box 2040, 3000 CA, Rotterdam, The Netherlands. E-mail: l.hollestein@erasmusmc.nl

Abbreviations: AJCC8, American Joint Committee on Cancer eighth edition; BWH, Brigham and Women's Hospital; CI, confidence interval; cSCC, cutaneous squamous cell carcinoma; PALGA, Nationwide Network and Registry of Histo- and Cytopathology; PPV, positive predictive value; SCC, squamous cell carcinoma; SNOMED-CT, Systemized Nomenclature of Medicine - Clinical Terms

Received 20 January 2022; revised 22 June 2022; accepted 11 July 2022; accepted manuscript published online 1 August 2022; corrected proof published online 15 September 2022

Table 1. Algorithm to Identify Advanced cSCC from Pathology Reports

Include	Exclude	Codelist[1]
Locally advanced primary cSCC T3/T4 (AJCC8)		
Primary cSCC code combined with skin/subcutis OR		1
PALGA sublocalization code likely to be a cSCC AND		2
any of the following criteria in the free text:	Unlikely primary cSCC localization	3
T3 or T4		4
Diameter >4 cm		5
Invasion depth (likely to be) >6 mm		6
Invasion beyond subcutaneous fat		7
Invasion in the muscles		8
Invasion in deep structures		9
Bone erosion or invasion		10
Nerve invasion ≥0.1 mm		11
Perineural invasion	Nerve invasion <0.1 mm	12
Angioinvasive growth		13
Invasion up to the bottom of the excision		14
Locally advanced primary cSCC T2b/T3 (BWH)		
Primary cSCC code combined with skin/subcutis OR		1
PALGA sublocalization code likely to be a cSCC		2
AND the free text indicating:	Unlikely primary cSCC localization	3
Bone erosion or invasion		10
At least two of the following criteria:		
Diameter ≥2 cm		15
Poor differentiation		16
Nerve invasion ≥0.1 mm or perineural invasion	Nerve invasion <0.1 mm	11 12
Invasion beyond subcutaneous fat		7
Invasion in the muscles		8
Invasion in deep structures		9
Local recurrence		
Primary cSCC code and free text indicating a recurrence		1 and 17
Metastasis		
Metastatic SCC code combined with skin/subcutis		18
Primary or (possibly) metastatic SCC code in the parotid, salivary, or submandibular gland		19
Free text indicating 'metastasis'		20
	Any unlikely metastatic cSCC localization	21
	Unlikely metastatic cSCC localization without skin/subcutis code	22
	Removal of the lungs/bronchus	23
	No malignancy/metastasis	24
	Unknown primary site	25

Abbreviations: AJCC8, American Joint Committee on Cancer, eight edition; BWH, Brigham and Women Hospital; cSCC, cutaneous squamous cell carcinoma; PALGA, Nationwide Network and Registry of Histo- and Cytopathology; SNOMED-CT, Systemized Nomenclature of Medicine – Clinical Terms
[1]See online Supplementary File S1 Codelists Dutch for original codelists and Supplementary File S2 Codelists SNOMED-CT for translated codelists.

RESULTS

Development of the algorithm

Identification of locally advanced primary cSCC. We identified locally advanced primary tumors staged as T3 or T4 according to the American Joint Committee on Cancer, eighth edition (AJCC8) tumor classification using three criteria, all of which had to be met: (i) a hierarchical histopathological code from the Nationwide Pathology Registry (i.e., Nationwide Network and Registry of Histo- and Cytopathology [PALGA] code) indicating a primary cSCC combined with a PALGA code for skin or subcutis or a PALGA sublocalization code that is likely to be a cSCC; (ii) the absence of a PALGA localization code in the first position that is likely to be another type of squamous cell carcinoma (SCC) (mucosa, cheek, maxilla, mandible, larynx, floor of the mouth); and (iii) the presence of any of the following high-risk features within the free-text conclusion of the pathology report: T3 or T4, tumor diameter > 4 cm, invasion depth not precisely known but at a minimum of 5.5 mm or an exact invasion depth >6.0 mm, invasion beyond the subcutaneous fat, invasion in muscles, invasion in deep structures, bone erosion or invasion, nerve invasion ≥0.1 mm, any perineural invasion (excluding in nerves <0.1 mm), angioinvasive growth, and invasion depth reaching the bottom of the excision (Table 1). A few criteria, such as a minimum invasion depth of 5.5 mm and perineural invasion, angioinvasion, and invasion depth reaching the bottom of the excision, are

Table 2. Description of the 186 Patients with Locally Advanced Primary (Stage III), Recurrent, or Metastatic cSCC (Stage III/IV) and the 184 Patients with Stage I/II cSCC

