



Article

Validation and Improvement of a Convolutional Neural Network to Predict the Involved Pathology in a Head and Neck Surgery Cohort

Dorian Culié ^{1,2,*}, Renaud Schiappa ², Sara Contu ², Boris Scheller ^{1,2}, Agathe Villarme ¹, Olivier Dassonville ¹, Gilles Poissonnet ¹, Alexandre Bozec ^{1,2} and Emmanuel Chamorey ²

¹ Head and Neck Surgery Department, Antoine Laccassagne Center, 06100 Nice, France

² Epidemiology, Biostatistics and Health Data Department, Antoine Laccassagne Center, 06100 Nice, France

* Correspondence: dorian.culie@nice.unicancer.fr

Abstract: The selection of patients for the constitution of a cohort is a major issue for clinical research (prospective studies and retrospective studies in real life). Our objective was to validate in real life conditions the use of a Deep Learning process based on a neural network, for the classification of patients according to the pathology involved in a head and neck surgery department. 24,434 Electronic Health Records (EHR) from the first visit between 2000 and 2020 were extracted. More than 6000 EHR were manually classified in ten groups of interest according to the reason for consultation with a clinical relevance. A convolutional neural network (TensorFlow, previously reported by Hsu et al.) was then used to predict the group of patients based on their pathology, using two levels of classification based on clinically relevant criteria. On the first and second level of classification, macro-average performances were: 0.95, 0.83, 0.85, 0.97, 0.84 and 0.93, 0.76, 0.83, 0.96, 0.79 for accuracy, recall, precision, specificity and F1-score versus accuracy, recall and precision of 0.580, 580 and 0.582 for Hsu et al., respectively. We validated this model to predict the pathology involved and to constitute clinically relevant cohorts in a tertiary hospital. This model did not require a preprocessing stage, was used in French and showed equivalent or better performances than other already published techniques.

Keywords: neural network; cohort constitution; patient classification



Citation: Culié, D.; Schiappa, R.; Contu, S.; Scheller, B.; Villarme, A.; Dassonville, O.; Poissonnet, G.; Bozec, A.; Chamorey, E. Validation and Improvement of a Convolutional Neural Network to Predict the Involved Pathology in a Head and Neck Surgery Cohort. *Int. J. Environ. Res. Public Health* **2022**, *19*, 12200. <https://doi.org/10.3390/ijerph191912200>

Academic Editor: Rodney P. Jones

Received: 30 August 2022

Accepted: 22 September 2022

Published: 26 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Artificial intelligence (AI) is currently a major area of interest and research, particularly in the medical field and in oncology. Some teams have developed tools to assist in the diagnosis, choice of treatment and prognosis [1,2]. However, the establishment of cohorts of patient with a clinical interest, to build solid databases, is a mandatory prerequisite for any AI analysis.

Building patient cohorts around defined pathologies or phenotypes is a major challenge [3] in real word medical activity. Those cohorts, in addition to serving as input for AI algorithms, are also very useful to improve clinical trial recruitment [4,5], outcome prediction, survival analysis and carrying out real life and retrospective studies. They can also provide a reflection of the activity of a medical team [6,7]. The identification of patients from the beginning of their care for the construction of homogeneous cohorts is therefore a major issue, a request from medical teams and a lack in current clinical practice.

Different techniques for identifying patients of interest have been described, but a simple and rapid tool is urgently needed [8]. Other approaches have been developed to predict patient phenotypes by ICD codes (International Statistical Classification of Disease and Related Health Problems). Ferrao et al. [9] developed a data mining approach to support clinical coding in the ICD-9. In their study, they compared different approaches to predict ICD-9 codes. Venkataraman et al. [10] also compared different approaches to determine ICD-9 codes automatically in a cohort of human and veterinary records. They

trained long short-term memory (LSTM) recurrent neural networks (RNNs) and compared them with decision tree and random forest techniques [11]. Tam et al. developed a high-quality procedure based on combined processes and on structured and unstructured data to create clinically defined EHR-derived cohorts. This process was initially built to identify patients with acute coronary syndrome (ACS) in patients in a large general hospital [12]. A review of the literature [13] summarizes all published approaches to automatically identify patients based on their phenotype. It lists different possibilities to perform automatic classification: natural language processing (NLP)-based techniques, rule-based systems, statistical analyses, data mining, machine learning and combined techniques. It thus appears that rule-based systems are gradually being abandoned in favor of AI approaches. In this review, no single AI technique was superior, and all had similar efficiencies. All also had their own weaknesses. Furthermore, none of these previously published techniques has been validated in the literature, such publications are needed before considering the use of these tools in routine clinical practice.

To our knowledge, none of these reported studies has been validated, either for the choice of technique, external reproducibility by other teams, on a different medical theme or in a different language. Convolutional Neural Networks have already been used to classify different aspects in medical fields, but few reports are available for patient identification based on the first medical report.

