Journal of Imaging Informatics in Medicine
https://doi.org/10.1007/s10278-024-01247-y

ORIGINAL PAPER

Integrating VAI-Assisted Quantified CXRs and Multimodal Data to Assess the Risk of Mortality

Yu-Cheng Chen[1] · Wen-Hui Fang[2,3] · Chin-Sheng Lin[4] · Dung-Jang Tsai[3,5] · Chih-Wei Hsiang[6] · Cheng-Kuang Chang[6] · Kai-Hsiung Ko[6] · Guo-Shu Huang[6] · Yung-Tsai Lee[7] · Chin Lin[3,5]

Received: 24 February 2024 / Revised: 27 July 2024 / Accepted: 21 August 2024

Abstract

To address the unmet need for a widely available examination for mortality prediction, this study developed a foundation visual artificial intelligence (VAI) to enhance mortality risk stratification using chest X-rays (CXRs). The VAI employed deep learning to extract CXR features and a Cox proportional hazard model to generate a hazard score ("CXR-risk"). We retrospectively collected CXRs from patients visited outpatient department and physical examination center. Subsequently, we reviewed mortality and morbidity outcomes from electronic medical records. The dataset consisted of 41,945, 10,492, 31,707, and 4441 patients in the training, validation, internal test, and external test sets, respectively. During the median follow-up of 3.2 (IQR, 1.2–6.1) years of both internal and external test sets, the "CXR-risk" demonstrated C-indexes of 0.859 (95% confidence interval (CI), 0.851–0.867) and 0.870 (95% CI, 0.844–0.896), respectively. Patients with high "CXR-risk," above 85th percentile, had a significantly higher risk of mortality than those with low risk, below 50th percentile. The addition of clinical and laboratory data and radiographic report further improved the predictive accuracy, resulting in C-indexes of 0.888 and 0.900. The VAI can provide accurate predictions of mortality and morbidity outcomes using just a single CXR, and it can complement other risk prediction indicators to assist physicians in assessing patient risk more effectively.

Keywords Chest X-ray · Deep learning · Mortality · Risk stratification · Survival analysis

Abbreviations

Afib	Atrial fibrillation
Alb	Albumin
ALT	Alanine aminotransferase
AST	Aspartate aminotransferase
Baso	Basophil
BUN	Blood urea nitrogen
CAP	Community-acquired pneumonia
COPD	Chronic obstruction pulmonary disease
Cr	Creatinine
CRP	C-reactive protein
eGFR	Estimated glomerular filtration rate
Eos	Eosinophil
Hb	Hemoglobin
HF	Heart failure
K	Potassium

✉ Yung-Tsai Lee
andrewytlee.cvs@gmail.com

✉ Chin Lin
xup6fup@mail.ndmctsgh.edu.tw

[1] Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan, ROC

[2] Department of Family and External Medicine, Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan, ROC

[3] Artificial Intelligence and Internet of Things Center, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan, ROC

[4] Division of Cardiology, Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan, ROC

[5] School of Medicine, National Defense Medical Center, Neihu Dist, No. 161, Min-Chun E. Rd., Sec. 6, Taipei 114, Taiwan, ROC

[6] Department of Radiology, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan, ROC

[7] Division of Cardiovascular Surgery, Cheng Hsin Rehabilitation and Medical Center, BeitouDist, No. 45, Zhenxing St, Taipei City 112, Taiwan, ROC

Published online: 24 October 2024

Lymph	Lymphocyte
MCV	Mean corpuscular volume
Mono	Monocyte
Na	Sodium
Neut	Neutrophil
PLT	Platelet
STK	Stroke
TB	Total bilirubin
TC	Total cholesterol
TG	Triglyceride
uPro	Dipstick test urine protein
WBC	White blood cell count

Introduction

Multimorbidity, which means the presence of multiple chronic conditions in one individual at the same time, is a growing issue among the world in past decade [1]. The prevalence of multimorbidity was varied and can be up to more than 80% in a certain region [2] and contributed to over 70% of outpatient service [3]. People with multimorbidity had been proved to have higher healthcare utility and cost, polypharmacy, more functional decline and disability, lower quality of life, and, most importantly, higher mortality rate [4, 5]. However, due to difference between the definitions from each study [4], the population with multimorbidity is highly heterogenic and their outcomes were varied. Currently, there was a lack of tools that can adequately differentiate this growing group of people according to their risk of future adverse outcomes, especially death.