186 Locally Advanced Primary, Recurrent, and Metastatic cSCC

Reason for Inclusion	Categories	n (%)
Locally advanced primary cSCC	Total	116 (62)
	T3 (AJCC8)	103 (89)
	T4 (AJCC8)	13 (11)
	T2b (BWH)	48 (41)
	T3 (BWH)	15 (13)
Recurrent tumor	Not classified within AJCC8/BWH	30 (16)
In-transit metastasis	Not classified within AJCC8/BWH	14 (8)
Histologically confirmed		12 (86)
Regional lymph node metastasis	N+	24 (13)
Histologically confirmed		22 (92)
Distant metastasis	M+	2 (1)
Histologically confirmed		1 (50)

184 Stage I/II cSCC

Reason for Inclusion	Categories	n (%)
T stage (AJCC8)	T1	160 (87)
	T2	24 (13)
T stage (BWH)	T1	149 (81)
	T2a	35 (19)
Extracted variables		
Differentiation	Well	59 (32)
	Moderate	102 (55)
	Poor	11 (6)
	Unknown/unreported	12 (7)
Vertical depth (mm)	<2 mm	68 (37)
	≥2 mm	79 (43)
	≥4 mm and <6 mm	13 (7)
	Bottom of biopsy	3 (2)
	Unknown/unreported	21 (11)
Clinical tumor diameter (cm)	< 2cm	149 (80)
	≥ 2cm and <4 cm	25 (15)
	Unknown/unreported	10 (5)

Abbreviations: AJCC8, American Joint Committee on Cancer, eight edition; BWH, Brigham and Women's Hospital; cSCC, cutaneous squamous cell carcinoma.

not official AJCC8 criteria but were included because these criteria increased the algorithm's sensitivity without a large decrease in the positive predictive value (PPV) in preliminary analyses (data not shown).

We identified locally advanced primary tumors staged as T2b or T3 according to the Brigham and Women's Hospital (BWH) alternative T-classification system by integrating three criteria: (i) PALGA code for primary cSCC combined with a PALGA code for skin or subcutis or PALGA sublocalization code that is likely to be a cSCC; (ii) the absence of PALGA localization code at the first position with a low likelihood of being a cSCC (mucosa, cheek, maxilla, mandible, larynx, floor of the mouth); and (iii) bone invasion or at least two of the following high-risk features within the pathology report's free-text conclusion: tumor diameter ≥2 cm, poor differentiation, perineural invasion in nerves ≥0.1 mm, invasion beyond the subcutaneous fat, invasion in deep structures, or invasion in muscles (Table 1).

Identification of recurrent cSCC. For the identification of recurrent cSCC, two criteria had to be met: (i) PALGA code for primary cSCC combined with a PALGA code for skin or subcutis or skin or subcutis in the free-text pathology conclusion and (ii) free text in the pathology conclusion indicating a recurrence (Table 1).

Identification of metastasis. We identified metastasis in three ways: (i) a PALGA code for metastatic SCC in combination with a PALGA code for skin or subcutis; (ii) PALGA code for primary SCC, metastatic SCC, metastatic carcinoma, or possible metastasis in combination with a PALGA code for parotid, salivary, or submandibular gland; and (iii) a free-text algorithm that identifies metastatic or malignant cells from the pathology conclusion in combination with squamous from the pathology conclusion or a primary SCC PALGA code (Table 1). Subsequently, we excluded pathology reports showing metastatic disease unlikely to have originated from an SCC of the skin (e.g., oral and pharyngeal cancers, lung cancer, or SCC of unknown origin) (Table 1). Codelists are provided in Supplementary File S1 (PALGA codes) and Supplementary File S2 (Systemized Nomenclature of Medicine — Clinical Terms [SNOMED-CT]).

Combined algorithm for advanced cSCC. All of the foregoing principles were included to enhance the algorithm's capabilities. This way, patients with multiple advanced cSCCs could still be identified even if one of their advanced cSCC reports was missing.

Validation of algorithm

Study population. We included 186 patients with advanced cSCC treated at the Erasmus MC Cancer Institute (Rotterdam, The Netherlands) between May 18, 2018 and October 9, 2020 (Table 2). The majority of patients had locally advanced primary cSCCs, of which 116 were classified as T3/T4 according to AJCC8, and 63 were classified as T2b/T3 according to BWH. There were 30 local recurrent tumors and 40 metastases. In addition, we included 184 patients treated at the Erasmus MC Cancer Institute between January 16, 2016 and September 23, 2020 with a T1/T2 cSCC according to AJCC8 and who were not T2b/T3 according to BWH (Table 2).