Head and neck oncology surgery involves the management of many different pathologies, and therefore of different patients. The grouping of these patients according to these subgroups is therefore an essential step before initiating any research project.

Our aim was to describe, evaluate and valid a rapid and useful automated process for identifying and classifying patients into clinically relevant groups, based on their reason for first consultation, in a head and neck oncology surgery department.

2. Materials and Methods

An Electronic Health Records (EHR) is a multimedia file that can be electronically stored, transmitted and managed. EHR of a medical visit is a text dictated by the surgeon to transcribe the interview. It includes the reason for the visit, the patient's complaints, the examinations carried out and the clinical examination of the patient. The data used in this study were text files of the EHR of the first visit. All files were free text in a French language. The files were previously pseudonymized for the patient and surgeon names.

Head and neck oncology surgery in our institution covers different fields. It includes thyroid and parathyroid gland surgery, skin cancer surgery, salivary gland surgery, upper aero-digestive tract (UADT) cancer management. UADT cancers include oral cavity, hypopharyngeal, laryngeal, oropharyngeal, nasopharyngeal and nasal cavity cancers (and all cervical masses).

EHR of all patients who had a consultation with one of the surgeons of the head and neck surgery department between 2000 and 2020 were extracted.

EHR of the first visit were then selected. A subset of the selected EHR was randomly chosen and manually categorized by expert (head and neck surgeons) into 10 clinically relevant group based on the reason for consultation:

- "Thyroid and parathyroid pathology": thyroid nodules, thyroid pathology (*Basedow* and *Hashimoto's* disease), parathyroid dysfunction;
- "Salivary gland pathology": major salivary gland or accessory salivary gland tumor;
- "Head and neck skin pathology": squamous cell carcinoma, basocellular carcinoma or melanoma located in the head and neck area;
- "Oral cavity": oral cavity cancer;
- "Hypopharynx/larynx": hypopharyngeal cancer, laryngeal cancer;
- "Oropharynx": oropharyngeal cancer;
- "Nasopharynx": nasopharyngeal cancer;
- "Isolated cervical lymph nodes": cervical nodes with unknown primary;
- "Nasal cavity and sinuses": nasal cavity and paranasal sinuses;

- “Other”: all the other reasons for the consultation.

Those EHR with manually labelled group constituted the initial dataset for this study.

We did not perform any preprocessing step. The entire project was done in Python. We used the similar deep learning model described by Hsu et al. [14]. Briefly, a deep neural network was built using the Keras Python deep learning library running on top of the TensorFlow deep learning framework. All hyperparameters and codes are available in the publication by Hsu et al. [14]. In short, the convolutional layer is preceded by a word embedding and tokenizer layer. The filter size was set to 5×400 , and the stride was set to 1. As it is compulsory to define a priori a modality of activation of each neuron, we chose the commonly used ReLu (Rectified Linear Unit). When training our model, we used Adam Optimizer, a parameter of the Keras algorithm that is an alternative optimization algorithm that provides more efficient neural network weights by running repeated cycles of adaptive moment estimation, with a moderate learning rate to avoid overfitting. An epoch of 30 was used to give the model sufficient time to complete convergence.

The dataset was randomly divided in a training set of 80% of the dataset and the remaining 20% for the test set.

The model was used in two steps for classification of EHR to get as close as possible to human reasoning (Figure 1). In a first step of classification, the model classified EHR according to five categories: “thyroid and parathyroid pathology”, “salivary gland pathology”, “head and neck skin pathology”, “upper aerodigestive tract (UADT) pathology” and “other”. The UADT group, although it corresponds to a well-established clinical entity, actually includes several anatomical locations. We hence applied a second level of classification in six subgroups was performed: “oral cavity”, “hypopharynx/larynx”, “oropharynx”, “nasopharynx”, “isolated cervical lymph nodes”, “nasal cavity and sinuses”.

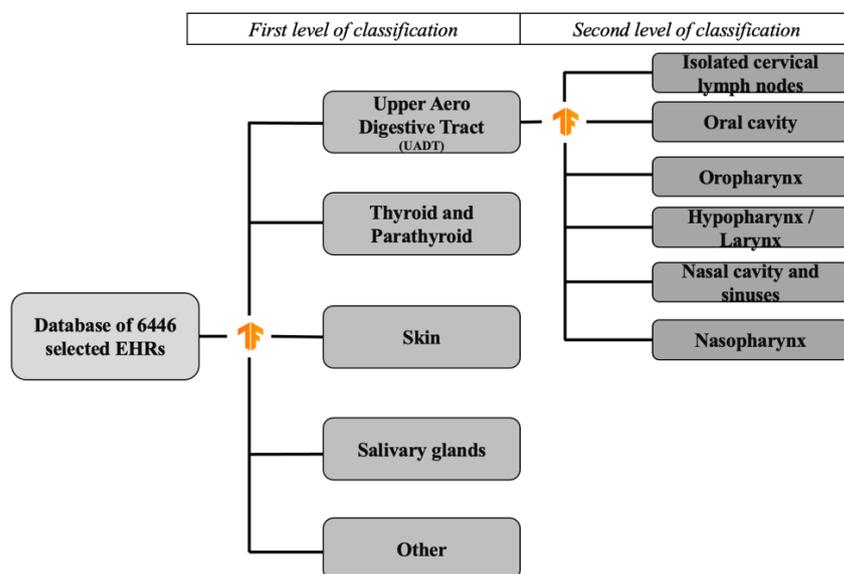


Figure 1. Flow chart of the two levels of classification.

