Assessment of Deep Learning Using Nonimaging Information and Sequential Medical Records to Develop a Prediction Model for Nonmelanoma Skin Cancer

Hsiao-Han Wang, MD; Yu-Hsiang Wang; Chia-Wei Liang, MS; Yu-Chuan Li, MD, PhD

**IMPORTANCE** A prediction model for new-onset nonmelanoma skin cancer could enhance prevention measures, but few patient data–driven tools exist for more accurate prediction.

**OBJECTIVE** To use machine learning to develop a prediction model for incident nonmelanoma skin cancer based on large-scale, multidimensional, nonimaging medical information.

**DESIGN, SETTING, AND PARTICIPANTS** This study used a database comprising 2 million randomly sampled patients from the Taiwan National Health Insurance Research Database from January 1, 1999, to December 31, 2013. A total of 1829 patients with nonmelanoma skin cancer as their first diagnosed cancer and 7665 random controls without cancer were included in the analysis. A convolutional neural network, a deep learning approach, was used to develop a risk prediction model. This risk prediction model used 3-year clinical diagnostic information, medical records, and temporal-sequential information to predict the skin cancer risk of a given patient within the next year. Stepwise feature selection was also performed to investigate important and determining factors of the model. Statistical analysis was performed from November 1, 2016, to October 31, 2018.

**MAIN OUTCOMES AND MEASURES** Sensitivity, specificity, and area under the receiver operating characteristic (AUROC) curve were used to evaluate the performance of the models.

**RESULTS** A total of 1829 patients (923 women [50.5%] and 906 men [49.5%]; mean [SD] age, 65.3 [15.7] years) with nonmelanoma skin cancer and 7665 random controls without cancer (3951 women [51.5%] and 3714 men [48.4%]; mean [SD] age, 47.5 [17.3] years) were included in the analysis. The 1-year incident nonmelanoma skin cancer risk prediction model using sequential diagnostic information and drug prescription information as a time-incorporated feature matrix could attain an AUROC of 0.89 (95% CI, 0.87-0.91), with a mean (SD) sensitivity of 83.1% (3.5%) and mean (SD) specificity of 82.3% (4.1%). Carcinoma in situ of skin (AUROC, 0.867; −2.80% loss) and other chronic comorbidities (eg, degenerative osteopathy [AUROC, 0.872; −2.32% loss], hypertension [AUROC, 0.879; −1.53% loss], and chronic kidney insufficiency [AUROC, 0.879; −1.52% loss]) served as more discriminative factors for the prediction. Medications such as trazodone, acarbose, systemic antifungal agents, statins, nonsteroidal anti-inflammatory drugs, and thiazide diuretics were the top-ranking discriminative features in the model; each led to more than a 1% decrease of the AUROC when eliminated individually (eg, trazodone AUROC, 0.868; −2.67% reduction; acarbose AUROC, 0.870; −2.50 reduction; and systemic antifungal agents AUROC, 0.875; −1.99 reduction).

**CONCLUSIONS AND RELEVANCE** The findings of this study suggest that a risk prediction model may have potential predictive factors for nonmelanoma skin cancer. This model may help health care professionals target high-risk populations for more intensive skin cancer preventive methods.
Nonmelanoma skin cancer (NMSC), comprising squamous cell carcinoma and basal cell carcinoma, is the most common type of malignant neoplasm in white individuals.1 Incidence rates of NMSC could exceed 100 per 100 000 person-years in many fair-skinned populations around the world,2,3 while the NMSC incidence among Asian individuals is 2.3 to 9.2 per 100 000 population.4-5 However, the incidence of NMSC in Hispanic and Asian individuals has continued to increase.6 When skin cancer occurs in nonwhite individuals, it often presents at a more advanced stage, with elevated morbidity and mortality.7 Established risk factors for skin cancer have been reported in previous literature: UV radiation,8 family history of skin cancer,9 smoking,10,11 ionizing radiation,12 immunosuppressive status (eg, after organ transplantation),13 and use of photosensitizing drugs.12,13,14 These studies were mainly epidemiologic risk analyses,12-14 examining the association between individual factors and the risk of skin cancer.