For a comprehensive evaluation, routine examinations such as medical history assessment, physical examinations, complete blood count, and biochemical tests are conducted in primary care. These assessments can be readily applied to hundreds of clinical risk scores, which are easily accessed from website such as MDCalc, UpToDate, and Medscape, to estimate the risk of developing various diseases in the future. Sometimes, it is still necessary to complement the evaluation of a patient's condition with additional imaging and physiological assessments, such as chest X-ray (CXR), electrocardiogram, sonography, and computed tomography. In those exams, CXR may be the most available and cheap and the results can be acquired within a brief time. Also, CXR is useful in diagnosing pulmonary and cardiovascular diseases which include but are not limited to chronic obstructive pulmonary disease, congestive heart failure, pneumonic infiltration, pleural effusion, and pulmonary artery enlargement [6]. Furthermore, it provides a lot of other information about mediastinum, diaphragm, subdiaphragmatic area, soft tissue, and bones [7]. However, interpreting CXRs requires radiologists. To facilitate easy integration into existing clinical risk scores, physicians may need a method to quantify CXR findings into a score, thus lowering the barrier for its application in primary care.

Within the field of image analysis, deep learning has emerged as the most rapidly growing method over the past decades, and it has been widely applied in biologic research and healthcare, such as proteomics [8, 9], oncology [10], radiology [11], disease diagnosis [12], and decision-making [13]. Due to its wide availability and considerable number of images, deep learning models (DLMs) have already been developed to assist the chest X-ray interpretation and were mostly applied to disease diagnosis, lesions localization, anatomic structure labeling, and interval change identification [14]. However, even DLMs had achieved state-of-the-art performance in CXR diagnosis [15, 16], its utility had not been extended to prognosis prediction until recently. The studies done in past few years proved that a DLM-enhanced single CXR-based risk score was effective in predicting prognosis [17, 18]. Previous studies had some limitations, such as a narrow study population and the lack of comparison with demographic, comorbidities, and laboratory data. Currently, scoring systems, like teh APACHE II Score [19, 20] and Charlson Comorbidity Index [21, 22], that use clinical variables and laboratory data to predict patient's outcome had been developed and validated for a long time and remain useful in today's clinical practice. Given the powerful image analysis capabilities of DLM, it is essential to develop a CXR quantification method based on DLM. However, training hundreds of individual DLMs separately to integrate CXR information with hundreds of clinical risk scores is evidently impractical. We need a method to quantify CXR so that it can be seamlessly integrated with existing clinical risk scores, thereby enhancing its practical value.

The foundation model is a type of DLM pre-trained on large database and capable of applying knowledge to other downstream tasks [23]. We believe that the proliferation of foundation models holds the potential to address the challenge of integrating DLM with diverse clinical risk scores. This study will utilize a publicly available base model for CXR, CheXzero [24], and demonstrate the relevance of these high-level features to mortality. We will also employ a simple linear model to integrate these features with patient characteristics to show the compatibility of this analysis technique with existing clinical risk scores.

Method

Population

This study was ethically approved by the institutional review board (IRB NO. C202105150) in Tri-Service General Hospital, Taipei, Taiwan (for blinded). We retrospectively collected data from two branches, NeiHu and Tingzhou, within

Hospital A system between Jan. 2011 and Apr. 2021. We selected patients visiting outpatient department and physical examination center with at least one CXR in posteroanterior (PA) view and associated radiological report during the study period. A total of 88,585 patients were included into the database. Figure 1 shows the generation of training, validation, internal test, and external test sets. Patients who had CXR exam in Tingzhou branch were assigned to external test set. The patients collected from Neihu branch were randomly assigned to training, validation, and internal test sets according to 4:1:3 ratio. To avoid over-representing sicker patients who received more CXR examinations, we applied the strategy previously mentioned by Raghunath et al. [25]. Instead of selecting the most recent CXR, one single CXR was sampled randomly per given patient in each set. This strategy will be more representative while deploying the model on a new CXR from a patient at a random time point within the life. At the end, there were 41,945, 10,492, 31,707, and 4441 pairs of patient-CXR in training, validation, internal test, and external test set, respectively. There was no overlap existing between datasets.