For each group, we analyzed performance according to the metrics defined below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{F1.score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{precision} + \text{recall}}$$

where TP is True Positive (TP: the model correctly predicts the label of interest), TN is the True Negative (TN: the model correctly predicts another class than the label of interest), FP is the False Positive (FP: the model incorrectly predicts the label of interest), FN is the False Negative (FN: the model incorrectly predicts another label than the label of interest). We used Confusion matrices to visualize the distribution of patients in the predicted groups based on the actual group (ground-truth labels in ordinate versus model prediction in abscissa).

Based on retrospective and pseudonymized data collection, this study complies with the French law corresponding to the Methodology of Reference 004 (MR004) for clinical research. A declaration of the database has been made to the French health data hub (N°F20211007110223) and information was given to patients before the start of the study. The study was conducted in compliance with the Helsinki Declaration.

3. Results

3.1. EHR Extraction and Manual Classification

107,282 head and neck oncology surgery consultation EHR were extracted. From these reports, 24,434 first visits were identified and 6446 (26%) were randomly selected and manually classified by an expert (Head and Neck surgeon) in the ten clinically relevant groups as reported in Table 1.

Table 1. Distribution of patients according to groups in the initial dataset.

Group	Number of Patients (%)
Thyroid and parathyroid pathology	2509 (38.92)
Salivary gland pathology	283 (4.39)
Head and neck skin pathology	841 (13.04)
Oral cavity	618 (9.58)
Hypopharynx/larynx	659 (10.22)
Oropharynx	431 (6.68)
Nasopharynx	38 (0.58)
Isolated cervical lymph nodes	363 (5.63)
Nasal cavity and sinuses	66 (1.02)
Other	638 (9.89)
<i>Total</i>	<i>6446 (100)</i>

3.2. First Classification Level

We first applied the model on EHR of the first consultation of 6446 patients. The algorithm trained on 5157 EHR (80%) and then tested on 1289 patients (20%).

Performances of this first level of classification are shown in Table 2 and Figure 2. Accuracy, recall, precision, specificity and F1-score of this step were 0.95, 0.83, 0.85, 0.97 and 0.84, respectively, for the macro-average. The best performance was observed in the "Thyroid and parathyroid" group (accuracy, recall, precision, specificity and F1-score were 0.96, 0.96, 0.95, 0.97, 0.96, respectively). The worst performance was observed in the "other" group (accuracy, recall, precision, specificity and F1-score were 0.93, 0.58, 0.70, 0.97, 0.63, respectively).

Table 2. Algorithm performance for first classification level.

Groups N (%)	UADT	Thyroid and Parathyroid Pathology	Head and Neck Skin Pathology	Other	Salivary Gland Pathology	Macro-Average
	438 (33.4)	538 (41.0)	174 (13.3)	111 (8.5)	51 (3.9)	1312 (100)
Accuracy	0.91	0.96	0.98	0.93	0.98	0.95
Recall	0.89	0.96	0.92	0.58	0.78	0.83
Precision	0.84	0.95	0.93	0.70	0.84	0.85
Specificity	0.92	0.97	0.99	0.97	0.99	0.97
F1-score	0.87	0.96	0.93	0.63	0.81	0.84

UADT: upper aerodigestive tract.