Several skin cancer risk prediction models used for communicating risk factors have been proposed for risk stratification and assisting prevention interventions. These studies used mostly self-reported risk factors, limited personal information, and traditional statistical tools (eg, logistic regression) to develop risk prediction models, with the AUC (area under curve) performances ranging between 0.62 to 0.86 for melanoma.15-17 The AUC for NMSC in such prediction models18,19 was also scarce (0.72-0.85). Only demographics, UV-related covariates, skin cancer history, and lesion-oriented parameters had been used for prediction.18-21

In the era of data-driven health care, effective use of machine learning and big medical data could provide better and more personalized health care.22-25 Machine learning is an extension of traditional statistical approaches and manages high-dimensional data (including images and large-scale electronic medical records [EMRs]), contributing to disease diagnosis,26,27 prediction of disease development,21,28,29 and prognosis prediction.30,31 Convolutional neural network (CNN), a deep machine learning approach used mostly in medical image classification,35 achieved a diagnostic performance in classifying skin cancer with a level of competence comparable with that of dermatologists.36 Instead of training CNN with images to facilitate lesion detection, CNN has been applied for analyzing EMRs to predict chronic diseases.23

The aim of this study is to use CNN to develop a prediction model for incident NMSC, based on nonimaging and multidimensional medical information. In traditional studies, the time-dependent variables were either binary (eg, with or without diabetes) or summarized in a single value (eg, sum and mean). As in the real world, with medical events evolving over time, we intended to capture the time sequentiality of clinical visits, diagnoses, and medications in this study. Each individual’s input EMR would be converted into an image-like matrix with longitudinal medical events. This is to our knowledge a novel approach to develop an NMSC risk prediction model via deep learning, using clinical diagnostic and medical records with temporal continuity to predict the skin cancer risk of a given patient within the next year.

**Key Points**

**Question** Can deep learning develop a risk-prediction model without information on UV exposure or specific lesions to predict nonmelanoma skin cancer in an Asian population?

**Findings** A deep learning convolutional neural network was trained using nonimaging and sequential medical records to predict the development of nonmelanoma skin cancer. The area under the receiver operating characteristic curve for the model was 0.89; several potential determining factors for prediction such as history of precancerous lesions and use of some photosensitizing medications were also identified.

**Meaning** A machine learning algorithm can accurately identify individuals at high risk of nonmelanoma skin cancer and may help to optimize prevention interventions and adherence.

**Methods**

**Data Set**

This study analyzed 2 million Clinical Declaration files and In-Hospital Declaration files from the Taiwan National Health Insurance Research Database (NHIRD) from January 1, 1999, to December 31, 2013. The NHIRD contains enrollment files and original claims data for reimbursement as well as information on demographic characteristics, International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) diagnosis and procedure codes, and details of prescriptions. This was an all-Asian population data set. This study was approved by the Taipei Medical University Institutional Review Board, which waived informed patient consent because all patient records and information were anonymized and deidentified before the analysis.

**Study Population and Definitions**

Using the NHIRD, we identified patients aged 20 to 90 years who had at least 3 years of records of more than 1 outpatient visit or admission claim during 2002-2013. The population with cancer was validated not only by ICD-9-CM code but also by intervention codes (eg, skin biopsy, skin surgery, or subsequent radiotherapy or chemotherapy) and national catastrophic illness card (which required a physician’s certificate and definite pathologic findings of cancer) if qualified. Nonmelanoma skin cancer had to be the patients’ first cancer. The ratio of patients with NMSC to controls without cancer was 1:4 (1829 patients with verified NMSC and 7665 controls without cancer).

The index date was defined as the date of the first diagnosis of skin cancer (ICD-9-CM code 173 for nonmelanotic skin cancers, comprising squamous cell carcinoma and basal cell carcinoma). For the control group, the index date was either matched with the cancer index date or the last day available in the database. The observation window consisted of 3 years of data. We thus use the past 3 years of medical information of the patient to predict the risk of new-onset NMSC 1 year later.
sponded to different diagnoses in terms of ICD-9-CM codes and indicated either a clinical visit or a drug prescription at that day. The image-like medical input matrix was further divided into 18 submatrices (C1 to C17 indicates 17 ICD-9-CM classifications according to organ systems; V indicates V-code supplementary classification). The hidden boxes indicate the feature maps of the filters generated in a convolutional layer. After convolution, max pooling (MP), and fully connected (FC) layers, the output prediction is binary (cancer or noncancer).