The Implementation of Visual Artificial Intelligence (VAI)

The CXR image was stored in DICOM format with a resolution more than 2000 × 2000 pixels. The major feature extraction architecture is based on CheXzero [24], including an image encoder and a language encoder pre-trained by CXR and associated radiological report pair. We used its image encoder, a vision transformer (ViT-B/32), to extract the embeddings for each CXR in our dataset. Each input image was resized to resolution of 256 × 256 to fit the network structure and encoded into a feature vector with length of 512. The weight, "best_64_0.0001_original_35000_0.864," was implemented without any fine-tuning process. The output feature vectors were then used to develop the Cox model for mortality prediction, called as CXR-risk score. The coefficient of Cox model is provided in Online Resource 1. The above details were implemented in a Python environment, version 3.10.10, utilizing the "torch" package version 2.0.1.

Baseline Information

The electronic medical records of each hospital supplied the baseline information. Patient's demographic data, such as age, gender, and body mass index (BMI), were extracted from medical records. The disease histories were based on an existed diagnosis according to the corresponding International Classification of Diseases, Ninth Revision and Tenth Revision (ICD-9 and ICD-10, respectively) or laboratory tests. They included diabetes mellitus (DM, ICD-9 codes 250.x and ICD-10 codes E11.x), hypertension (HTN, ICD-9 codes 401.x to 404.x and ICD-10 codes I10.x to I16.x), hyperlipidemia (HLP, ICD-9 codes 272.x and ICD-10 codes E78.x), chronic kidney disease (CKD, ICD-9 codes 585.x and ICD-10 codes N18.x), acute myocardial infarction (AMI, ICD-9 codes 410.x and ICD-10 codes I21.x), coronary artery disease (CAD, ICD-9 codes 410.x to 414.x, and 429.2, and ICD-10 codes I20.x to I25.x), heart failure (HF, ICD-9 codes 428.x and ICD-10 codes I50.x), stroke (STK, ICD-9 codes 430.x to 438.x and ICD-10 codes I60.x to I63.x), atrial fibrillation (Afib, ICD-9 codes 427.31 and ICD-10 codes I48.x), chronic obstruction pulmonary disease (COPD, ICD-9 codes 490.x to 496.x and ICD-10 codes J44.9), and malignancy (ICD-9). Moreover, patients with at least two records of more than or equal to 126 mg/dL of fasting glucose or more than or equal to 6.5% of glycated hemoglobin from 6 months were also considered to have DM. We also defined at least two records of estimated glomerular filtration rate less than 60 mL/min as CKD. The criteria of

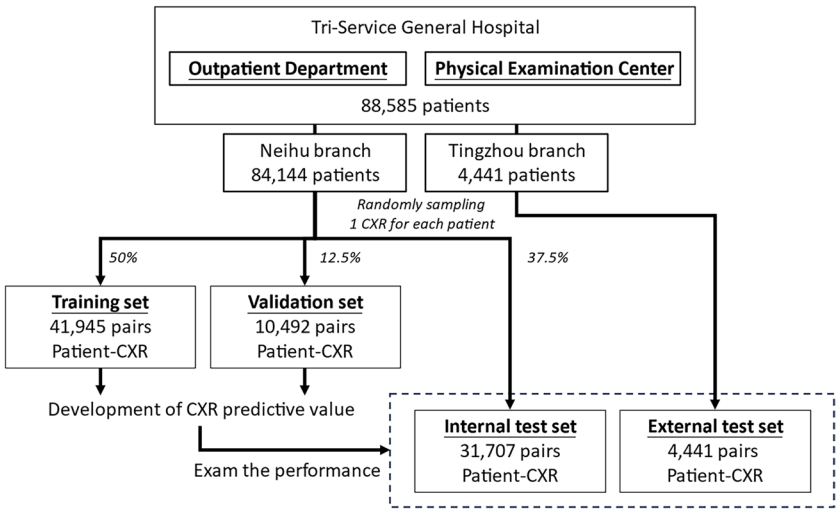


Fig. 1 Generation of datasets for training, validation, internal test, and external test. Schematic of the dataset creation and analysis strategy, which was devised to assure a robust and reliable dataset for training, validating, and testing of the network. Once a patient's data were placed in one of the datasets, that individual's data were used only in that set, avoiding "cross-contamination" among the training, validation, and test datasets. The details of the flow chart and how each of the datasets was used are described in the "Method" section