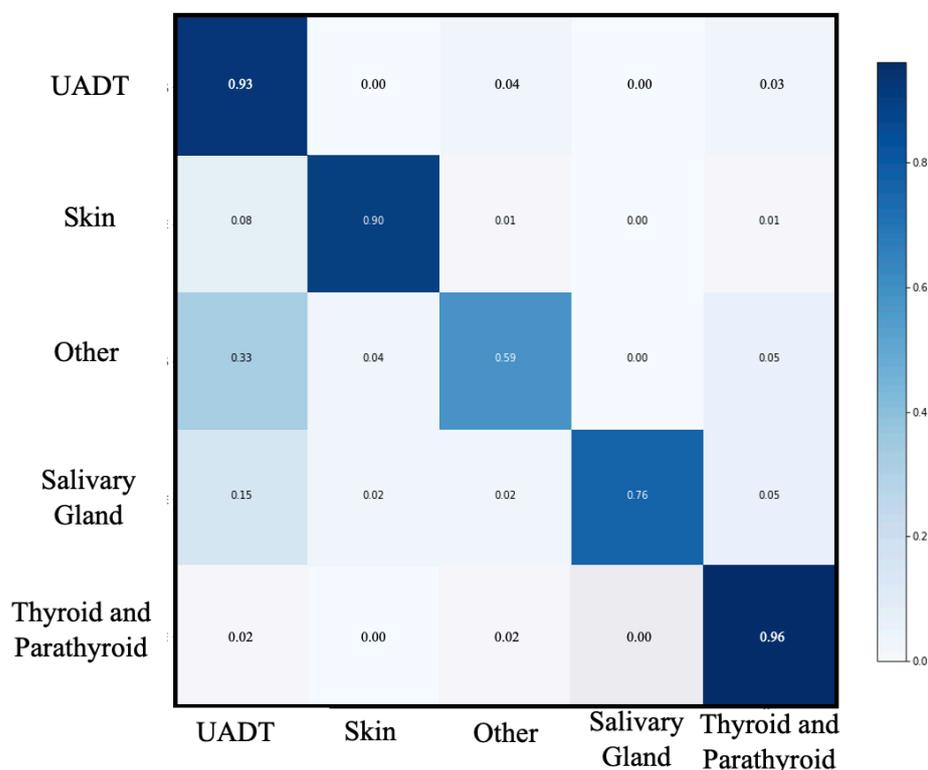


Figure 2. Confusion matrix of first level classification. UADT: upper aerodigestive tract.

3.3. Second Classification Level

We applied the same algorithm on the “UADT” group of 2175 patients. The algorithm trained on 1740 EHR (80%) and then tested on 435 patients (20%).

Performances of this second level of classification are shown in Table 3 and Figure 3. Accuracy, recall, precision, specificity and F1-score of this step were 0.93, 0.76, 0.83, 0.96 and 0.79, respectively for the macro-average. The best performance was observed in the “Hypopharynx and larynx” group (accuracy, recall, precision, specificity and F1-score were 0.91, 0.91, 0.82, 0.92, 0.86, respectively). The worst performance was observed in the “nasopharynx” group (accuracy, recall, precision, specificity and F1-score were 0.99, 0.63, 0.83, 1, 0.71, respectively).

Table 3. Algorithm performance for second classification level.

Groups N (%)	Isolated Cervical Lymph Nodes	Oral Cavity	Oro-Pharynx	Hypopharynx /Larynx	Nasal Cavity and Sinuses	Naso- Pharynx	Macro- Average
	71 (16.3)	120 (27.6)	88 (22.2)	142 (32.6)	8 (1.8)	6 (1.4)	435 (100)
Accuracy	0.93	0.90	0.88	0.91	0.99	0.99	0.93
Recall	0.77	0.81	0.70	0.91	0.73	0.63	0.76
Precision	0.77	0.84	0.73	0.82	1	0.83	0.83
Specificity	0.96	0.94	0.93	0.92	1	1	0.96
F1-score	0.77	0.82	0.72	0.86	0.84	0.71	0.79