Measures of performance

Sensitivity. The algorithm correctly identified 171 of 186 patients with advanced cSCC, which resulted in an overall sensitivity of 91.9% (95% confidence interval [CI] = 88.0–95.9) (Table 3). The majority of false negatives were caused by clinically identified features of advanced tumors that were not described or seen during pathological assessment, such as a clinical diameter >4 cm or imaging-detected bone invasion. All false negatives are summarized in Supplementary Table S1. The sensitivity of the three subgroups was lower, ranging from 23.3% for recurrent cSCC to 52.3% for T2b/T3 (BWH) locally advanced primary cSCC, 70.0% for metastases, and finally 86.2% for T3/T4 (AJCC8) locally advanced primary cSCC. The sensitivity of the algorithm for locally

Table 3. Performance Measures of the Algorithm

	TP	TP + FN	Sensitivity (95% CI)	TN	TN + FP	Specificity (95% CI)	TP	TP + FP	PPV (95% CI)
All cases combined	171	186	91.9% (88.0–95.9)	178	184	96.7% (94.2–99.3)	277	353	78.5% (74.2–82.8)
Locally advanced primary tumor AJCC8 T3/T4	100	116	86.2% (79.9–92.5)	179	184	97.3% (94.9–99.6)	221	268	82.5% (77.9–87.0)
Locally advanced primary tumor BWH T2b/T3	34	65	52.3% (40.1–64.5)	184	184	100.0% (100.0–100.0)	79	95	83.2% (75.6–90.7)
Recurrent tumor	7	30	23.3% (8.2–38.5)	184	184	100.0% (100.0–100.0)	22	28	78.6% (63.3–93.8)
Metastases	28	40	70.0% (55.8–84.2)	183	184	99.5% (98.4–100.5)	83	133	62.4% (54.2–70.7)

Abbreviations: AJCC8, American Joint Committee on Cancer, eight edition; BWH, Brigham and Women's Hospital; CI, confidence interval; FN, false negative; FP, false positive; PPV, positive predictive value; TN, true negative; TP, true positive.

advanced primary cSCC was higher for AJCC8-based advanced tumors (86.2%, 95% CI = 79.9–92.5) than for BWH-based advanced tumors (52.3%, 95% CI = 40.1–64.5) because only one high-risk feature is required to identify T3/T4 (AJCC8) tumors, whereas the identification of at least two high-risk features or bone invasion is required for most T2b/T3 (BWH) tumors. Only 46% of T2b/T3 (BWH) tumors had more than two high-risk features identified from pathology reports, whereas 80% of T3/T4 (AJCC8) tumors had at least one of four high-risk features identified. Supplementary Table S2 includes a detailed summary of the type and number of features that could be extracted from pathology reports. Supplementary Table S3 shows the stratified sensitivity per subgroup when patients with multiple advanced cSCCs could be identified by any of their advanced cSCC reports.

Specificity. The algorithm falsely identified six patients as advanced cSCC among 184 patients with low-stage cSCC, whereas 178 were correctly categorized as low stage, resulting in a specificity of 96.7% (95% CI = 94.2–99.3) (Table 3). All false-positive cases are summarized in Supplementary Table S4. Stratified analysis revealed a specificity >97% for all subgroups.

PPV. Among the 353 patients who were categorized as advanced cSCC by the algorithm, 277 actually had T3/T4 (AJCC8) locally advanced primary (n = 221), T2b/T3 (BWH) locally advanced primary (n = 79), recurrent cSCC (n = 22), or metastatic cSCC (n = 83), resulting in an overall PPV of 78.5 (95% CI = 74.2–82.8). The PPV was highest among those identified as T2b/T3 (BWH) locally advanced primary cSCC (PPV = 83.2%, 95% CI = 75.6–90.7) and was lowest among those identified as metastatic cSCC (PPV = 62.4%, 95% CI = 54.2–70.7). The PPV for metastatic cSCC increased when only PALGA codes were used instead of PALGA codes with free text (PPV = 79.4, 95% CI = 69.4–89.4). However, this reduced the sensitivity for metastatic cSCC to 52.5% (95% CI = 37.0–68.0) with a specificity of 100% (data not shown). As a result, we used a combination of PALGA codes and free text for the final algorithm.

DISCUSSION

In this study, we have developed and validated a rule-based algorithm on the basis of hierarchical histopathological codes and free text that automatically identifies patients with advanced cSCC from pathology reports with a very favorable sensitivity of 91.9% and a specificity of 96.7%. The PPV or the percentage of all identified pathology reports that are certain advanced cSCC cases was almost 80% for all advanced cSCC combined. Such a high PPV is critical when the algorithm is used to identify patients with advanced cSCC for cancer registries or other observational studies to avoid reading too many patient files of low-risk patients and thereby wasting registration time.

The sensitivity of specific subgroups was lower. For example, if only the part of the algorithm to detect metastasis was used, 70% of all metastatic cSCC would be detected instead of more than 90%. The combined algorithm has a higher sensitivity because most patients with metastatic cSCC also have a pathology report for a locally advanced primary or recurrent cSCC and will be identified in this manner when a pathology report for metastasis is missing, for example, in the case of imaging-detected metastasis without histological confirmation. Thus, when this algorithm is used to assess the prevalence of specific subgroups of advanced cSCC, it should be taken into account that the stratified sensitivity was lower and that, for example, 30% of metastatic cSCC may have been missed. The stratified PPV was equally high for most subgroups, except for metastasis. Of all patients who were identified as having cSCC metastasis, 38% were false positives. Reasons for this included that the algorithm misidentified reports of patients with mucosal SCC metastasis or reports where it was reported to be unclear whether the tumor was a new primary cSCC or a skin metastasis but was in fact a new primary cSCC. Nevertheless, our algorithm is thought to save a huge amount of time. Even if the algorithm is only used to detect metastases, it still saves a lot of time compared with opening all the files of patients who develop cSCC every year.