new diagnoses event were defined as newly recorded ICD code of disease that had not been identified in the medical record before the CXR taken. The time-to-event period was then defined as the interval between CXR taken and disease diagnosed.

We collected the nearest laboratory values from the tests within 1 month, including estimated glomerular filtration rate (eGFR), creatinine (Cr), blood urea nitrogen (BUN), sodium (Na), potassium (K), hemoglobin (Hb), mean corpuscular volume (MCV), white blood cell count (WBC), platelet (PLT), neutrophil (Neut), lymphocyte (Lymph), monocyte (Mono), eosinophil (Eos), basophil (Baso), aspartate aminotransferase (AST), alanine aminotransferase (ALT), albumin (Alb), total bilirubin (TB), triglyceride (TG), total cholesterol (TC), C-reactive protein (CRP), and dipstick test urine protein (uPro).

The missing rate of each variable ranged from 0 to 85%, with an average of 34%. Missing values were imputed using multiple imputations by chained equations [26], which have been proven effective for handling data with high missing rates [27].

The structuralized reports associated to all included CXRs were labeled as "positive" or "negative" of following 31 image findings, which included consolidation change, pneumonia, emphysematous change, pneumothorax, atelectasis, scalloping of the diaphragm, costophrenic angle blunting, pleural effusion, atherosclerosis, cardiomegaly, prominence of hilar shadow, pulmonary edema, aneurysm, degenerative joint disease, fracture, spondylosis, osteophyte formation, osteoporosis, osteoarthritis, widening of the mediastinum, malignancy, inflammatory, pigtail or drainage, sternotomy, port a implantation, perm catheter insertion, pacemaker, tracheostomy, vertebroplasty, endotracheal tube, and nasogastric tube.

Outcomes of Interest

The most important outcome we focused on is all-cause mortality and the median follow-up periods of the mortality event was about 3–4 years. We also collected the newly diagnosed disease, including STK, HF, Afib, CKD, malignancy, and infection disease. All the events mentioned above were defined from electronic medical records in our hospital. Data for alive visits were censored at the patient's last known hospital alive encounter to limit bias from incomplete records.

Clinical Models

To compare and integrate with the VAI, three clinical models, derived from demographic variables, medical history, laboratory variables, and labeled CXR report, were developed through training and validation sets and evaluated in internal and external test sets. The clinical models were built through a forward, stepwise Akaike Information Criterion (AIC) method [28], as illustrated in Online Resource 2, that each variate was added into the model by turns and the variable made the model having minimal AIC in the training set was selected as candidate model in each cycle. The process was repeated until all covariates were used. All candidate models were validated in validation set and the model having minimal AIC value was selected as the final model.

Each clinical and CXR-risk score was used as a single, continuous variable to develop the combined models rather than the components of them as independent variables.