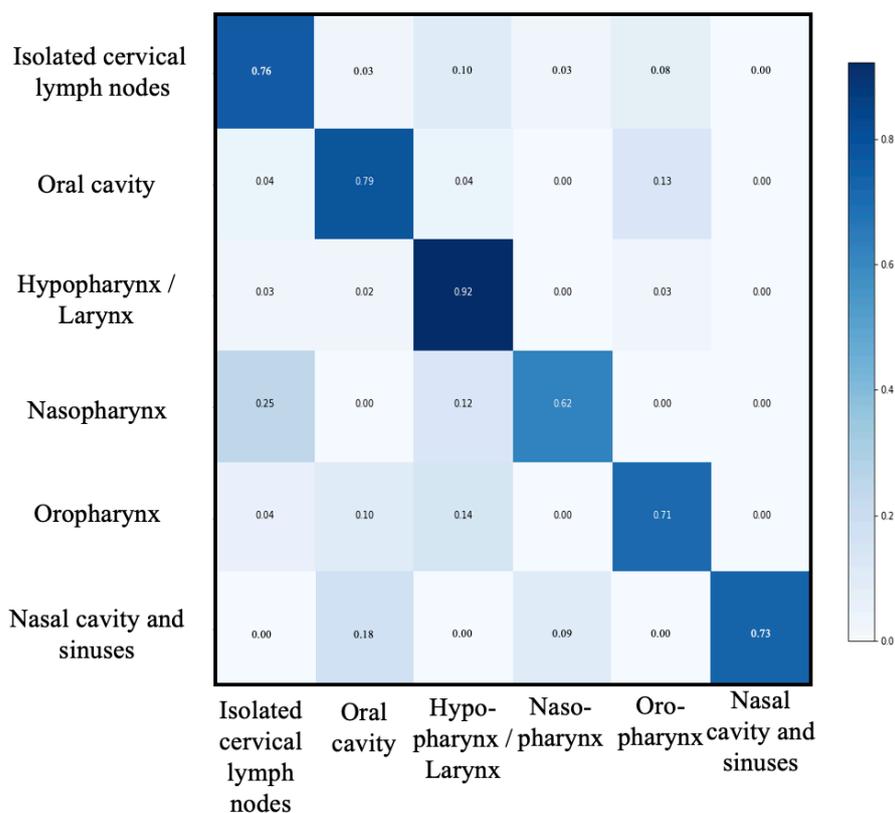


Figure 3. Confusion matrix of second level classification. UADT: upper aerodigestive tract.

3.4. Extension of the Process

This classification process was then applied to the rest of the cohort of 17 988 patients and the classification results were as follow:

- "Thyroid and parathyroid pathology": 6798 patients (37.79%)
- "Salivary gland pathology": 525 patients (2.91%)
- "Head and neck skin pathology": 2448 patients (13.60%)
- "other": 1230 patients (6.84%)
- "UADT": 6987 patients (38.84%)
- "Oral cavity": 2083 patients (29.81%)
- "hypopharynx/larynx": 2311 patients (33.05%)
- "oropharynx": 1379 patients (19.73%)
- "nasopharynx": 48 patients (0.69%)
- "Isolated cervical lymph nodes": 1062 patients (15.19%)
- "Nasal cavity and sinuses": 104 patients (1.48%)

4. Discussion

Many different pathologies are involved in the same medical department, so the constitution of patient cohorts with clinical consistency is of major interest. Indeed, the patient cohorts identified are useful for the inclusion of prospective therapeutic trials, for retrospective and real-life studies, and for building local data hubs for clinical or medico-economic evaluation [4,6].

A similar approach with an algorithm based on TensorFlow, an open AI resource, was recently described by Hsu et al. [14]. They applied a convolution neural network model to predict the International Statistical Classification of Disease and Related Health Problems (ICD-9) on subjective component of the progress note of 168,000 medical records of a single center. Previous studies have demonstrated that ICD classification have several limitations and is not sufficient for cohort constitution [15–17]. Indeed, for many medical conditions, less than 10% of the affected individuals' EHR contain the respective International Classification of Diseases (ICD) diagnosis codes [18,19]. ICD can also be miscoded at the time of data analysis or done after data extraction. In fact, diagnostic coding accuracy also varies by setting, provider type, and whether the code was assigned by a billing specialist [20–23]. Miscoding would lead to measurement error and missing data would contribute to selection bias and neutralize the statistical power available by mining the data contained in HER [12,24–27].

Therefore, we based our classification on clinically relevant criteria to constitute more homogeneous cohorts according to phenotype. Our classification was based on surgeons' notes, as physicians' notes are valuable sources of patient information [28,29].

In this study, we only applied this deep learning approach on first outpatient records to better identify the treated pathology. In addition, we obtained better results than Hsu et al. [14] whose best performance yielded an accuracy of 0.580, a recall of 0.580, and a precision of 0.582 vs. 0.95, 0.83 and 0.85, respectively for macro-average and 0.96, 0.96 and 0.95 for the best group ("Thyroid and parathyroid pathology") of the first level of classification in our study. The performance of our model was particularly high for the most represented pathology (thyroid pathology) in the first level of classification, explained by a larger training set. Moreover, in the second level of classification, the model achieved acceptable performance in a group where the pathologies involved are anatomically very similar. This classification is however very useful for the data analysis of these patients because even if they can be grouped within the same entity, each of these cancers has a different diagnostic, prognostic and therapeutic management. In addition, we compared the prediction with manually reviewed and labelled data by medical experts, which is considered as the 'ground truth' by most of the studies reporting performance on data [13].

Our study provides additional validation in a language other than English [14]. We first applied the process to determine five major and clinically relevant categories. We then used output classification of the UADT group as an input for a second TensorFlow process to determine subcategories in this group. The lower number of output categories gave better performances than a one-step classification.

The data mining approach of Ferrao et al. [9] concluded that logistic regression models had higher performances, while decision trees had a higher precision (lower rate of false positive) and support vector machines a higher recall (lower rate of false negative). Regardless of technique, published performances were poorer for predicting ICD-9 codes than our process to predict the relevant pathology.

Venkataraman et al. [10] approach, based on LSTM RNNs found that the "neoplasia" category gave the best performances among the different categories, explaining our good results as our patients were all in this category. Second, their LSTM RNNs approach had the best performances, with the highest F1-score obtained of 0.91. As a reminder, the F1 score is a score used to evaluate categorization tools in artificial intelligence that takes into account recall and precision. The best result of this study was in line with ours, even if our approach did not include a metatmap processing, a clinical natural language processing tool.

