공학석사 학위논문

흉부 방사선 영상에서 환자 확인 오류를 자동 검출하기 위한 지도 대조 학습을 통한 환자 유사도 측정 기법에 대한 연구

A Study on Patient Similarity Measurement with Supervised Contrastive Learning for Preventing Patient Misidentification in Chest Radiographs

울산대학교 대학원

의 과 학 과

김 지 영

흉부 방사선 영상에서 환자 확인 오류를 자동 검출하기 위한 지도 대조 학습을 통한 환자 유사도 측정 기법에 대한 연구

지도교수 김남국

이 논문을 공학석사 학위논문으로 제출함

2024 년 2 월

울 산 대 학 교 대 학 원
의 과 학 과
김 지 영

김지영의 공학 석사학위 논문을 인준함

심사위원 홍 길 선 (인)

심사위원 김 남 국 (인)

심사위원 이 준 구 (인)

울 산 대 학 교 대 학 원

2024 년 2 월

Abstract

Backgrounds: Accurate patient identification is important for patient safety and accurate diagnosis and treatment. Although reports of patient identification errors are rare in clinical settings, even a single error in patient identification can lead to fatal consequences such as incorrect treatment and surgery on the wrong patient. Therefore, the goal of this study is to propose a deep learning model that automatically detects patient identification errors in paired chest radiographs (CXRs), evaluate it, and compare and analyze it with experts.

Methods: We developed deep learning models for patient identification using a dataset of 240,004 CXRs. These models were validated using multiple datasets, including internal, CheXpert, and Chest ImaGenome datasets, which include different populations. We assessed the model's performance in terms of its ability to discern changes in disease status. To evaluate the model, we compared its performance to that of three junior radiology residents (Group I), two senior radiology residents (Group II), and two board-certified expert radiologists (Group III) in identifying patients from paired CXRs. Additionally, we conducted a one-sided non-inferiority test with a margin of 0.05 to further assess the model's performance.

Results: Our similarity-based deep learning model, SimChest, showed the most impressive patient identification performance across various datasets, regardless of changes in disease status (internal validation [area under the receiver operating characteristic curve (AUC) range 0.992–0.999], CheXpert [0.933–0.948], and Chest ImaGenome [0.949–0.951]). In comparison, radiologists demonstrated an average accuracy of 0.900 (with a 95% confidence interval ranging from 0.852-0.948) in identifying patients from paired CXRs, with performance levels increasing in accordance with experience (mean accuracy: Group I [0.874], Group II [0.904], Group III [0.935] and SimChest [0.904]). SimChest's performance was found to be non-inferior to the average performance of radiologists, with a P-value for non-inferiority of 0.015.

Conclusion: The results of this diagnostic study suggest that deep learning models can effectively identify misidentification using a pair of CXRs and perform at a level that is not significantly inferior to that of human experts.

차 례

표 및 그림 차례

Introduction

Accurate patient verification is essential for patient safety and to ensure accurate diagnosis and treatment. Although patient misidentification is reported to occur infrequently in real-world practice [1, 2], a single incident of misidentification can lead to catastrophic consequences, including misdiagnosis, incorrect treatment, wrong-site surgery, and wrong-patient surgery [3]. To prevent patient misidentification errors, all team members involved in the procedure and the patient confirm the identity of the patient and the procedure during the common "time-out" process [4]. Despite these efforts, Henneman [5] reported that only 61% of healthcare workers detected the misidentification error in simulation settings once it had occurred. Therefore, the use of additional patient verification methods, other than human-involved procedures, may be beneficial in reducing the risk of medical errors.

Several studies have shown that computerized algorithms can be used to recognize and identify patients. Jeon et al. found that a facial recognition algorithm could improve patient verification [6]. In addition, Silverstein et al. used a commercial facial recognition algorithm to identify individual patients [7], while Ampamya et al. presented decent unique patient matching performance with an open-source facial recognition system [8]. However, using facial recognition to screen for misidentification errors has several drawbacks. It may violate patient privacy, because facial photographs are not typically collected as part of routine practice. In addition, the integrity of facial images in medical settings can be easily compromised by factors such as painful expressions or facial trauma. Chest radiographs (CXRs), one of the most commonly taken medical images, can be used to screen for misidentification errors [9]. Morishita et al. used histogram correlation values of CXR [10], while Koa et al. used anatomical features to identify individual patients [11]. However, these methods using CXRs can also often fail to identify patients because variations in the CXRs, such as pose, breath-hold level, and gross abnormalities can confuse the models.

Recent evidence has shown that deep learning (DL) models can accurately predict demographic factors from medical images, such as age [12], sex [13], and race [14], which

were long believed to be nearly impossible for physicians. These findings suggest that DL models can directly recognize individual patient identities. Packhäuser et al. showed that a DL algorithm has a potential to verify and re-identify patients more accurately than existing computerized algorithms [15]. As the famous phrase “To err is human [16]” suggests, catastrophic misidentification errors do occur, although they are rare. Given that human systems sometimes fail to detect identification errors [5] and human performance may deteriorate due to fatigue, DL-based human identification can play an important role in screening for misidentification errors.

In this study, we aimed to investigate the ability of DL models to precisely screen misidentified patients. Therefore, we developed a similarity-based DL algorithm for patient identification using CXR, and validated it with various datasets from different centers, races, and clinical conditions. We also compared the algorithm’s performance with that of human experts and applied this model to another task such as image-based patient retrieval [17].

Materials and Methods

Definition of change/no-change status of disease in a CXR pair

The disease change-status expressed in a pair of CXRs can affect the performance of a patient identification model. Therefore, the change/no-change status of each pair of CXRs was labelled to evaluate the model’s performance according to this status. The pairs were labelled as follows: change was labelled using (increas*, decreas*, new*, improve*, aggravate*, progress*, resolv*, disappear*, heal*, and enlarge*) and the no-change was labelled using (no interval, no significant, and no remarkable) keywords in the radiologic reports. Through the visual validation by an expert thoracic radiologist, the change/no-change labelling accuracy obtained from radiology report was approximately 80% in the randomly sampled data from training dataset.

Data collection

The institutional review board (IRB) approved the study design and the requirement for written informed consent was waived due to the retrospective nature of the study. CXRs collected at the radiology department of a tertiary hospital in South Korea between 2011 and 2018 were included. Considering the model size and resources, 240K pairs of CXRs from 54,556 subjects were randomly sampled, including 120K each for change/no-change pairs.

We randomly sampled 240 pairs of CXRs from the internal dataset to assess the clinical comparability of our deep learning model with human experts. The dataset used for the reader study included 180 same (correctly identified) subjects (11 and 169 pairs respectively for each change/no-change) and 60 different (misidentified) subjects, the prevalence of which was determined by an expert radiologist with 20 years' experience, considering that patient mismatches are less likely to occur in real-world settings. Change/no-change labels and position of subjects (PA, posterior-anterior and AP, anterior-posterior) labels were determined by a radiology resident with two years' experience, who was not involved in the performance assessment.