Statistical Analysis

We presented the characteristics of different datasets as means and standard deviations, numbers of patients, or percentages, where appropriate. They were compared using either analysis of variance or the chi-square test, as proper. Since the main purpose of model is screening, sensitivity and specificity were chosen to estimate the performance of AI model. The receiver operating characteristic (ROC) curve and area under curve (AUC) were used to present the ability to detect mortality events at 3 months, 1 year, and 5 years. Concordance index, which is a proportion with value from 0 to 1, was used to present the mortality prediction results of model for better readability. It is defined as the proportion of all usable patient pairs in which the predictions and outcomes are concordant, indicating the probability that a patient with a longer predicted survival time lives longer than a patient with a shorter predicted survival time. The use of the concordance index facilitates clinicians in estimating the confidence level regarding whether high-risk patients have a higher risk in real clinical settings, thereby guiding adjustments in clinical practice. The ΔAIC [29] and likelihood ratio test [30] were used to statistically compare the discriminatory power of non-nested and nested models, when appropriate. The best model within all non-nested model was forced to have $\Delta AIC = 0$. Models having $\Delta AIC \leq 2$ have substantial support, those in which $4 \leq \Delta AIC \leq 7$ have considerably less support, and models having $\Delta AIC > 10$ have essentially no support. Risk stratification analysis was done by setting cut points at 50th and 85th percentile of CXR-risk score in validation set which were then applied to both of internal and external test sets. Kaplan–Meier survival analysis and Cox proportional hazard models were applied in following survival analysis. All statistical analyses were completed in R version 3.4.4. The significance level was set as $p < 0.05$.

Results

Table 1 presents the baseline characteristics of population in training, validation, internal test, and external test sets. Among our study population, more than half of them, ranging from 55.2 to 73.4%, have at least one morbidity.

We first examined the performance of "CXR-risk score" produced by proposed VAI in mortality event prediction. Figure 2a shows the area under curve of receiver operating characteristic curve of 3-month, 1-year, and 5-year mortality. In the internal test set, the AUCs were 0.920, 0.891, and 0.874 for mortality prediction at 3 months, 1 year, and 5 years, respectively. The results in the external test set were 0.947, 0.881, and 0.881, correspondingly. The concordance index for model assessment is presented in Fig. 2b and the values in internal and external test sets were 0.859 (95% confidence interval (CI), 0.851–0.867) and 0.870 (95% CI, 0.844–0.896), respectively. For risk stratification and following survival analysis, the cut points were selected at 50th and 85th percentile in the validation set of standardized CXR-risk score as shown in Fig. 2c. The distributions of three risk groups in two test sets are shown in Fig. 2d. A total of 49.7%, 35.6%, and 14.6% of population in internal test set were classified into a "Low," "Moderate," and "High" risk group. In external test set, the number was 49.7%, 35.6%, and 14.6% of each group, respectively.

After risk stratification, the result of survival analysis is illustrated in Fig. 3. The "Low" risk group was set as the reference group in this analysis. The hazard ratios of the "High" and "Moderate" risk group were 35.74 (95% CI, 30.27–42.20) and 6.05 (95% CI, 5.09–7.20) in internal test set and 51.07 (95% CI, 25.85–100.90) and 9.17 (95% CI, 4.56–18.47) in external test set. And we also present the Kaplan–Meier curve of newly developed morbidities and infectious disease of each risk group in Online Resource 3. In summary, this risk score can be widely used in multiple clinical outcomes.

Another stratification analysis was done according to population's demography and underlying morbidity and is presented in Fig. 4. The performance of CXR-risk score had concordance index between 0.710 and 0.903 among all subgroups in both internal and external test sets. The most significant performance reduction was shown in patient with history of AMI (internal, 0.815; external, 0.741). The similar performances were presented in the remaining subgroups.

In Fig. 5, we presented the concordance index of each variable from training set. The development of the "Clinical-risk" model is showed in Fig. 5a and all of the demographic variants, age, gender, BMI, and five diseases were included. Age and malignancy were chosen in the early phase of model development. Figure 5b reveals the concordance index of the "Clinical-risk" model in internal and external test sets, which were 0.847 (95% CI, 0.839–0.854) and 0.866 (95% CI, 0.844–0.887), respectively. Figure 5c demonstrates the development of the "Lab-risk" model which finally used eight parameters, including eGFR, albumin, and hemoglobin, in its coefficients. The concordance index of the "Lab-risk" model in internal and external test sets was 0.769 (95% CI, 0.758–0.779) and 0.808 (95% CI, 0.776–0.840), as presented in Fig. 5d. The "Report-risk" is presented in Fig. 5e and f, 15 out of 31 image findings included in the model, such as "costophrenic angle blunting," "Port-A implantation," and "atherosclerosis." The concordance index of "Report-risk" model in internal and external test sets was 0.826 (95% CI, 0.817–0.834) and 0.838 (95% CI, 0.810–0.867). This result shows comparable ability to predict mortality by single "CXR-risk" compared with those clinical and laboratory data.