The authors compared their model to another AI approach, although the gold standard remains a comparison with human-controlled data [11].

For Tam et al. process, it was initially constructed to identify patients with acute coronary syndrome (ACS) in patients in a large general hospital. This model had a high sensitivity but contrary to our model, did not have a second manual or computerized control to confirm the ACS. Manual classification of EHR by medical experts for the learning and testing sets was a strength of our study. In addition, the Tam et al. [12] approach was useful for identifying the phenotype of patients in a primary care medical center. However, this model does not seem suitable for predicting a classification of all patients in a department of a second or tertiary medical center. Our model effectively classified patients with similar phenotypes among patients in the head and neck surgery department.

Hsu et al. [14] had to perform a preprocessing step to standardize Chinese and English text, modify punctuation and perform segmentation. Chiaramello et al. [30] had to perform an English translation from Italian clinical notes with Google Traduction before using the MetaMap steps (tokenization, parsing, variant generation, candidate retrieval, candidate evaluation, and mapping construction). Our approach had the advantage of not requiring a preprocessing step that can be time consuming and lead to errors. EHR were directly and manually classified by medical experts based on the specified pathology to form the input database.

The constitution of patient cohorts based on phenotype and only on the first medical examination could constitute a limitation of our study. Nevertheless, as a tertiary medical center, the locations of neoplasia are most often determined. Otherwise, patients were classified in the “other” group. This group therefore consisted of patients for whom medical expertise was required (patients without tumors) or with an undetermined neoplasia location. Another limitation of this study is that it predicts only one class of pathology involved. Some patients may present with two or more pathologies at the first visit. This is also a limitation for manual classification because our goal was to build cohorts of patients, and only one pathology was labeled for each patient. Even if this event is rare, another neural network could be used to identify patients with multiple cancers at this stage.

The major strength of our study is the open access and ease of use of the resource, which makes this model widely reproducible. Indeed, many authors use AI without a deep understanding of its mathematical, statistical and programming principles [31]. We therefore used this previously reported methodology to validate it in a different center and language than Hsu et al. [14], which can now be easily reproduced elsewhere, and for several different pathologies. The manual classification of EHR for training and validation sets was also a strength of our model, with a limited working time for this step (two days for a head and neck surgeon). This work constitutes an essential first step in the construction of any health data hub. This tool can thus be the first brick of a real time patient data processing. Other research work is necessary to continue the exploitation of the information contained in medical reports in order to extract and structure them.

5. Conclusions

We report here a reliable model based on the TensorFlow algorithm, a convolutional neural network with an open resource access, used to build phenotype-based patient cohorts in the head and neck surgery department of a tertiary cancer medical center, without using ICD codes. We report good performances of this process with macro-average performance on the first and second levels of classification of 0.95, 0.83, 0.85, 0.97, 0.84 and 0.93, 0.76, 0.83, 0.96, 0.79 for accuracy, recall, precision, specificity and F1-score. We have therefore validated this previously described approach with a better performance, in another language and without any preprocessing step. This approach represents an easy-to-use tool based on supervised deep learning. In addition, its main limitation was the need of a large initial dataset, with manually reviewed EHR by a medical expert. This model can now be used to build data hubs, to facilitate inclusion in clinical trials, to perform epidemiological and real-life studies, to assist administrative workers for business

accounting, and can be applied to other AI models as an automatic data extraction tool. This process should then be applied to all the pathologies treated in our center in order to build real time patient cohort.

Author Contributions: Conceptualization, D.C. and R.S.; methodology, D.C., R.S. and E.C.; software, R.S. and S.C.; validation, S.C., A.B. and E.C.; formal analysis, D.C.; data curation, D.C., A.V., B.S., O.D. and G.P.; writing—original draft preparation, D.C.; writing—review and editing, R.S., S.C., A.B. and E.C.; visualization, A.V., B.S., O.D. and G.P.; project administration, R.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Antoine Lacassagne Center (protocol F20211007110223, 10 July 2021).