Comparison with literature

Various computational techniques have been explored to extract cancer-related information from pathology free text (Spasić et al., 2014). The majority focused on colorectal, breast, prostate, and lung cancer (Buckley et al., 2012; Coden et al., 2009; Currie et al., 2006). To the best of our knowledge, this is the only automated pathology algorithm that concentrates on advanced cSCC. Eide et al. (2012) employed pathology reports' free-text retrieval capacity to identify the incidence of keratinocyte cancers to validate medical claims data algorithms but not (high-risk) cSCC in particular. Thompson et al. (2020) used supervised learning methods to build a web application that automatically extracts diagnostic information for keratinocyte cancers, such as (subtype) diagnosis and site, from free-text pathology. Their objective

was to estimate incidences accurately in the absence of nationwide registration, not to identify or extract cSCC high-risk features, particularly (Thompson et al., 2020).

Strengths and limitations

Strengths of the study include the manual registration of 186 patients with locally advanced primary (stage III), recurrent, or metastatic cSCC (stage III/IV) and 184 patients with stage I/II cSCC from the medical patient files of the Erasmus MC Cancer Institute. Because this dataset included patients with cSCC with clinically diagnosed advanced cSCC but no histological confirmation (e.g., imaging-detected bone invasion), this was critical for an accurate estimation of the algorithm's sensitivity. Furthermore, it was of vital importance to be able to retrieve the complete history of all pathology reports by linking them to a nationwide database of pathology reports (PALGA) to include primary cSCC of referred patients.

Information from pathology reports is complex to retrieve automatically because most reports are written in narrative format and because the pathologists' nomenclature for describing a diagnosis or lack thereof varies greatly between pathologists. The sensitivity of our algorithm could have been higher because of several high-risk features that were present during pathological assessment but were not reported, such as tumor diameter. Nationwide implementation of synoptic reporting for tumor characteristics would therefore greatly improve data quality and collection. Synoptic reporting is currently used in 29% of all cSCC pathology reports in the Netherlands, but more laboratories have agreed to use it in the near future (Swillens et al., 2019).

However, also in case of poor synoptic reporting rates, the algorithm can still identify patients with advanced cSCC from pathology reports accurately. Given that SNOMED-CT was reported to be utilized in over 50 countries in 2013 and that synoptic reporting is likely to grow in the future, we believe that our rule-based algorithm can be used globally after external validation (Lee et al., 2013).

Another obstacle that we encountered during the analysis was that the algorithm identified more patients with advanced cSCC than those we had initially included in our selection. All pathology reports of patients who were identified by the algorithm but not included in the sensitivity dataset were therefore manually reviewed. However, scoring a pathology report as a true positive was done in a conservative way. For example, if it was unclear from the pathology report whether it was a new primary cSCC or a skin metastasis, we included the report as false positive, whereas if we had had the clinical information, this may have been a true positive. This is likely to have resulted in an underestimation of the PPV. The algorithm has yet to be externally validated, which will require data from both a nationwide pathology registry as well as data from a single institution. To enable external validation and thereby increase its international applicability, the algorithm has been translated into corresponding international SNOMED-CT and English free text.

This study shows that patients with advanced cSCC can be accurately identified from pathology reports, allowing cost-effective–targeted surveillance of patients with advanced cSCC. Although external validation still has to take place, this rule-based algorithm opens up future large-scale epidemiological research on advanced cSCC.

MATERIAL AND METHODS

Definition of advanced cSCC

In this study, advanced cSCC was defined as locally advanced primary cSCC (either T3/T4 according to AJCC8 or T2b/T3 according to BWH), recurrent cSCC, or metastatic cSCC (skin, nodal, or distant metastasis). CSCC that had been staged according to the seventh edition of the American Joint Committee on Cancer were included if they fulfilled the AJCC8 T3/T4 criteria (i.e., T3/T4 or T1/T2 with perineural invasion, tumor depth >6 mm, invasion beyond subcutaneous fat, or minor bone erosion).

Study population and data sources

Sensitivity dataset. To determine sensitivity, we retrieved data on patients with advanced cSCC from the clinical patient files of the Erasmus MC Cancer Institute. These patients were identified by reviewing the records from the multidisciplinary skin cancer board meetings between May 18, 2018 and October 9, 2020, where all patients with advanced cSCC were discussed weekly. Subsequently, we retrieved all pathology reports that met the criteria related to cSCC (either primary, recurrent, or metastatic) from these patients from PALGA (see Supplementary Table S5) (Casparie et al., 2007). Pathology reports from other pathology laboratories were also included because patients may have been diagnosed with advanced cSCC in another hospital before being sent to the Erasmus MC Cancer Institute.