**Prediction Model Construction and Evaluation**

Within the observation window of each patient, we used age, sex, ICD-9-CM diagnostic codes, World Health Organization-Anatomical Therapeutic Chemical (WHO-ATC) prescription codes, and the total numbers of clinical encounters found in that period to create features. We also used 1092 ICD-9-CM codes comprising 17 organ systems plus V-code supplementary classification (used when circumstances other than a disease or injury result in an encounter or are recorded by clinicians as problems or factors that influence care), and World Health Organization Anatomical Therapeutic Chemical prescription codes. The horizontal axis represents the temporal information. Each dot in the matrix indicates an ICD-9-CM code diagnosed at the corresponding clinical visit or a prescription at that day. The image-like medical input matrix was further divided into 18 submatrices (C1 to C17 indicates 17 ICD-9-CM classifications according to organ systems; V indicates V-code supplementary classification). The hidden boxes indicate the feature maps of the filters generated in a convolutional layer. After convolution, max pooling (MP), and fully connected (FC) layers, the output prediction is binary (cancer or noncancer).

**Results**

The mean (SD) age in the NMSC group was 65.3 (15.7) years, with 923 women (50.5%) and 906 men (49.5%) (Table 1). In the randomly sampled controls without cancer, the mean (SD) age was 47.5 (17.3) years, with 3951 women (51.5%) and 3714 men (48.4%). Each patient’s image-like medical input matrix (Figure 1) comprised 3 years of sequential medical events (ICD-9-CM codes in each clinical visit and prescription medications taken each day). The overall mean (SD) annual number of outpatient service visits was 13.4 (3.6). The numbers of annual ICD-9-CM codes multiplied by clinical visits were 26.8 in the controls without cancer and 43.4 in the NMSC group. The numbers of medications multiplied by prescription days per person per year were 388.0 in the controls without cancer and 863.4 in the NMSC group.
The CNN model with only ICD-9-CM diagnostic input information (model 1) yielded a mean (SD) AUROC of 0.820 (0.014) (Table 2). When taking only medications as input features (model 2), the AUROC reached 0.879 (0.012). The CNN prediction model with both ICD-9-CM and drug prescriptions (model 3) as input features achieved an AUROC of 0.894 (0.007) (Figure 2). Model 3 also achieved mean (SD) recall (sensitivity) of 83.1% (3.5) at precision (positive predictive value) of 0.571. The mean (SD) specificity reached 82.3% (4.1) when using both diagnostic and medication information. Also, the final risk probability score was between 0 (no disease) and 1 (disease). The risk score threshold was 0.166 in model 1, 0.288 in model 2, and 0.324 in model 3. As a sensitivity analysis, we further built models with case-control ratios as 1:6, 1:8, and 1:20. Multiple samplings with a ratio of 1:4 were performed as well. The AUROC performances were similar, ranging from 0.885 to 0.894 when using both diagnostic and drug information as inputs.

The features presented in Table 3 were identified through stepwise elimination of the diagnostic and drug variables, individually. The AUROC loss ranged from the minimum of −0.001% to a maximum of −2.80%. Variables with the most discriminative effects for the model to make prediction (>2% loss) and documented photosensitizing drugs with fair determining potential (>0.8% loss) are listed in Table 3. For other drugs reported to have an association with NMSC, such as voriconazole or metformin, the AUROC loss was too subtle (<0.01% loss) to affect the model performance. Table 3 showed that age was not the most important factor of the cancer prediction model, with an AUROC of 0.882 (−1.18% loss). Carcinoma in situ of skin (ICD-9-CM code 232; AUROC, 0.867; −2.80% loss) and other chronic comorbidities (eg, degenerative osteopathy [AUROC, 0.872; −2.32% loss], hypertension [AUROC, 0.879; −1.53% loss], and chronic kidney insufficiency [AUROC, 0.879; −1.52% loss]) served as more discriminative factors for the prediction. Medications such as trazodone, acarbose, systemic antifungal agents, statins, tricyclic antidepressants, nonsteroidal anti-inflammatory drugs, thiazide diuretics, and β-blockers were important factors, as the top-ranking discriminative features in the model (eg, trazodone AUROC, 0.868; −2.67% reduction; acarbose AUROC, 0.870; −2.50% reduction; and systemic antifungal agent AUROC, 0.875; −1.99% reduction). Other established photosensitizing drugs such as fluoroquinolones, trimethoprim-sulfamethoxazole, and antihypertensive drugs such as angiotensin-converting enzyme inhibitors, angiotensin receptor blockers, and calcium channel blockers were less decisive but still potentially clinically relevant predictors in the prediction model.