Informed Consent Statement: This study complies with the French law corresponding to the Methodology of Reference 004 (MR004) for clinical research. Information was given to patients before the start of the study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, J.; Wu, J.; Zhao, Z.; Zhang, Q.; Shao, J.; Wang, C.; Qiu, Z.; Li, W. Artificial intelligence-assisted decision making for prognosis and drug efficacy prediction in lung cancer patients: A narrative review. *J. Thorac. Dis.* **2021**, *13*, 7021–7033. [[CrossRef](#)]
2. Li, D.; Pehrson, L.M.; Lauridsen, C.A.; Tøttrup, L.; Fraccaro, M.; Elliott, D.; Zając, H.D.; Darkner, S.; Carlsen, J.F.; Nielsen, N.B. The added effect of artificial intelligence on physicians' performance in detecting thoracic pathologies on CT and chest X-ray: A systematic review. *Diagnosics* **2021**, *11*, 2206. [[CrossRef](#)]
3. Kho, A.N.; Pacheco, J.A.; Peissig, P.L.; Rasmussen, L.; Newton, K.M.; Weston, N.; Crane, P.K.; Pathak, J.; Chute, C.G.; Bielinski, S.J.; et al. Electronic Medical Records for Genetic Research: Results of the eMERGE Consortium. *Sci. Transl. Med.* **2011**, *3*, 79re1. [[CrossRef](#)] [[PubMed](#)]
4. Hassanzadeh, H.; Karimi, S.; Nguyen, A. Matching patients to clinical trials using semantically enriched document representation. *J. Biomed. Inform.* **2020**, *105*, 103406. [[CrossRef](#)] [[PubMed](#)]
5. Spasic, I.; Krzeminski, D.; Corcoran, P.; Balinsky, A. Cohort Selection for Clinical Trials from Longitudinal Patient Records: Text Mining Approach. *JMIR Med. Inform.* **2019**, *7*, e15980. [[CrossRef](#)]
6. Mathias, J.S.; Gossett, D.; Baker, D.W. Use of electronic health record data to evaluate overuse of cervical cancer screening. *J. Am. Med. Inform. Assoc.* **2012**, *19*, e96. [[CrossRef](#)]
7. Strom, B.L.; Schinnar, R.; Jones, J.; Bilker, W.B.; Weiner, M.G.; Hennessy, S.; Leonard, C.E.; Cronholm, P.F.; Pifer, E. Detecting pregnancy use of non-hormonal category X medications in electronic medical records. *J. Am. Med. Inform. Assoc.* **2011**, *18* (Suppl. S1), 81–86. [[CrossRef](#)] [[PubMed](#)]
8. Peissig, P.L.; Santos Costa, V.; Caldwell, M.D.; Rottschreit, C.; Berg, R.L.; Mendonca, E.A.; Page, D. Relational machine learning for electronic health record-driven phenotyping. *J. Biomed. Inform.* **2014**, *52*, 260–270. [[CrossRef](#)] [[PubMed](#)]
9. Ferrão, J.C.; Oliveira, M.D.; Janela, F.; Martins, H.M.G.; Gartner, D. Can structured EHR data support clinical coding? A data mining approach. *Health Syst.* **2020**, *10*, 138–161. [[CrossRef](#)] [[PubMed](#)]
10. Venkataraman, G.R.; Pineda, A.L.; Bear Don't Walk, O.J.; Zehnder, A.M.; Ayyar, S.; Page, R.L. FasTag: Automatic text classification of unstructured medical narratives. *PLoS ONE* **2020**, *15*, e0234647. [[CrossRef](#)] [[PubMed](#)]
11. Schuemie, M.J.; Sen, E.; 't Jong, G.W.; Van Soest, E.M.; Sturkenboom, M.C.; Kors, J.A. Automating classification of free-text electronic health records for epidemiological studies. *Pharmacoepidemiol. Drug Saf.* **2012**, *21*, 651–658. [[CrossRef](#)]
12. Tam, C.S.; Gullick, J.; Saavedra, A.; Vernon, S.T.; Figtree, G.A.; Chow, C.K.; Cretikos, M.; Morris, R.W.; William, M.; Morris, J.; et al. Combining structured and unstructured data in EMRs to create clinically-defined EMR-derived cohorts. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 91. [[CrossRef](#)] [[PubMed](#)]
13. Shivade, C.; Raghavan, P.; Fosler-Lussier, E.; Embi, P.J.; Elhadad, N.; Johnson, S.B.; Lai, A.M. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 221–230. [[CrossRef](#)] [[PubMed](#)]
14. Hsu, J.L.; Hsu, T.J.; Hsieh, C.H.; Singaravelan, A. Applying Convolutional Neural Networks to Predict the ICD-9 Codes of Medical Records. *Sensors* **2020**, *20*, 7116. [[CrossRef](#)] [[PubMed](#)]
15. Singh, J.A.; Holmgren, A.R.; Noorbaloochi, S. Accuracy of Veterans Administration databases for a diagnosis of rheumatoid arthritis. *Arthritis Rheum.* **2004**, *51*, 952–957. [[CrossRef](#)] [[PubMed](#)]
16. Kandula, S.; Zeng-Treitler, Q.; Chen, L.; Salomon, W.L.; Bray, B.E. A bootstrapping algorithm to improve cohort identification using structured data. *J. Biomed. Inform.* **2011**, *44* (Suppl. S1), S63–S68. [[CrossRef](#)]