Specificity dataset. To determine the specificity of the algorithm, we selected a random sample of patients with low-stage/nonadvanced, stages I and II cSCC according to AJCC8 and who were not T2b/T3 according to BWH from the Erasmus MC Cancer Institute between January 16, 2016 and September 23, 2020. These patients were identified by reviewing all patient records with a Diagnostic Related Group code for skin cancer in combination with a specific diagnosis of cSCC. These patients were also linked to PALGA to retrieve the same selection of pathology reports as previously mentioned.

PPV dataset. To determine the PPV, we retrieved all pathology reports of cSCC in the Erasmus MC Cancer Institute from PALGA during the same time period. To identify all patients with cSCC on the basis of pathology reports, we applied the selection criteria presented in Supplementary Table S5. Thereafter, all cSCC-related pathology reports in the Erasmus MC Cancer Institute were retrieved using the same criteria as those used for patients in the sensitivity and specificity dataset (see Supplementary Table S5).

PALGA data

The data from PALGA included the report's conclusion, which was either free-text based or automatically generated if synoptic reporting was used. In addition to the conclusion, the pathologist assigned one or more diagnostic rules to each report as a standard, which consisted of a combination of diagnostic terms (localization, procedure, disease) from the PALGA thesaurus (https://www.palga.nl/palga-on-line-thesaurus.html). The diagnostic terms are automatically translated into one or more PALGA codes from a hierarchical coding system on the basis of SNOMED-CT, a well-established international terminology system that allows language-based data exchange both nationally and

internationally. Examples of PALGA free-text conclusions and diagnostic rules can be found in Supplementary Table S6. We provided the PALGA codes as well as the SNOMED-CT for our algorithm and translated Dutch free text into English to facilitate external validation, thereby increasing the international applicability of our rule-based algorithm.

Data extraction from Erasmus MC Cancer Institute medical files

Data from the medical files of patients with advanced cSCC included type of advanced cSCC (i.e., locally advanced, recurrent, and metastatic cSCC); tumor location; tumor diameter (cm); pathology features (e.g., tumor differentiation, invasion depth (mm); presence of invasion beyond subcutaneous fat; perineural invasion $\geq$0.1 mm; lymphovascular invasion and bone invasion; presence of in-transit, regional, or distant metastasis; date of pathology diagnosis; and pathology record number. Clinical factors, such as imaging-detected bone invasion, were also recorded.

For stage I/II cSCC, the following data were retrieved: tumor location, tumor diameter (cm), pathology features (tumor differentiation, invasion depth [mm]), date of pathology diagnosis, and pathology record number. Patients were excluded if invasion depth was unreported or if it reached the bottom of the biopsy unless an invasion depth >6 mm was thought to be very unlikely (e.g., a superficial biopsy of a tumor <1 cm in clinical diameter). Similarly, patients with an unreported clinical tumor diameter were only included in stage I/II selection if the postoperative defect size suggested that the tumor should have been <4 cm.

Statistical analyses

To calculate a specificity or sensitivity of 85% as a single proportion with a 95% CI between 80 and 90%, we aimed to include 193 advanced cSCC and 193 stage I/II cSCC. Patient and tumor characteristics were presented as means and proportions.

The sensitivity, specificity, and PPV of the algorithm with a 95% CI were calculated. Measures of performance were stratified by the type of advanced cSCC (i.e., locally advanced primary cSCC according to AJCC8 or BWH, recurrent cSCC, and metastatic cSCC). The algorithm was developed using SAS 9.1.3 (SAS Institute, Cary, NC). Descriptive statistics were used to characterize the study cohort and were performed using Statistical Package for the Social Sciences 25.0 statistical software (SPSS, Chicago, IL). This study was approved by the scientific committees of the Erasmus MC Cancer Institute (MEC-2020-0054), PALGA, and the Dutch Clinical Research Foundation (W20.048/NMWO20.02.007) and was conducted with waived informed consent.

Data availability statement

The data used to support the findings of this study are available from the Erasmus MC Cancer Institute and Nationwide Network and Registry of Histo- and Cytopathology, but they are under license and hence not publicly available. The authors can provide data on reasonable request and with permission from the Erasmus MC Cancer Institute and Nationwide Network and Registry of Histo- and Cytopathology.

ORCIDs

Celeste Eggermont: http://orcid.org/0000-0002-1669-0283
Marlies Wakkee: http://orcid.org/0000-0001-8578-901X
Annette Bruggink: http://orcid.org/0000-0003-1354-6988
Quirinus Voorham: http://orcid.org/0000-0002-2631-7976
Kay Schreuder: http://orcid.org/0000-0002-4224-800X
Marieke Louwman: http://orcid.org/0000-0001-9011-6741
Antien Mooyaart: http://orcid.org/0000-0001-9810-5780
Loes Hollestein: http://orcid.org/0000-0001-8922-6791

CONFLICT OF INTEREST

MW served on Sanofi Genzyme's advanced cutaneous squamous cell carcinoma advisory board and was compensated. MW and CE presented this study at a Sanofi Genzyme internal meeting and were compensated. The remaining authors state no conflicts of interest.