One internal validation and two external validation datasets were used to evaluate the model's performance and robustness in screening for misidentified patients. The internal validation dataset consisted of 290 pairs of the same subjects (157 and 133 pairs for each change/no-change), and 290 different subjects, all of which were not seen in the training. CheXpert [18] and the gold standard Chest-ImaGenome [19] (CIG) dataset were used for the external validation. A total of 200 pairs of CXR of the same subjects (100 pairs for each change/no-change) and 200 pairs of different subjects were randomly sampled from the CheXpert. Finally, 290 pairs of same subjects (174 and 116 pairs for each change/no-change) of all gold standard dataset of CIG were used and 290 different subjects were randomly sampled. For each dataset for retrieval, 200 same pairs were randomly sampled from the internal validation dataset, CheXpert dataset, and CIG dataset of patient identification dataset. Because the identification dataset of CheXpert consisted of 200 same pairs and 200 different

pairs, all same pairs of CheXpert were used for retrieval. All change/no-change labels for the internal validation and CheXpert were completely reviewed by a board-certified thoracic radiologist with 6 years' experience. Labels for CIG were reviewed by radiology and internal medicine clinicians [19]. All change dataset was consisted of the same pairs with change labels and the no-change dataset with no-change labels. The different pairs were common to each dataset. Flow diagram of data collection is depicted in Figure 1.

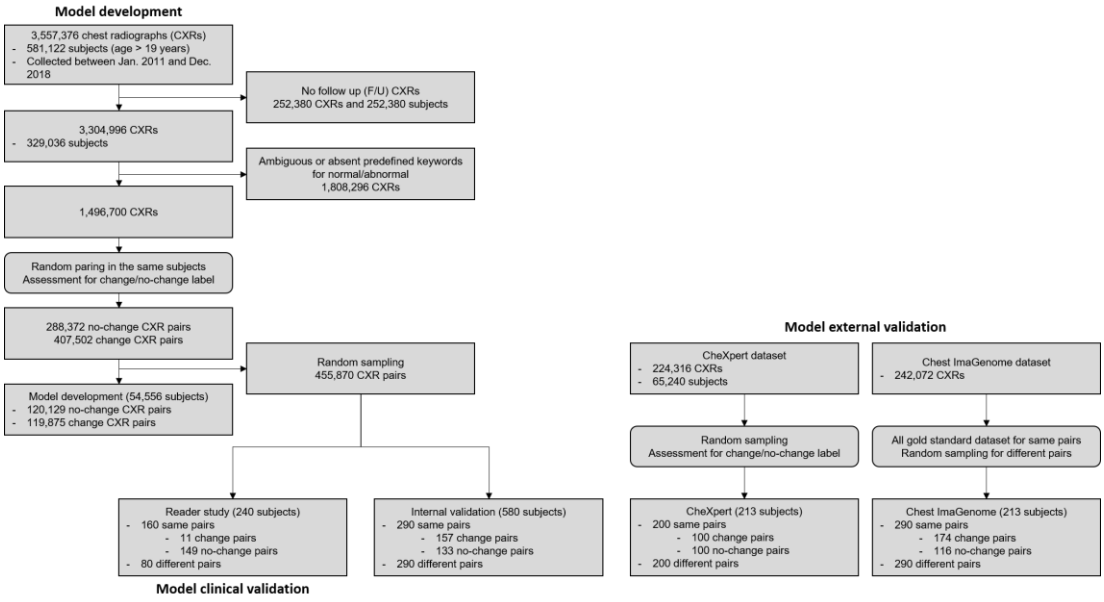

Figure 1. Flow diagram of data collection for model development, clinical validation, and external validation.

Patient identification model development

We developed a similarity-based DL network, SimChest for patient identification. During training, SimChest optimizes the supervised contrastive loss [20] to allocate short distances to CXRs from the same subject and large distances to CXRs from different subjects in the latent space. After training, SimChest was used to obtain feature vectors from the reference and comparison CXRs. Finally, the cosine similarity between the two vectors was calculated. This similarity index was used for the screening for misidentification errors. The training strategy of SimChest is depicted in Figure 2.

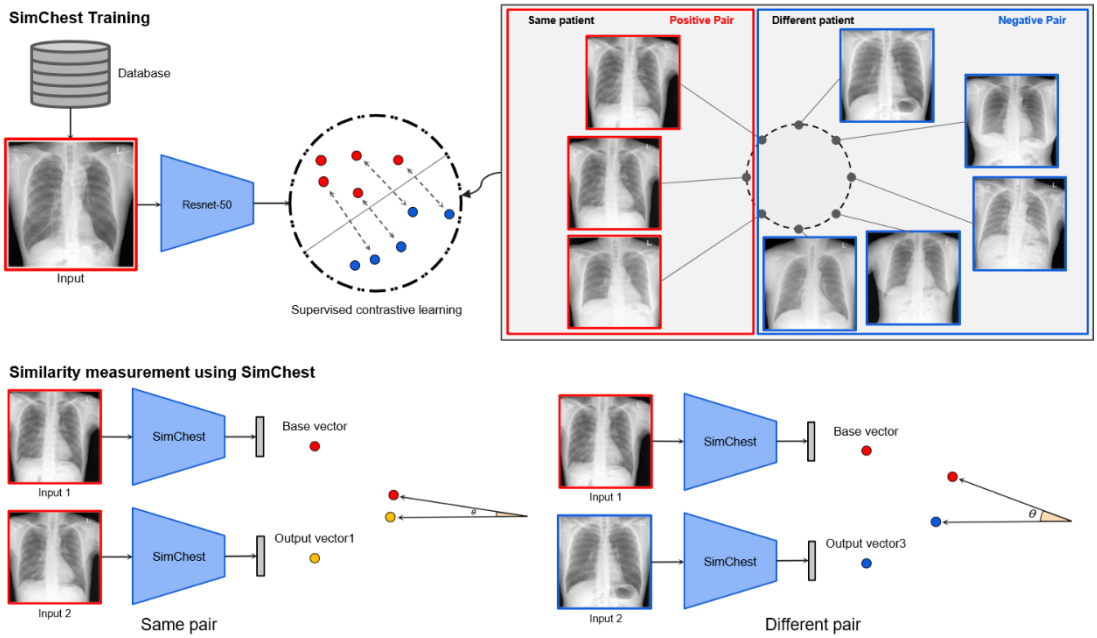

Figure 2. Training strategy and similarity measurement method of SimChest.

To obtain a radiologic fingerprint and measure the similarity between two different CXRs, SimChest optimizes supervised contrastive loss. The supervised contrastive loss is described as follows:

$$\boldsymbol{Supervised\ Contrastive\ Loss} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \boldsymbol{log} \frac{\boldsymbol{exp}\,(z_i \cdot z_p/\tau)}{\sum_{a \in A(i)} \boldsymbol{exp}\,(z_i \cdot z_a/\tau)} \quad (1)$$

Here, i denotes each image from the entire dataset I, which also contains augmented images. $A(i)$ means the dataset I excluding i. z_i denotes the feature vector from i and the z_p from the positive key of the query. z_a denotes the feature vector from the negative key of the query, and τ means temperature, which was set to 0.07. $P(i) \equiv \{p \in A(i) : \tilde{y}_p = \tilde{y}_i\}$ is the set of augmented images but affiliated same class with i and $|P(i)|$ is the cardinality of i.