17. Perry, T.; Zha, H.; Oster, M.E.; Frias, P.A.; Braunstein, M. Utility of a Clinical Support Tool for Outpatient Evaluation of Pediatric Chest Pain. *AMIA Annu. Symp. Proc.* **2012**, *2012*, 726. Available online: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3540476/> (accessed on 3 March 2022). [[PubMed](#)]
18. Callahan, A.; Shah, N.H.; Chen, J.H. Research and Reporting Considerations for Observational Studies Using Electronic Health Record Data. *Ann. Intern. Med.* **2020**, *172* (Suppl. S11), S79. [[CrossRef](#)] [[PubMed](#)]
19. Wei, W.Q.; Teixeira, P.L.; Mo, H.; Cronin, R.M.; Warner, J.L.; Denny, J.C. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J. Am. Med. Inform. Assoc.* **2016**, *23*, e20. [[CrossRef](#)] [[PubMed](#)]
20. Fisher, E.S.; Whaley, F.S.; Krushat, W.M.; Malenka, D.J.; Fleming, C.; Baron, J.A.; Hsia, D.C. The accuracy of Medicare’s hospital claims data: Progress has been made, but problems remain. *Am. J. Public Health.* **1992**, *82*, 243–248. [[CrossRef](#)]
21. Reker, D.M.; Hamilton, B.B.; Duncan, P.W.; Yeh, S.C.J.; Rosen, A. Stroke: Who’s counting what? *J. Rehabil. Res. Dev.* **2001**, *38*, 281–289. Available online: <https://pubmed.ncbi.nlm.nih.gov/11392661/> (accessed on 3 March 2022). [[PubMed](#)]
22. Chescheir, N.; Meints, L. Prospective study of coding practices for cesarean deliveries. *Obstet. Gynecol.* **2009**, *114 Pt 1*, 217–223. [[CrossRef](#)] [[PubMed](#)]
23. Al Achkar, M.; Kengeri-Srikantiah, S.; Yamane, B.M.; Villasmil, J.; Busha, M.E.; Gebke, K.B. Billing by residents and attending physicians in family medicine: The effects of the provider, patient, and visit factors. *BMC Med. Educ.* **2018**, *18*, 136. [[CrossRef](#)] [[PubMed](#)]
24. Xu, H.; Fu, Z.; Shah, A.; Chen, Y.; Peterson, N.B.; Chen, Q.; Mani, S.; Levy, M.A.; Dai, Q.; Denny, J.C. Extracting and Integrating Data from Entire Electronic Health Records for Detecting Colorectal Cancer Cases. *AMIA Annu. Symp. Proc.* **2011**, *2011*, 1564. Available online: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243156/> (accessed on 4 March 2022).
25. Fernández-Breis, J.T.; Maldonado, J.A.; Marcos, M.; Legaz-García, M.D.C.; Moner, D.; Torres-Sospedra, J.; Esteban-Gil, A.; Martínez-Salvador, B.; Robles, M. Leveraging electronic healthcare record standards and semantic web technologies for the identification of patient cohorts. *J. Am. Med. Inform. Assoc.* **2013**, *20*, e288–e296. [[CrossRef](#)]
26. Virani, S.S.; Akeroyd, J.M.; Ahmed, S.T.; Krittanawong, C.; Martin, L.A.; Slagle, S.; Gobbel, G.T.; Matheny, M.E.; Ballantyne, C.M.; Petersen, L.A.; et al. The Use of Structured Data Elements to Identify ASCVD Patients with Statin-Associated Side Effects: Insights from the Department of Veterans Affairs. *J. Clin. Lipidol.* **2019**, *13*, 797. [[CrossRef](#)]
27. Ford, E.; Carroll, J.A.; Smith, H.E.; Scott, D.; Cassell, J.A. Extracting information from the text of electronic medical records to improve case detection: A systematic review. *J. Am. Med. Inform. Assoc.* **2016**, *23*, 1007–1015. [[CrossRef](#)]
28. Li, L.; Chase, H.S.; Patel, C.O.; Friedman, C.; Weng, C. Comparing ICD9-Encoded Diagnoses and NLP-Processed Discharge Summaries for Clinical Trials Pre-Screening: A Case Study. *AMIA Annu. Symp. Proc.* **2008**, *2008*, 404. Available online: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2656007/> (accessed on 4 March 2022).
29. Friedman, C.; Shagina, L.; Lussier, Y.; Hripcsak, G. Automated encoding of clinical documents based on natural language processing. *J. Am. Med. Inform. Assoc.* **2004**, *11*, 392–402. [[CrossRef](#)]
30. Chiaramello, E.; Pincioli, F.; Bonalumi, A.; Caroli, A.; Tognola, G. Use of “off-the-shelf” information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes. *J. Biomed. Inform.* **2016**, *63*, 22–32. [[CrossRef](#)]
31. Faes, L.; Wagner, S.K.; Fu, D.J.; Jack, D.; Liu, X.; Korot, E.; Ledsam, J.R.; Back, T.; Chopra, R.; Pontikos, N.; et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: A feasibility study. *Lancet Digit. Health* **2019**, *1*, e232–e242. [[CrossRef](#)]