### Table 1. Demographics of the Sampled Data Set

| Characteristic | NMSC Group (n = 1829) | Control Group (n = 7665) |
|---------------|-----------------------|-------------------------|
| Age, mean (SD), y | 65.3 (15.7) | 47.5 (17.3) |
| Male, No. (%) | 906 (49.5) | 3714 (48.5) |
| Race/ethnicity | All Asian | All Asian |
| Total diagnosis counts, No. | 551 033 | 211 672 |
| Annual accumulation | | |
| ICD-9-CM counts per person, No. | 43.4 | 26.8 |
| Medications per person, No.* | 863.4 | 388.0 |

Abbreviations: ICD-9, International Classification of Diseases, Ninth Revision; Clinical Modification; NMSC, nonmelanoma skin cancer.

* Medications multiplied by prescription days.

**Discussion**

This NMSC risk prediction model developed via a deep learning approach appeared to demonstrate robust discrimination. The AUROC reached 0.894, with a sensitivity of 83.1% and specificity of 82.3%. This predictive model did not require information on traditional risk factors such as sun exposure, smoking, or family history of skin cancer. This model used diagnostic and drug data readily available in the EMR system. Other potential predictive parameters including comorbid chronic diseases and drugs were also identified. This machine learning–based NMSC prediction tool may facilitate determination of which patients are likely to develop NMSC, potentially allowing clinicians to intervene before disease advancement for high-risk patients, while sparing unneeded screening for low-risk individuals.

Our NMSC prediction model uses CNN, a deep learning approach, incorporating sequential multidimensional EMR data of each given patient. Detection models with lesion-oriented parameters for actinic keratosis and basal cell carcinoma had been reported but mainly for aiding diagnosis instead of mass screening.20 A research group in Australia reported 10 predictors identified through logistic regression and built a prediction model with an acceptable AUC of 0.80. The predictors were demographic and UV-related factors, along with history of skin cancer. However, among those without a history of skin cancer, the result for the prediction was an AUC of 0.72.18 Wang et al20 had developed a prediction tool to provide a probability of developing cutaneous squamous cell carcinoma within the next 3 years for non-Hispanic white individuals, with an AUROC of 85%. The prediction tool also used history of skin cancer and inherited covariates. Our current study may provide better prediction for incident NMSC. Roffman et al31 also used machine learning for NMSC prediction, based on personal health survey data. However, a traditional artificial neural network was trained with only 13 limited demographic and comorbidity parameters to predict NMSC. These parameters were mostly binary (yes or no) one-time inputs with no sequentiality or severity information. Therefore, the study yielded an AUROC of only 0.81, sensitivities of 86.2% to 88.5%, and specificities of 62.2% to 62.7%, with specificities 10% to 20% inferior to our models and sensitivities less than 10% better than ours. All our models produced better AUROCs (0.82-0.89; Table 2).