ACKNOWLEDGMENTS

We would like to thank Nadia Rbia for her help with reviewing the patient files and the extraction of the required variables. This investigator-initiated study received funding from Sanofi Genzyme.

AUTHOR CONTRIBUTIONS

Conceptualization: CE, MW, AB, RV, KS, ML, AM, LH; Funding Acquisition: MW, LH

Formal Analysis: CE, AG, LH; Supervision: MW, LH; Writing — Original Draft Preparation: CE, LH; Writing — Review and Editing: CE, MW, AB, RV, KS, ML, AM, LH

Disclaimer

The funders had no role in study design, data collection, analysis, or report writing.

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at www.jidonline.org, and at https://doi.org/10.1016/j.jid.2022.07.008

REFERENCES

Adalsteinsson JA, Olafsdottir E, Ratner D, Waldman R, Feng H, Ungar J, et al. Invasive and in situ squamous cell carcinoma of the skin: a nationwide study in Iceland. Br J Dermatol 2021;185:537–47.

Buckley JM, Coopey SB, Sharko J, Polubriaginof F, Drohan B, Belli AK, et al. The feasibility of using natural language processing to extract clinical information from breast pathology reports. J Pathol Inform 2012;3:23.

Casparie M, Tiebosch AT, Burger G, Blauwgeers H, van de Pol A, van Krieken JH, et al. Pathology databanking and biobanking in the Netherlands, a central role for PALGA, the nationwide histopathology and cytopathology data network and archive. Cell Oncol 2007;29:19–24.

Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, et al. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. J Biomed Inform 2009;42:937–49.

Currie AM, Fricke T, Gawne A, Johnston R, Liu J, Stein B. Automated extraction of free-text from pathology reports. AMIA Annu Symp Proc 2006;2006:899.

Eide MJ, Tuthill JM, Krajenta RJ, Jacobsen GR, Levine M, Johnson CC. Validation of claims data algorithms to identify nonmelanoma skin cancer. J Invest Dermatol 2012;132:2005–9.

Glaser AP, Jordan BJ, Cohen J, Desai A, Silberman P, Meeks JJ. Automated extraction of grade, stage, and quality information from transurethral resection of bladder tumor pathology reports using natural language processing. JCO Clin Cancer Inform 2018;2:1–8.

Guorgis G, Anderson CD, Lyth J, Falk M. Actinic keratosis diagnosis and increased risk of developing skin cancer: a 10-year cohort study of 17,651 patients in Sweden. Acta Derm Venereol 2020;100:adv00128.

Hanauer DA, Miela G, Chinnaiyan AM, Chang AE, Blayney DW. 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 2007;205:690–7.

Jouhet V, Defossez G, Burgun A, le Beux P, Levillain P, Ingrand P, et al. Automated classification of free-text pathology reports for registration of incident cases of cancer. Methods Inf Med 2012;51:242–51.

Lee D, Cornet R, Lau F, de Keizer N. A survey of SNOMED CT implementations. J Biomed Inform 2013;46:87–96.

Lomas A, Leonardi-Bee J, Bath-Hextall F. A systematic review of worldwide incidence of nonmelanoma skin cancer. Br J Dermatol 2012;166:1069–80.

Nguyen AN, Moore J, O'Dwyer J, Philpot S. Assessing the utility of automatic cancer registry notifications data extraction from free-text pathology reports. AMIA Annu Symp Proc 2015;2015:953–62.

Spasić I, Livsey J, Keane JA, Nenadić G. Text mining of cancer-related information: review of current status and future directions. Int J Med Inform 2014;83:605–23.

Stang A, Khil L, Kajüter H, Pandeya N, Schmults CD, Ruiz ES, et al. Incidence and mortality for cutaneous squamous cell carcinoma: comparison across three continents. J Eur Acad Dermatol Venereol 2019;33(Suppl 8):6–10.

Swillens JEM, Sluijter CE, Overbeek LIH, Nagtegaal ID, Hermens RPMG. Identification of barriers and facilitators in nationwide implementation of standardized structured reporting in pathology: a mixed method study. Virchows Arch 2019;475:551–61.

Thompson BS, Hardy S, Pandeya N, Dusingize JC, Green AC, Millane A, et al. Web application for the automated extraction of diagnosis and site from pathology reports for keratinocyte cancers. JCO Clin Cancer Inform 2020;4:711–23.

Tokez S, Hollestein L, Louwman M, Nijsten T, Wakkee M. Incidence of multiple vs first cutaneous squamous cell carcinoma on a nationwide scale and estimation of future incidences of cutaneous squamous cell carcinoma. JAMA Dermatol 2020;156:1300–6.