Comparison of model performance with human expert

To confirm the validity of screening for patient misidentification errors using a pair of CXRs and to compare the model performance with human experts, a reader study was conducted. Seven radiologists participated in reader study, including three junior radiology residents (Group I), two senior radiology residents (Group II), and two board-certified radiologists, each

with 18 years and 14 years of experiences (Group III). Each radiologist independently evaluated the same or different status of each pair of CXRs, with sufficient amount of time to read the CXRs. Additional performance analyses were performed according to disease change-status and position of subject to determine if the screening performance of human readers and the DL model varied according to these statuses.

Patient retrieval

One internal validation dataset and two external validation datasets were also used to evaluate the models' performance for image-based patient retrieval. For each dataset, 200 same pairs were randomly sampled. DL-based algorithms were also used for patient retrieval. First, each algorithm extracted the feature vector of each CXR from the given dataset. Second, once a base CXR was selected, its feature vector could be used as a query. Finally, the cosine distance between the query and the feature vectors from the rest of the given dataset was calculated. Subsequently, the top-1 and top-5 ACCs were calculated for patient retrieval, indicating whether the same patient was included in the top N similar logits calculated from each patient of the given dataset. We did not present a null hypothesis test for patient retrieval because it was not meaningful given the large size of the contingency table (i.e., all p values were <0.001).

External validation and implementation details

We used one internal and two external validation datasets to determine whether the results were generalizable to patients from different hospitals and to patients of different races. In addition, we compared our method with existing DL methods to evaluate the performance of our proposed method. We selected two DL methods for comparison: contrastive learning (MoCo v2) and supervised learning methods (Siamese neural network, SNN [15]). Because SimChest optimizes supervised contrastive loss, commonly used contrastive learning method and supervised learning method were compared.

MoCo v2 was selected for the contrastive learning-based algorithm, and it was trained using contrastive loss. The loss is described as follows:

$$\boldsymbol{Contrastive\ Loss} = -\boldsymbol{log}\frac{\boldsymbol{exp}\,(\boldsymbol{q}\cdot\boldsymbol{k}_{+}/\boldsymbol{\tau})}{\sum_{\boldsymbol{i}=\boldsymbol{0}}^{\boldsymbol{K}}\boldsymbol{exp}\,(\boldsymbol{q}\cdot\boldsymbol{k}_{\boldsymbol{i}}/\boldsymbol{\tau})} \quad (2)$$

where the q denotes the feature vector from the query image and the k_+ denotes the feature vector from the positive key. The k_i except k_+ denotes the feature vector from the negative key of the query, and τ means temperature, which was set to 0.07. This loss makes q and positive key similar but q and negative key different, which helps the model cluster the same CXR in the embedding space to be nearby and the different CXRs to be far. After contrastive learning, MoCo v2 was fine-tuned with a fully connected layer for patient identification.

A Siamese neural network of Resnet-50 was selected for the supervised learning-based algorithm. Because a pair of CXRs was used as an input of Siamese Resnet-50, the L1 distance between two feature output vectors from the Siamese Resnet-50 was calculated. The sigmoid activation function was applied to the L1 distance, and finally, cross-entropy was calculated as a loss function. This loss is described as follows:

$$\boldsymbol{D} = \sum|\boldsymbol{sigmoid}(\boldsymbol{output_1}) - \boldsymbol{sigmoid}(\boldsymbol{output_2})| \quad (3)$$

$$\boldsymbol{Loss} = -[\boldsymbol{gt}\times\boldsymbol{log}(\boldsymbol{sigmoid}(\boldsymbol{D})) + (\boldsymbol{1}-\boldsymbol{gt})\times\boldsymbol{log}\,(\boldsymbol{1}-\boldsymbol{sigmoid}(\boldsymbol{D}))] \quad (4)$$

Here, $output$ denotes the feature output vector of Siamese Resnet-50 and D is the L1 distance value between $output_1$ and $output_2$. In equation (4), gt denotes the ground-truth value whether the given pair CXR is the same or different.

One of the most widely used convolutional neural network architecture, ResNet-50 was used for SimChest and other deep learning-based methods, considering the size of the dataset and training time. The models were implemented in Python version 3.6.9 using PyTorch version 1.6.0. The models were trained using a stochastic gradient descent optimizer with a learning rate of 5e-2 and a warmup of three epochs with a weight decay of 1e-4. Data augmentations, such as shifting (±6.2%), zooming (±2%), rotation (±10°), sharpening (±5), motion blur, median blur, optical distortion, Gaussian noise, and contrast limited adaptive histogram equalization, were used. All images were resized to 512 × 512 pixels. The batch

size of the overall experiment was set to the maximum for the GPU (Quadro RTX 8000) memory.

Statistical analysis

Sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV), accuracy (ACC), and area under the receiver operating characteristics curve (AUC) were calculated for the quantitative results of screening for misidentified patient. Means ± standard deviations (SDs) for ACC were calculated for the reader groups.

Two-sample proportion tests were performed between the total sample and each conditioned subgroup to compare the performance in each condition. Additional non-inferiority tests were performed to compare the performance of average human readers and the model. The null hypothesis was that the difference between the accuracy of average human readers was 5% or more (non-inferiority margin). DeLong's test for two correlated receiver operating characteristics (ROC) curves was performed to compare the misidentified patient screening results between our method and other existing DL methods. A two-sided $P<0.05$ was considered statistically significant except for non-inferiority test. All statistical analyses were performed using R version of 4.2.1.

Results

Subject characteristics

The training, internal validation, and reader study datasets all consisted of only Asian participants. The 213 participants in the CheXpert dataset consisted of 15 Asian, 14 Black, 133 White, and 51 other or unknown race. The CIG dataset did not provide with the race information. Vendor information used for CXR acquisition was only provided for the training dataset. The internal validation, CheXpert, CIG, and reader study dataset were all anonymized for the vendor information. The 240,004 CXRs in the training dataset consisted of 166,260 CXRs acquired with GE Healthcare equipment, 35,203 with Fujifilm Healthcare, 28,970 with

Canon Inc., 3,763 with Samsung Electronics, 173 with DK Medical Systems, one with Siemens Healthineers, and 5,634 with unknown equipment. Data profiles are summarized in Table 1.

Table 1. Summary of datasets used for misidentified patient screening experiments.

	Training dataset	**Reader study**	**Internal validation**	**CheXpert**	**CIG**
CXR (Patient)	240,004 (54,556)	480 (240)	1,160 (580)	800 (213)	1,160 (580)
Pairs					
Same pair	NA	160	290	200	290
Different pair	NA	80	290	200	290
Age, year	62.230 ± 13.647	56.617 ± 13.690	59.760 ± 14.151	59.366 ± 17.861	NA
Sex					
Female	17,043 (31.2)	113 (47.1)	196 (33.6)	79 (37.1)	NA
Male	21,638 (39.7)	119 (49.6)	212 (36.6)	134 (62.9)	NA
Anonymized	15,875 (29.1)	8 (3.3)	173 (29.8)	0 (0)	580 (100)
Change status					
Change	119,875 (49.9)	11 (6.9)	157 (54.1)	100 (50)	174 (60)
No-change	120,129 (50.1)	149 (93.1)	133 (45.0)	100 (50)	116 (40)

Note: CIG=Chest ImaGenome dataset. NA=Not applicable.