Electronic medical record data or electronic health record data are systematic collections of longitudinal patient health information. Electronic health records provide opportunities for developing and refining risk prediction algorithms.2,4 Machine learning approaches are recognized as tools for constructing electronic health record–based risk models to select
important predictors automatically from numerous features and manage random errors.\textsuperscript{24,28} By using electronic medical records and machine learning, incident hypertension,\textsuperscript{28} and new chronic kidney disease\textsuperscript{29} could be accurately predicted. The 1-year incident hypertension risk model\textsuperscript{28} using XGboost classification tree attained an AUROC of 0.917 and 0.870 in the retrospective and prospective cohorts. This real-time predictive analytic model\textsuperscript{28} has even been deployed in the state of Maine to target high-risk populations and to tailor the treatments solutions. The present use of CNN on EMR data for cancer prediction appears to be novel and effective. By using deep learning, we analyzed more than 50 million sets of real-life medical data.\textsuperscript{23,34} Convolutional neural network has been more commonly used on image data and for the classification of skin neoplasms.\textsuperscript{26,27,35} Cheng et al\textsuperscript{23} first built CNN models using electronic health record matrices to predict disease development. Patient data were represented as time-embedded matrices. Instead of extracting single values from time series, this method kept the continuous temporal associations of clinical events. With the smart CNN structure, the algorithm can extract important features and weigh them automatically in the prediction phase. In the study by Cheng et al,\textsuperscript{23} the AUROC performance of predicting the onset of congestive heart failure was 0.767 and of predicting chronic obstructive pulmonary disease was 0.738. By adding extensive and sequential drug information, an AUROC of 0.89 could be accomplished in our model.

The final risk probability score was between 0 (no disease) and 1 (disease). The thresholds in the current study ranged from 0.1 to 0.43 using different models. The risk score threshold was determined by obtaining the maximal sum of sensitivity and specificity. If an arbitrary threshold of 0.5 was used, sensitivity would be diminished, leaving out more suspicious cases. Ye et al\textsuperscript{28} stratified the prediction scores for incident hypertension into 5 risk categories, with a score greater than 0.2 as the high-risk category and those greater than 0.4 as the very-high-risk group. Another value to be questioned is the prediction window. Our study used 1-year risk prediction. Predicting half-year, 2-year, or 3-year risk of developing NMSC could also be done. One-year prediction was not too far away to raise concerns, nor too near for proper clinical actions to be initiated.

Most of the discriminative factors identified in this study for model prediction could be consistent with the previous literature. All classes of antihypertensives have been reported to be associated with phototoxic and/or photoallergic cutaneous reactions and increases in the risk of skin cancer risk in Table 2. Performances of Convolutional Neural Network With Different Input Features

| Input Features                  | Mean (SD) | AUROC | Sensitivity, % | Specificity, % | Precision |
|--------------------------------|-----------|-------|----------------|----------------|-----------|
| Diagnosis only                 | 0.820 (0.014) | 79.6 (3.1) | 71.2 (2.7)    | 0.443 (0.039) |
| Medication only                | 0.879 (0.012) | 81.7 (2.6) | 78.4 (4.9)    | 0.488 (0.128) |
| Diagnosis and medication       | 0.894 (0.007) | 81.1 (3.5) | 82.3 (4.1)    | 0.571 (0.049) |

Table 3. Important Features for Prediction of Nonmelanoma Skin Cancer

| Feature                                      | AUROC (% Decrease) ^a |
|----------------------------------------------|------------------------|
| All features included                        | 0.894                  |
| Age                                          | 0.882 (–1.18)          |
| Comorbidities                                |                        |
| Carcinoma in situ of skin                   | 0.867 (–2.80)          |
| Degenerative osteopathy                     | 0.872 (–2.32)          |
| Hypertension                                | 0.879 (–1.53)          |
| Chronic kidney insufficiency                | 0.879 (–1.52)          |
| Cirrhosis                                    | 0.881 (–1.40)          |
| Chronic obstructive pulmonary disease        | 0.884 (–1.02)          |
| Medications                                  |                        |
| Trazodone                                    | 0.868 (–2.67)          |
| Acarbose                                     | 0.870 (–2.50)          |
| Systemic antifungal agents (griseofulvin or terbinafine) | 0.875 (–1.99)          |
| Statins                                      | 0.877 (–1.78)          |
| Tricyclic antidepressants                    | 0.879 (–1.54)          |
| NSAIDs                                       | 0.881 (–1.17)          |
| Thiazide diuretics                           | 0.884 (–1.14)          |
| ß-blockers                                   | 0.886 (–0.90)          |
| Fluoroquinolones                             | 0.886 (–0.89)          |
| ACE inhibitors and ARBs                     | 0.886 (–0.89)          |
| Trimethoprim-sulfamethoxazole                | 0.886 (–0.85)          |
| Calcium channel blocker (diltiazem)          | 0.887 (–0.78)          |

Abbreviations: ACE, angiotensin-converting enzyme; ARB, angiotensin receptor blocker; AUROC, area under the receiver operating characteristic curve; NSAIDs, nonsteroidal anti-inflammatory drugs.