Tokez S, Wakkee M, Kan W, Venables ZC, Mooyaart AL, Louwman M, et al. Cumulative incidence and disease-specific survival of metastatic cutaneous squamous cell carcinoma: a nationwide cancer registry study. J Am Acad Dermatol 2022;86:331–8.

Venables ZC, Nijsten T, Wong KF, Autier P, Broggio J, Deas A, et al. Epidemiology of basal and cutaneous squamous cell carcinoma in the U.K. 2013–15: a cohort study. Br J Dermatol 2019;181:474–82.

Wehner MR. Underestimation of cutaneous squamous cell carcinoma incidence, even in cancer registries. JAMA Dermatol 2020;156:1290–1.

Supplementary Table S1. Description of the False Negatives

Patient	Reason for Inclusion (AJCC8)	Reason not Detected by Algorithm
76	T3/T4	The single high-risk factor was clinical tumor diameter (>4 cm), which was recorded in the patient file but not indicated in the pathology report.
88	T3/T4	Deep invasion clinically detected by imaging (not during pathological assessment).
96	T3/T4	The single high-risk factor was clinical tumor diameter (>4 cm), which was recorded in the patient file but not indicated in the pathology report
106	T3/T4	Perineural growth was seen during MMS, which was recorded in the patient file but not in the pathology report.
212	T3/T4	Bone invasion clinically detected by CT scan (not during pathological assessment).
225	T3/T4	Perineural growth was seen during MMS, which was recorded in the patient file but not in the pathology report.
245	T3/T4	Bone invasion clinically detected by CT scan (not during pathological assessment).
247	T3/T4	The single high-risk factor was clinical tumor diameter (>4 cm), which was recorded in the patient file but not indicated in the pathology report.
321	T3/T4	Perineural growth was seen during MMS, which was recorded in the patient file but not in the pathology report.
352	T3/T4	The single high-risk factor was a clinical tumor diameter (>4 cm), which was recorded in the patient file but not indicated in the pathology report
229	Recurrence	No mention of recurrence in the pathology report
367	Recurrence	No mention of recurrence in the pathology report.
315	Skin metastasis	The skin metastasis is described in the pathology report as a large-cell malignancy matching the SCC localization.
213	Lymph node metastasis	Detected by imaging, not by histopathology (FNA was inconclusive).
357	Lymph node metastasis	This pathology report was missing in our selection because it lacked a morphology code of SCC. Only a code for carcinoma was included.

Abbreviations: AJCC8, American Joint Committee on Cancer, eighth edition; CT, computed tomography; FNA, fine needle aspiration; MMS, Mohs micrographic surgery; SCC, squamous cell carcinoma.

Supplementary Table S2. Description of Type and Number of High-Risk Features Detected in Pathology Reports

High-Risk Features	Clinical Files with Primary cSCC Pathology Date (n)	Reported Risk Factors in Matched Pathology Report, n (%)
AJCC8	111	111
Diameter		
≤4 cm	18	
>4 cm	22	2 (10)
Unknown/unreported	71	
Invasion depth		
≤6 mm	44	
>6 mm	30	44 (147)[1]
Unknown/unreported	37	
Invasion beyond subcutaneous fat		
Yes	65	15 (23)
No	43	
Unknown/unreported	3	
Muscle invasion[2]		
Yes	31	20 (65)
No	78	
Unknown/unreported	2	
Bone erosion/bone invasion		
Yes	16	1 (6)
No	93	
Unknown/unreported	2	
Perineural invasion		
Yes	47	38 (81)
No	63	
Unknown/unreported	1	
Angioinvasion[3]		
Yes	11	3 (27)
No	97	
Unknown/unreported	3	
Any AJCC8 high-risk feature	100	80 (80)
BWH	64	64
Diameter		
≤2 cm	1	
>2 cm	27	3 (11)
Unknown/unreported	36	
Differentiation		
Well	6	
Moderate	31	
Poor	26	25 (96)
Unknown/unreported	1	
Perineural invasion		
Yes	34	27 (79)
No	30	
Unknown/unreported	0	
Invasion beyond subcutaneous fat[4]		
Yes	53	26 (49)
No	11	
Unknown/unreported	0	
Bone erosion/bone invasion		
Yes	14	1 (7)
No	50	

(*continued*)

Supplementary Table S2. Continued

High-Risk Features	Clinical Files with Primary cSCC Pathology Date (n)	Reported Risk Factors in Matched Pathology Report, n (%)
Number of risk factors		
0 or 1	3	36
≥2	61	28 (46)

Abbreviations: AJCC8, American Joint Committee on Cancer, eighth edition; BWH, Brigham and Women's Hospital; cSCC, cutaneous squamous cell carcinoma.

[1]There were an additional 14 reports because we included reports that specified invasion depth of at least 5.5 mm and also because there were a number of unknown invasion depths that may have been detected in other pathology laboratories.

[2]Muscle invasion is not listed as an official risk factor in the AJCC8, but it does imply that it occurs beyond the subcutaneous fat. We included this risk factor separately because it increased the sensitivity of the algorithm without a large decrease in the PPV.