[a] Data are presented as number (percentage) of participants unless otherwise indicated.

Performance of the deep learning model compared to radiologists

We found that radiologists can screen if a pair of CXRs has been misidentified or not. Given sufficient time, all radiologists performed well in screening misidentified patients (mean ACC 0.900, 95% CI 0.852 – 0.948). In addition, the results reveals that the longer the radiologists

had practiced, the better they were at screening misidentified patients. Board-certified radiologists performed best (Group III 0.935), followed by senior radiology residents (Group II 0.904) and junior radiology residents (Group I 0.874). SimChest demonstrated similar accuracy to that of senior radiology residents (SimChest 0.904). In addition, SimChest achieved non-inferior performance compared to average radiologists (one-sided P for non-inferiority 0.015). The performance of radiology readers and SimChest is compared in Table 2, Figure 3.

Table 2. Performance of SimChest and human readers to screen for misidentified patient from paired CXRs

	SimChest	**Group I**	**Group II**	**Group III**	**All readers**
Sensitivity	0.917	0.794	0.750	0.858	0.800 ± 0.111
Specificity	0.900	0.900	0.956	0.961	0.933 ± 0.0385
PPV	0.753	0.722	0.857	0.881	0.806 ± 0.106
NPV	0.970	0.930	0.920	0.953	0.934 ± 0.036
Accuracy	0.904	0.874	0.904	0.935	0.900 ± 0.048

Note: CXR=Chest radiograph. PPV=Positive predictive value. NPV=Negative predictive value. Group I=Junior radiology residents. Group II=Senior radiology residents. Group III=Board-certified expert radiologists.

[a] Mean of all reader group is shown with 95% confidence interval.

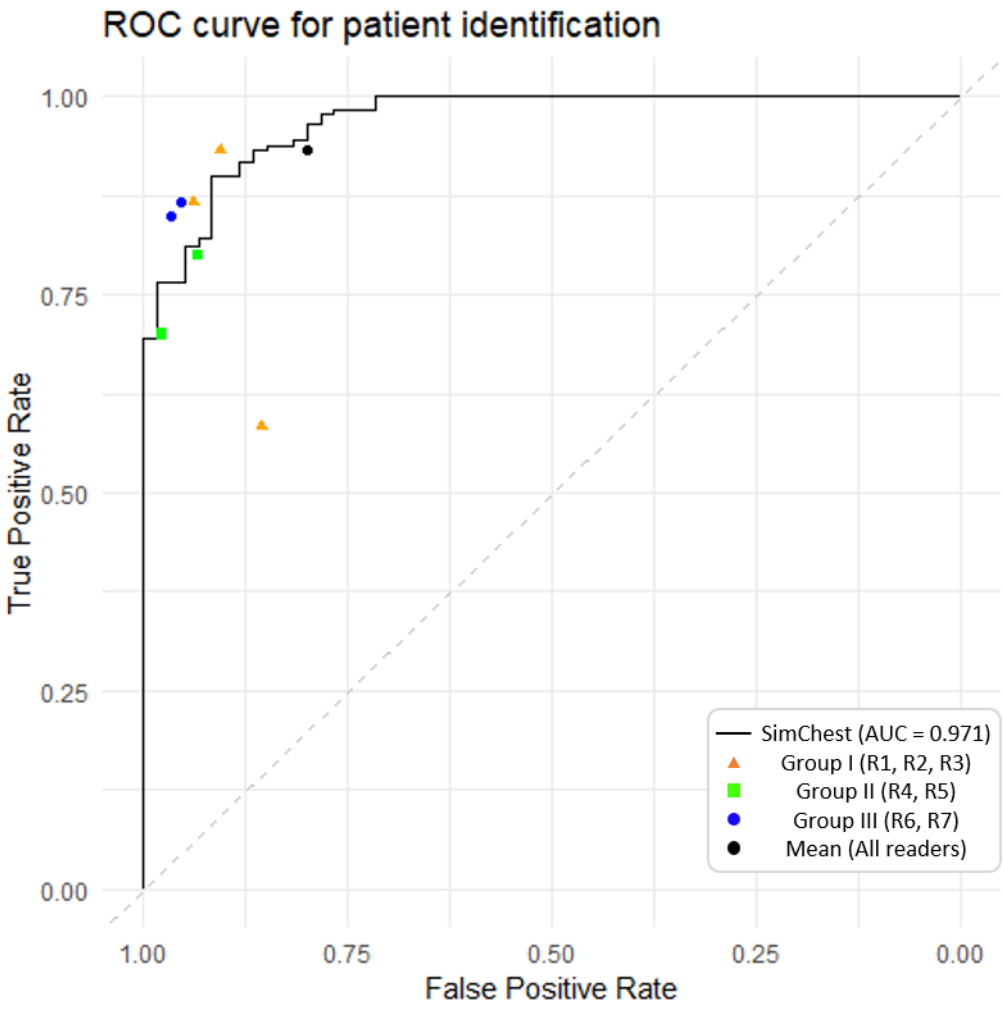

Figure 3. Receiver operating characteristics curve of SimChest and comparison with human readers.

Note: AUC=Area under receiver operating characteristics curve. Group I=junior radiology residents. Group II=Senior radiology residents. Group III=Board-certified expert radiologists.

Performance analyses according to disease change status and subject position

In the additional performance analyses by disease change-status, DeLong's test for two ROCs showed that there was no statistical difference between the performance of the total dataset (AUC 0.971, 95% CI 0.951 – 0.990) and the change dataset (AUC 0.942, 95% CI 0.885 – 1.000; P-value 0.369), and the no-change dataset (AUC 0.972, 95% CI 0.954 – 0.991; P-value 0.892). SimChest demonstrated non-inferior performance compared to average human readers

on both the change (one-sided P for non-inferiority 0.003) and no-change datasets (one-sided P for non-inferiority 0.016). In the performance analyses by position of subjects, the DeLong's test for two ROCs showed that there was no statistical difference between the performance of the total dataset (AUC 0.971, 95% CI 0.951 – 0.990) and the PA only dataset (AUC 0.989, 95% CI 0.978 – 1.000; P-value 0.098), and the AP involved dataset (AUC 0.902, 95% CI 0.818 – 0.987; P-value 0.126). SimChest showed non-inferior performance compared to average human readers in the PA only dataset (one-sided P for non-inferiority 0.002), but not in the AP involved dataset (one-sided P for non-inferiority 0.468). The analyses are summarized in Table 3.