^a The AUROC decrease of the convolutional neural network when a feature was removed.

Figure 2. Area Under the Receiver Operating Characteristic Curve (AUROC) of the Model Using Sequential Diagnostic and Medication Information
long-term users. A meta-analysis showed associations between skin cancer and calcium channel blockers and β-blockers; thiazide diuretics, angiotensin-converting enzyme inhibitors and angiotensin receptor blockers showed no significant associations. However, a recent study from Denmark reported an association between use of hydrochlorothiazide and increased risk of NMSC, especially for squamous cell carcinoma. Metformin has been reported to be associated in a dose-response manner with decreases in the risk of skin cancer (including melanoma and NMSC) in patients with type 2 diabetes. Longer duration of statin use was also associated with higher risk of NMSC, especially basal cell carcinoma in men. Metformin was not a decisive factor in our prediction model, which was built for the general population, while another antidiabetic agent, acarbose, was a more significant factor (Table 3). Use of nonsteroidal anti-inflammatory drugs demonstrated a trend toward lower risk of NMSC, especially in women. Psychiatric treatments, such as trazodone and tri cyclic antidepressants, and antimicrobial agents, including quinolones and antifungal agents, have been documented as photosensitizing drugs. Chronic diseases such as osteoarthritis, hypertension, and chronic obstructive pulmonary disease (Table 3) could be proxy measures for aging, smoking, occupational factors, and even UV irradiation.

Skin cancer is complex, with many factors associated with its development. Instead of focusing on a single risk factor, the present CNN prediction model could weigh all diagnoses and medications to make an accurate prediction. However, the results in Table 3 could not provide statistical inference as regression models, designed for finding associations (eg, increased or decreased risk) between variables and outcomes.

Our study used clinical visits, diagnostic information, and prescriptions along with sequential temporal information. Future studies may further incorporate procedures, laboratory test results, radiographic test results, pathologic test reports, medical expenditures, and accessible personal health information. This predictive analytic model could be deployed in health care and hospital information systems, which may help health care professionals to target high-risk populations and optimize prevention strategies. For instance, the model could be applied with My Health Bank, an online government-launched program providing each individual with his or her medical records under Taiwanese National Health Insurance. Personal risk of skin cancer can be annually assessed based on previous 3-year health data and educational guidelines (eg, sun protection and lifestyle modifications) can also be enhanced.

Limitations
This study has several limitations. The NHIRD did not include the following information: lifestyles, exposure (eg, occupation or sunlight), family history, genetic parameters, and laboratory data. Information on the types, pathologic characteristics, grading, and staging of NMSC was not available. Therefore, separate prediction of basal cell carcinoma and squamous cell carcinoma could not be performed. The limited ethnicity of the study population is also a critical limitation. However, by using nonimaging variables, this model still holds the potential to be generalizable beyond the Taiwanese population. Further investigation on different thresholds or modified models under the same concept for other ethnic groups will be necessary before generalization to different ethnic groups. Further external validation using more recent NHIRD data sets or other Asian databases can be performed.

Conclusions
We developed a risk prediction model using sequential diagnostic and medication informatics to predict NMSC. Several discriminative factors for prediction, such as history of precancerous lesions and use of some photosensitizing medications, were also identified. This predictive analytic model may help health care professionals to target high-risk populations and optimize prevention strategies. However, this innovative and clinically relevant prediction model, as a proof of concept at this stage, still requires further validation efforts. In addition, the feasibility and cost-effectiveness of the system requires more clinical testing before it can be deployed into clinical practice.