[3]Angioinvasive growth is not an official risk factor in AJCC8, but it was included as a separate variable because it increased the algorithm's sensitivity without a large decrease in the PPV.

[4]Includes muscle invasion.

Supplementary Table S3. Performance Measures of the Algorithm Taking Multiple Advanced cSCC per Patient into Account

Subgroups	TP	TP + FN	Sensitivity (95% CI)
All cases combined	171	186	91.9% (88.0–95.9)
Locally advanced primary tumor AJCC8 T3/T4	106	116	91.4% (86.3–96.5)
Locally advanced primary tumor BWH T2b/T3	61	65	93.8% (88.0–99.7)
Recurrent tumor	27	30	90.0% (79.2–100.8)
Metastases	37	40	92.5% (84.3–100.7)

Abbreviations: AJCC8, American Joint Committee on Cancer; CI, confidence interval; cSCC, cutaneous squamous cell carcinoma; FN, false negative; TP, true positive; TN, true negative.

Supplementary Table S4. Description of the False Positives

Patient	Reason for Inclusion (AJCC8)	Reason Identified by the Algorithm
23	T1	Perineural growth
42	T1	Bottom excision
158	T1	Pathology report of primary tumor, which is described in the pathology report as a possible metastasis
207	T1	Error in conclusion, invasion depth of 1.6 cm within biopsy instead of 1.6 mm
279	T1	Bottom excision
43	T2	Re-excision of positive margins but described as recurrence
68	T2	Perineural growth

Abbreviation: AJCC8, American Joint Committee on Cancer, eighth edition.

Supplementary Table S5. Selection of Pathology Reports

Codes (i.e., any of these codes, OR)	Description per Code		Codes (i.e., Any of these Codes, OR)	Description per Code
M80513	All primary SCC	AND	T01_	Skin, any
M80523			T02_	
M80543				
M807_2				
M807_3				
M80704				
M80813				
OR				
M80513	All primary SCC	AND	TY_	Any topography code
M80523			T04_	Breast
M80543			T52_	Lip
M807_2			TXY_	Ear
M807_3			TXX_	Eyelid
M80704			T08_	Lymph node
M80813				
OR				
T08_	Any lymph nodes			
OR				
M8___6	Any metastasis			
M9___6				
OR				
T00060	Unknown localization of primary tumor			
OR				
M80009	Unknown primary or metastatic			

Abbreviation: SCC, squamous cell carcinoma.

Supplementary Table S6. Patient Selection for Positive Predictive Value

Codes (i.e., any of these Codes, OR)	Description per Code		Codes (i.e., Any of these Codes, OR)	Description per Code
M80513	All primary SCC	AND	T01_	Skin, any
M80523			T02_	
M80543				
M807_2				
M807_3				
M80704				
M80813				
OR				
M80513	All primary SCC	AND	TY_	Any topography code
M80523			T04_	Breast
M80543			T52_	Lip
M807_2			TXY_	Ear
M807_3			TXX_	Eyelid
M80704			T08_	Lymph node
M80813				

Abbreviation: SCC, squamous cell carcinoma.

Supplementary Table S7. Example of PALGA Free-Text Conclusion and PALGA Codes

Free Text of the Conclusion	PALGA Code 1[1]	PALGA Diagnosis 1 (Translation of PALGA Code 1)	PALGA Code 2	PALGA Diagnosis 2 (Translation of PALGA Code 2)
Lymph node puncture neck level 4 left: no malignancy.	T08000*TY0600*TYY990*P31430*M09450	lymph node*neck*left*puncture*no malignancy		
Left upper eyelid skin biopsy: moderately differentiated squamous cell carcinoma. Invasion depth at least 4.8 mm. There is perineural growth around nerve branches with a diameter >0.1 mm. No angioinvasive growth.	T01000*TXX810*TYY990*P11400*M80703*P30731	skin*eyelid*left*biopsy*squamous cell carcinoma		
"I: Skin excision (crown left): well-differentiated squamous cell carcinoma. Infiltration depth: 3.0 mm. Perineural growth absent. (lymph) angioinvasive growth absent. Deep structures invasion: absent. Cutting surfaces: free. TNM classification (7th edition): pT1. II: Skin excision (crown back left): Morbus Bowen. Cutting surfaces: not free. Positive cutting edge location: pointed end.	T01000*P11200*M80703*E99997*M09410	skin*excision*squamous cell carcinoma*local protocol*cut surfaces free	T01000*P11200*M80812*E99997*M09420	skin*excision* morbus bowen*local protocol*cut surfaces not free

Abbreviations: PALGA, Nationwide Network and Registry of Histo- and Cytopathology; TNM, tumor, node, metastasis.

The first letter of the PALGA code indicates the type of code: T = localization code; TY = sublocalization code; M = morphology code; P = procedure code.

[1]These reports are originally written in Dutch and have been translated into English.