Table 3. Accuracies of SimChest and human readers to screen for misidentified patient from paired CXRs according to disease changes status and subject position (PA/AP)

	Disease change status		**Subject position**		**Total**
	Change (N=71)	**No-change (N=229)**	**PA only (N=164)**	**AP involved (N=76)**	
Group I	0.775	0.884	0.917	0.781	0.874
Group II	0.761	0.908	0.930	0.849	0.904
Group III	0.880	0.932	0.963	0.875	0.935
All readers	0.801 ± 0.113	0.905 ± 0.042	0.934 ± 0.051	0.827 ± 0.061	0.900 ± 0.048
SimChest	0.901	0.908	0.957	0.789*	0.904

Note: CXR=Chest radiograph. PA=Posterior-Anterior projection. AP=Anterior-Posterior projection. PPV=Positive predictive value. NPV=Negative predictive value. Group I=Junior radiology residents. Group II=Senior radiology residents. Group III=Board-certified expert radiologists.

[a] (*), *P*-value<0.05; (**), *P*-value<0.01; (***), *P*-value<0.001.

[b] Two-sample proportion test was conducted between total sample and each condition to compare the performance in each condition.

[c] Mean of all reader group is shown with 95% confidence interval.

Performance and generalizability of deep learning models

SimChest and the other DL models were evaluated on internal, CheXpert, and CIG datasets. The performances are summarized in Table 4. In internal validation dataset, SimChest

outperformed the other deep learning models with the AUC of 0.996 (SNN 0.961, P-value<0.001; MoCo v2 0.966, P-value<0.001). SimChest showed robust performance on CheXpert with the AUC of 0.941, outperforming MoCo v2 (AUC 0.906, P-value=0.011) and SNN (AUC 0.830, P-value <0.001). SimChest also showed robust performance on CIG dataset with the AUC of 0.950, outperforming MoCo v2 (AUC 0.924, P-value=0.037) and SNN (AUC 0.853, P-value<0.001).

Table 4. AUCs of deep learning models to screen for misidentified patient from paired CXRs

	SNN	**MoCo v2**	**SimChest**
Internal validation			
Change	0.958 ± 0.018***	0.963 ± 0.016***	0.992 ± 0.007
No-change	0.964 ± 0.016***	0.968 ± 0.015***	0.999 ± 0.001
Total	0.961 ± 0.015***	0.966 ± 0.014***	0.996 ± 0.003
CheXpert			
Change	0.801 ± 0.051***	0.897 ± 0.036*	0.933 ± 0.030
No-change	0.860 ± 0.041***	0.915 ± 0.030*	0.948 ± 0.028
Total	0.830 ± 0.040***	0.906 ± 0.029*	0.941 ± 0.021
CIG			
Change	0.849 ± 0.034***	0.923 ± 0.024*	0.941 ± 0.021
No-change	0.860 ± 0.036***	0.925 ± 0.025	0.951 ± 0.024
Total	0.853 ± 0.029***	0.924 ± 0.022*	0.950 ± 0.016

Note: AUC=Area under receiver operating characteristic curve. CIG=Chest ImaGenome dataset. SNN=Siamese Neural Network.

[a] (*), *P*-value<0.05; (**), *P*-value<0.01; (***), *P*-value<0.001.

[b] DeLong's test for paired receiver operating characteristics curves was performed to compare the performance of SimChest with the other models.

Table 5. Performance of deep learning models to retrieve a patient from the set of CXRs

	SNN	**MoCo v2**	**SimChest**
Internal validation			
Top 1 ACC	0.315	0.185	0.885
Top 5 ACC	0.560	0.420	0.965
CheXpert			
Top 1 ACC	0.125	0.100	0.505
Top 5 ACC	0.230	0.230	0.820
CIG			

Top 1 ACC	0.080	0.105	0.400
Top 5 ACC	0.240	0.240	0.595

Note: CIG=Chest ImaGenome dataset. ACC=Accuracy. SNN=Siamese Neural Network.

a Top N ACC indicates whether the same patient was included in the top N similar logits calculated from each patient of the given dataset

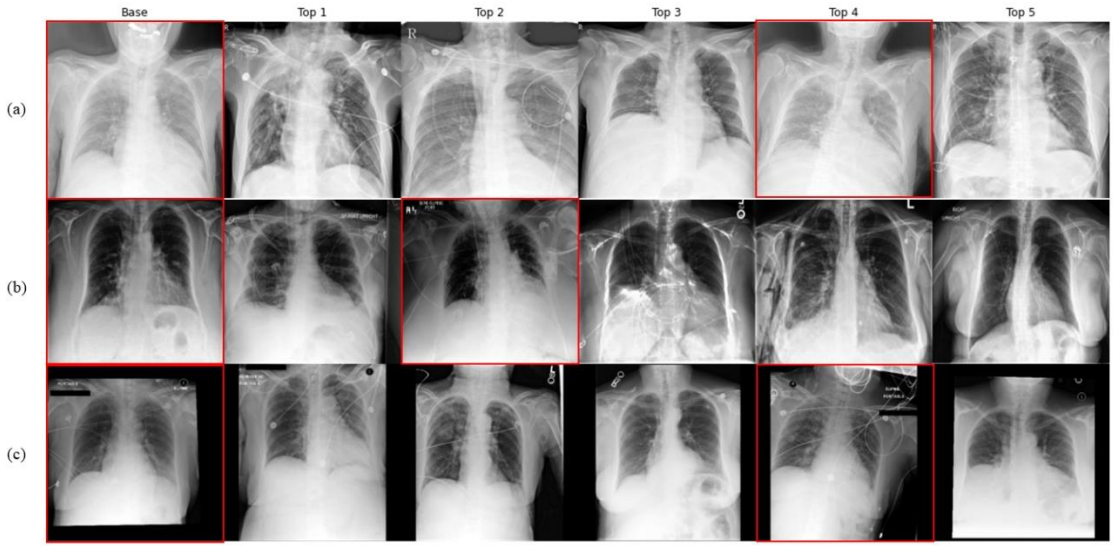

Figure 4. Sample images of the base CXR and the top 5 result CXRs for each dataset in the retrieval task.

Note: (a) represents the internal dataset, (b) is the CheXpert dataset, and (c) is the CIG dataset. In each row, the red boxes represent the base CXR and the CXR of the same patient as the base patient.

In addition, the ability to retrieve the most similar CXR from the given dataset was also evaluated in this study. SimChest showed the best and most robust performance across all datasets. Other DL models were unable to retrieve the patient in the external validation datasets, and the results are summarized in Table 5. The sample images of the base CXR for each dataset ant the top 5 result CXRs are depicted in Figure 4.

The saliency maps acquired using gradient weighted class activation map (Grad-CAM) [21] of SimChest are shown in Figure 5. Figure 6 shows the class activation maps of Siamese Resnet-50 and MoCo v2. The saliency map of SimChest focused mainly on the ribs and

thoracic cage, and this trend is similar to the results of the reader test compared to Siamese Resnet-50 and MoCo v2.

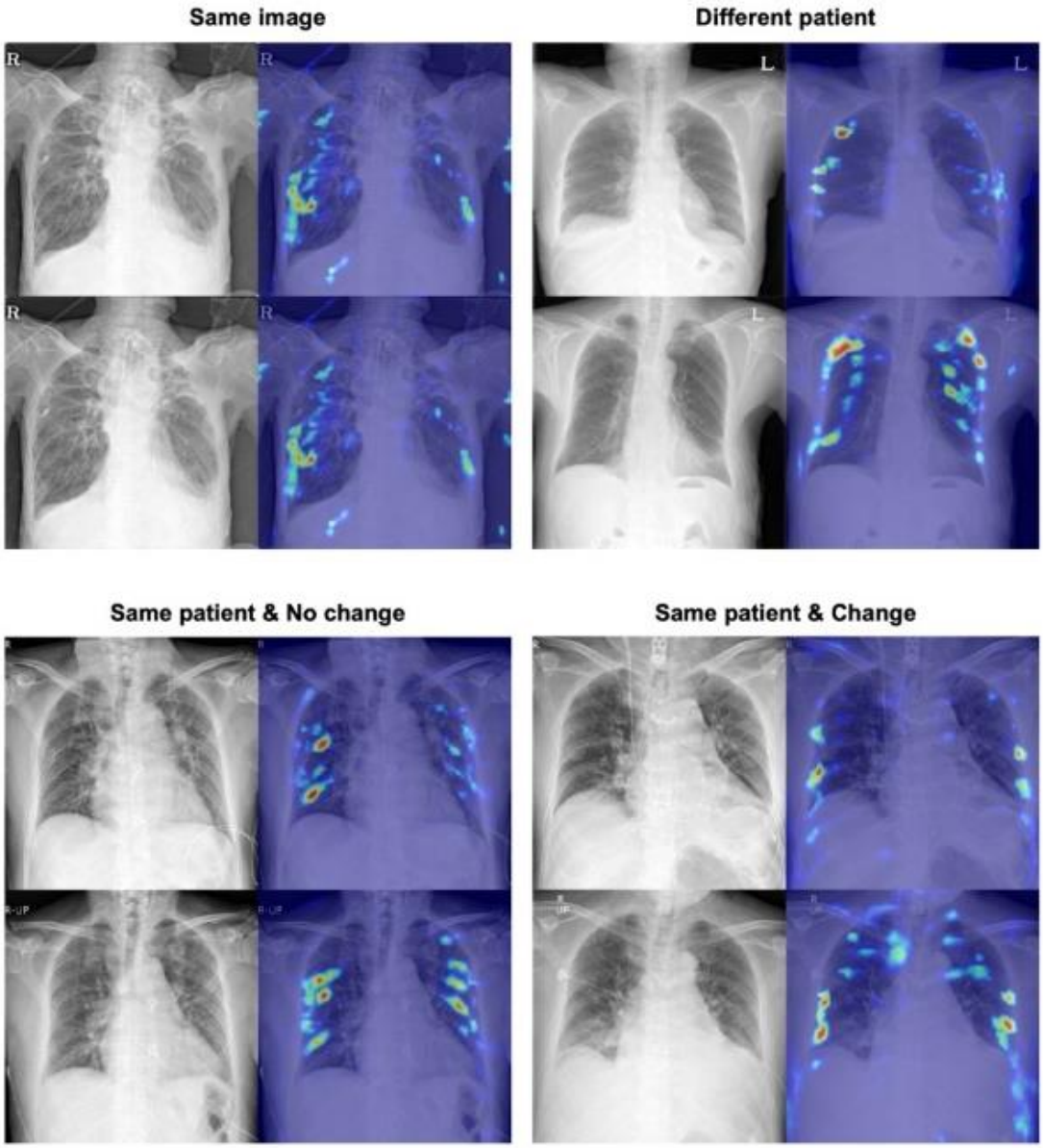

Figure 5. Grad-CAM of SimChest on the following: same image, same patient with change/no-change of disease, and different patient.

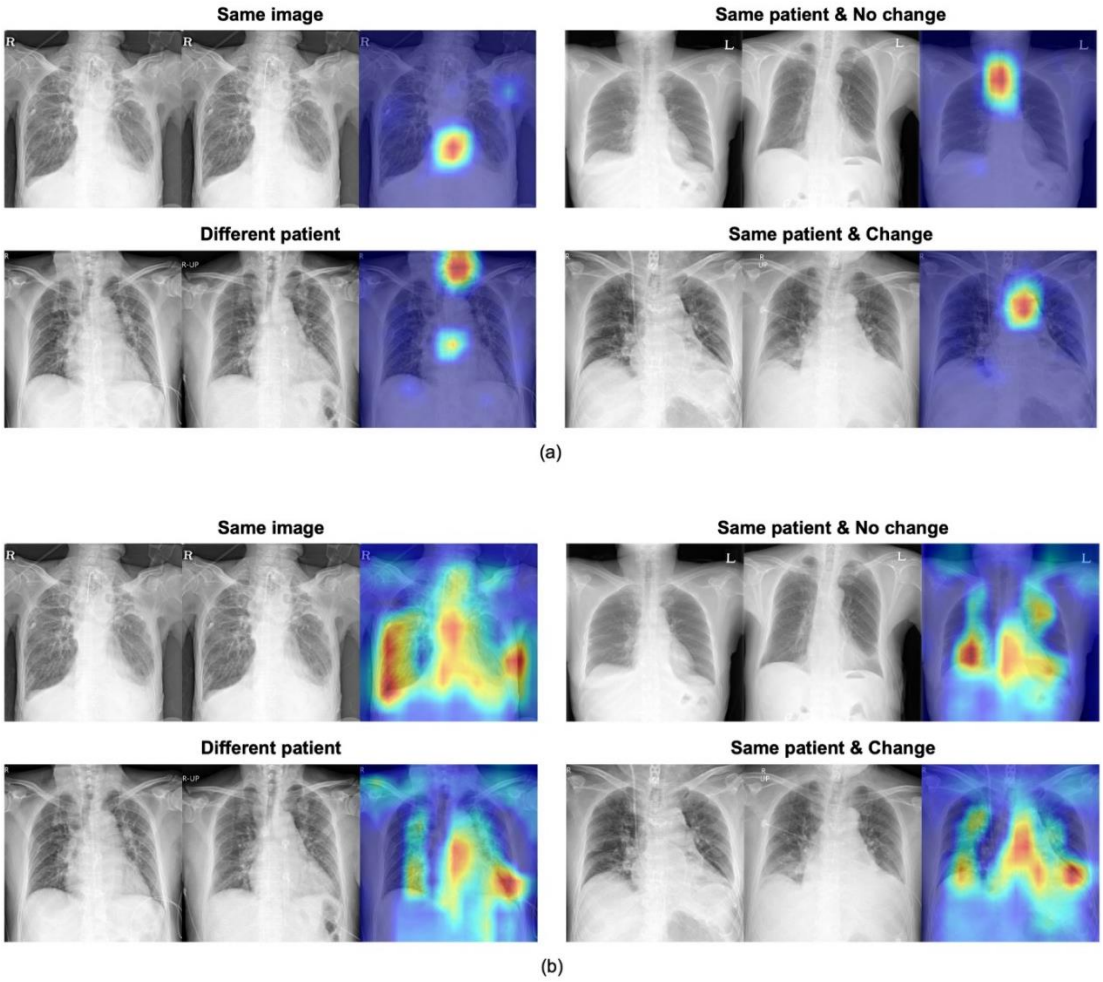

Figure 6. Grad-CAM of Siamese Neural Network (a) and MoCo v2 (b) on the following: same image, same patient with change/no-change of disease, and different patient.

Discussion

In this study, we showed that both the DL models and the radiologists can screen for misidentified patients from paired CXRs and compared the performance with that of DL model. This finding is noteworthy because human experts typically do not perform this task, and as a result, clinical feasibility have not been proved previously. Furthermore, the results of our model were comparable to those of senior radiology residents. In addition, we provided additional insight into the performance of model and human readers through subgroup analyses by disease change-status and positions of subjects.

We also showed that the generalizability of the DL model across various clinical settings, medical imaging vendors, and subject populations by utilizing multiple private and public datasets. In addition, our model showed the consistent performance regardless of the disease change/no-change status of the CXR pairs. Finally, SimChest showed the most robust results across all external validation datasets. Furthermore, considering that the DL models were trained on the CXRs of Asian race, SimChest showed generalizable performance on the CheXpert and CIG datasets, which consisted of Black, White, Hispanic and other races. This can be interpreted as it being a race-agnostic model.

We also demonstrated that our similarity-based DL model, SimChest, can retrieve the most similar CXR from a given set of CXRs on the internal validation, CheXpert, and CIG datasets. Accurate patient identification and retrieval through this similarity-based method bear significance not only within the task itself but also in the clinical context. Two clinical scenarios for utilizing SimChest are proposed, as illustrated in Figure 7.

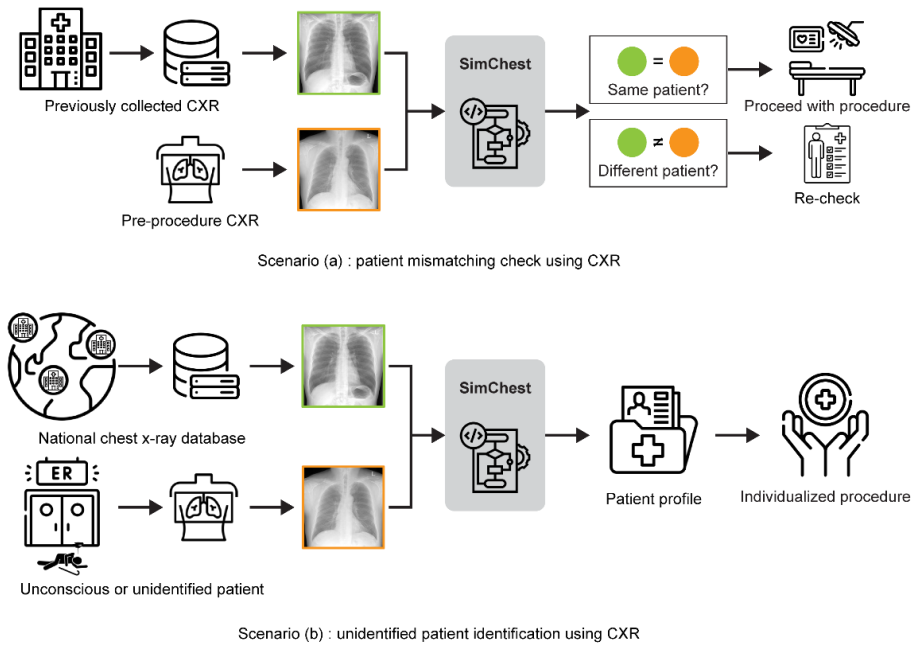

Figure 7. Possible scenario for using the patient identification algorithm in clinical practice.

In the preoperative or interventional setting, SimChest can serve as an additional tool to verify the patient's identity during the "time-out" process as depicted in scenario (a). Physicians can proceed with the procedure if SimChest confirms that the two CXRs belong to the same patient. However, if SimChest indicates that the two CXRs belong to different patients, physicians should re-evaluate the similarity logit and repeat the "time-out" process with the medical team. This process can help prevent misidentification errors.

SimChest also has potential applications in emergency situations, as depicted in scenario (b), where rapid patient identification can be of benefit to unconscious patients. If we establish a national database that stores latent vectors of CXRs representing each individual, SimChest can be utilized to promptly identify unconscious and unidentified patients in the emergency department. By identifying individual patient and matching patient information, physicians can promptly order diagnostic tests to determine the underlying cause of the loss of consciousness and provide personalized treatment.

Our study demonstrated promising results of SimChest using paired CXRs. Patient misidentification can occur in all medical departments. Although it did not reach the level of board-certified radiologists, SimChest achieved a comparable performance to radiology residents. In addition, the non-inferiority test demonstrated that SimChest showed non-inferior performance to average radiologists regardless of disease change-status. However, It was not determined whether the model performed non-inferiorly to average radiologists regardless of patient position.

Given the widespread use of CXR as a prevalent imaging modality among physicians and radiologists, our findings present a promising approach to improving patient safety by mitigating misidentification errors. Considering the infrequent incidence of misidentification errors in the real-world [1], false alarm in well-matched cases triggered by SimChest may potentially lead to clinician and radiologist fatigue. However, as it is difficult to screen misidentification errors once they have occurred [5], the implementation of this autonomous safety device may help to reduce such errors independently of human intervention. Further

studies will be conducted to address false positives and enhance generalizability, aiming to alleviate this fatigue.

This study has some limitations. First, due to the retrospective nature of this study, the study results may not be generalizable to real-world practice. Further prospective analysis in clinical practice is needed to confirm the study results. Second, although SimChest showed robust performance regardless of disease change-status, whether the model can perform non-inferiorly compared to average radiologists regardless of patient position remained undetermined in this study. Further research is needed to develop models that can perform at the level of radiologists regardless of patient position. Finally, the patient retrieval performance of SimChest in the external validation datasets was not good enough to be used as it is. This may be due to the differences in the acquisition settings of the CXRs or differences in race. Figure 4 illustrates the sample images of the base CXR for each dataset and the top 5 result CXRs, and these belong to the top 5 but the top 1 case. When examining the external datasets (b) and (c), CXRs corresponding to the top 5 are more similar to the base CXR compared to the internal dataset. However, distinguishing whether the CXRs from the base patient and the reference patient are the same is challenging due to variations in posture, breathing level, and other factors. Additionally, in the case of CheXpert dataset, difference in trends may arise due to preprocessing, while in the case of CIG, the dataset is predominantly obtained in AP in ICU settings, resulting in distinct characteristics from the internal dataset.

However, patient retrieval can be used in real-world settings as content-based image retrieval [17] to find similar cases or to find patients in a defined database as shown in scenario (b) of Figure 7. Therefore, the decent performance shown in the internal validation dataset may show its usability for fine-tuning in the hospital-based cohorts, community-based cohorts, or national cohorts.

Conclusion

Our study showed that the DL model can precisely screen for misidentified patients using paired CXRs. Our results were robust across disease change-status, patient position, multiple datasets, multiple races, and multiple clinical settings. In addition, radiologists can also screen for misidentified patients from paired CXRs and SimChest performed non-inferiorly to human radiologists. Furthermore, the performance of SimChest in patient retrieval had shown potential to be applied in content-based image retrieval or identification of unidentified patient from a given cohort database.

Acknowledgement

This study was supported by grants of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, (HI18C0022, HI18C2383) and Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Republic of Korea (1711134538, 20210003930012002).

We would like to express our gratitude to Hee Jun Park, Chan Yang Cho, Su Kyeong Yeon, So Yeon Kim, Seongken Kim, Ju Hee Lee, and Min Jung Kim for participating as human readers for the reader study.

References

1. Rubio, E.I. and L.J.A.J.o.R. Hogan, *Time-out: it's radiology's turn—incidence of wrong-patient or wrong-study errors.* AJR Am J Roentgenol, 205(5):941-6 (2015)
2. Sadigh, G., et al., *Evaluation of near-miss wrong-patient events in radiology reports.* AJR Am J Roentgenol, 205(2):337-43 (2015)
3. Beyea, S.C.J.A.j., *Patient identification--a crucial aspect of patient safety.* AORN journal 78, 478-481 (2003)
4. Papadakis, M., A. Meiwandi, and A.J.I.J.o.S. Grzybowski, *The WHO safer surgery checklist time out procedure revisited: Strategies to optimise compliance and safety.* Int J Surg, 69:19-22. (2019)
5. Henneman, P.L., et al., *Patient identification errors are common in a simulated setting.* Annals of emergency medicine 55, 503-509 (2010)
6. Jeon, B., et al., *A facial recognition mobile app for patient safety and biometric identification: Design, development, and validation.* JMIR Mhealth Uhealth, 7(4):e11472 (2019)
7. Silverstein, E. and M.J.M.p. Snyder, *Implementation of facial recognition with Microsoft Kinect v2 sensor for patient verification,* Med Phys, 44(6):2391-2399 (2017)
8. Ampamya, S., J.M. Kitayimbwa, and M.C.J.I.J.o.M.I. Were, *Performance of an open source facial recognition system for unique patient matching in a resource-limited setting.* Int J Med Inform, 141:104180 (2020)
9. Morishita, J., et al., *Potential usefulness of biological fingerprints in chest radiographs for automated patient recognition and identification* Acad Radiol, 11(3):309-15 (2004)
10. Morishita, J., et al., *An automated patient recognition method based on an image-matching technique using previous chest radiographs in the picture archiving and communication system environment.* Medical physics 28, 1093-1097 (2001).
11. Kao, E.-F., et al., *Automated patient identity recognition by analysis of chest radiograph features,* Academic Radiology 20, 1024-1031 (2013).

12. Raghu, V.K., et al., *Deep learning to estimate biological age from chest radiographs.* JACC Cardiovasc Imaging, 14(11):2226-2236 (2021)
13. He, H., et al. *Model and predict age and sex in healthy subjects using brain white matter features: a deep learning approach*. in 2022 IEEE 19th International Symposium on Biomedical Imaging (2022)
14. Gichoya, J.W., et al., *AI recognition of patient race in medical imaging: a modelling study,* Lancet Digit Health, 4(6):e406-e414 (2022)
15. Packhäuser, K., et al., *Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest X-ray data.* Sci Rep 12, 14851 (2022)
16. Donaldson, M.S., J.M. Corrigan, and L.T. Kohn, *To err is human: building a safer health system.* National Academies Press (US) (2000)
17. Choe, J., et al., *Content-based image retrieval by using deep learning for interstitial lung disease diagnosis with chest CT,* Radiology, 302(1):187-197 (2022)
18. Irvin, J., et al. *Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison*, Proceedings of the AAAI conference on artificial intelligence (2019)
19. Wu, J.T., et al., *Chest ImaGenome dataset for clinical reasoning,* PhysioNet, (2021)
20. Khosla, P., et al., *Supervised contrastive learning,* Adv. Neural Inf. Process, 33, 18661–18673 (2020)
21. Selvaraju, R.R., et al. *Grad-cam: Visual explanations from deep networks via gradient-based localization*, Proceedings of the IEEE international conference on computer vision. (2017)

국문요약

배경: 환자의 정확한 식별은 환자 안전과 정확한 진단 및 치료를 위해 중요하다. 비록 실제 임상 현장에서 환자 식별 오류에 대한 보고가 드물긴 하지만, 한 번이라도 환자 식별의 오류가 발생하면 부정확한 치료와 잘못된 환자를 수술하는 등의 치명적인 결과를 초래할 수 있다. 따라서 본 연구의 목표는 쌍으로 된 흉부 방사선 사진에서 환자 식별 오류를 자동으로 검출하는 딥러닝 모델을 제안하여 이를 평가하고 전문가와 비교 분석하는 것이다.

방법: 우리는 240,004 개의 흉부 방사선 사진을 활용하여 환자 식별을 위한 딥러닝 모델을 개발하였다. 딥러닝 모델은 지도 대조 학습을 통해 잠재 공간에서 같은 환자의 흉부 방사선 사진의 경우 거리를 가깝게, 다른 환자의 흉부 방사선 사진의 경우 거리를 멀게 하는 방향으로 학습을 하였다. 이렇게 학습한 모델은 내부 검증 데이터셋, CheXpert 및 Chest ImaGenome 데이터셋을 비롯한 여러 데이터셋을 사용하여 검증되었으며, 각 데이터셋은 여러 인종을 포함한다. 모델의 성능은 질병의 상태 변화에 따라서도 분석되었으며, 모델의 성능을 평가하기 위해 세 명의 주니어 방사선 전문의 그룹(그룹 I), 두 명의 시니어 방사선 전문의 그룹(그룹 II) 및 두 명의 인증된 전문 방사선 전문의 그룹(그룹 III)의 쌍으로 된 흉부 방사선 사진에서 환자를 식별하는 성능과 비교하였다. 또한, 비열등성 검정을 통해 비교 분석하였다.

결과: 유사성 기반 딥러닝 모델인 SimChest 는 질병의 상태 변화 여부와 관계없이 다양

한 데이터셋에서 가장 뛰어난 환자 식별 성능을 보였다(내부 검증 데이터셋 [수신자 작동 특성 곡선 아래의 영역 (AUC) 범위 0.992-0.999], CheXpert [0.933-0.948], 및 Chest ImaGenome [0.949-0.951]). 방사선 전문의들은 쌍으로 된 흉부 방사선 사진에서 평균 정확도 0.900(95% 신뢰 구간 0.852-0.948)로 환자를 식별할 수 있으며, 이는 전문의들의 수련 경험의 증가와 함께 향상되었다. 그룹 I 의 평균 정확도는 0.874, 그룹 II 는 0.904, 그룹 III 는 0.935 였으며, SimChest 의 평균 정확도는 0.904 였다. SimChest 의 성능은 방사선 전문의들의 평균 성능과 유사성을 가지며, 비열등성 검정에서 P-값은 0.015 로 나타났다. 이는 SimChest 가 방사선 전문의들의 성능을 크게 못지않은 수준으로 달성했음을 의미한다.

결론: 이 진단 연구는 딥러닝 모델이 쌍으로 된 흉부 방사선 사진을 사용하여 환자 식별 오류를 자동으로 검출하며, 이는 방사선 전문의의 수준에 비열등함이 입증되었다. 이 연구는 실제 임상 현장에서 환자 식별을 통해 환자 안전을 향상시키는 데 활용될 수 있다.