The comparison of the performance of a VAI-enhanced "CXR-risk" model and clinical models, "Clinical-risk," "Lab-risk," and "Report-risk" model, is demonstrated in Fig. 6. The "CXR-risk" model exhibits the highest predictive ability for mortality compared to the "Clinical-risk," "Lab-risk," and "Report-risk" models. We found that the "CXR-risk" model gains significant improvement in performance after combination with each of the three clinical models. The combination of "CXR-risk" and "Clinical-risk" yielded concordance index of 0.883 (95% CI, 0.877–0.889) in internal test set and 0.893 (95% CI, 0.872–0.915) in external test set. About combination of "CXR-risk" and "Lab-risk," the values were 0.866 (95% CI, 0.859–0.874) in internal test set and 0.883 (95% CI, 0.858–0.908) in external test set. The concordance values of "CXR-risk" plus "Report-risk" in internal and external test sets were 0.874 (95% CI, 0.867–0.880) and 0.879 (95% CI, 0.855–0.904), respectively. Combining all of "Clinical-risk," "Lab-risk," and "Report-risk" to "CXR-risk" showed significant improvement in model performance in both internal and external test sets with concordance about 0.888 (95% CI, 0.881–0.894) and 0.900 (95% CI, 0.878–0.921), respectively.

Discussion

In this study, we introduced a VAI-based CXR-risk scoring system which demonstrated a strong correlation with the risk of mortality events, successfully identifying high-risk populations within the general population. Additionally, this model proved to be effective in risk stratification as compared with the model developed from clinical demography, laboratory variant, and image reports. Moreover, combining these models can result in better performance than each one individually. Notably, the VAI-base CXR-risk score not only excelled in predicting mortality but also showed correlation with the risk of developing new morbidities in the future. Also, our study demonstrates a novel way with great accessibility and serviceability of non-numeric data quantification and combination with other clinical values through a simple liner model.

Table 1 Baseline characteristics

	Training ($n=41,945$)	Validation ($n=10,492$)	Internal test ($n=31,707$)	External test ($n=4441$)	p-value
Demography					
Gender (male)	21,397 (51%)	5424 (51.7%)	16,188 (51.1%)	2112 (47.6%)	< 0.001
Age (years)	56.1 ± 17.7	56.1 ± 17.5	56.2 ± 17.6	59.2 ± 18.7	< 0.001
BMI (kg/m^2)	24 ± 4.2	24.1 ± 4.2	24.1 ± 4.2	23.9 ± 4.3	0.019
Comorbidity					
AMI	538 (1.3%)	125 (1.2%)	385 (1.2%)	25 (0.6%)	0.001
STK	3287 (7.8%)	813 (7.7%)	2522 (8%)	459 (10.3%)	< 0.001
CAD	7017 (16.7%)	1728 (16.5%)	5240 (16.5%)	868 (19.5%)	< 0.001
HF	2206 (5.3%)	538 (5.1%)	1659 (5.2%)	299 (6.7%)	< 0.001
Afib	1121 (2.7%)	297 (2.8%)	841 (2.7%)	136 (3.1%)	0.349
DM	6954 (16.6%)	1765 (16.8%)	5380 (17%)	955 (21.5%)	< 0.001
HTN	13,184 (31.4%)	3249 (31%)	9979 (31.5%)	1872 (42.2%)	< 0.001
CKD	4969 (11.8%)	1265 (12.1%)	3726 (11.8%)	552 (12.4%)	0.546
HLP	11,957 (28.5%)	2840 (27.1%)	9079 (28.6%)	1679 (37.8%)	< 0.001
COPD	7423 (17.7%)	1814 (17.3%)	5475 (17.3%)	832 (18.7%)	0.067
Malignancy	7544 (18%)	1965 (18.7%)	5624 (17.7%)	902 (20.3%)	< 0.001
Morbidity ≥ 1	26,496 (63.2%)	6558 (62.5%)	20,030 (63.2%)	3168 (71.3%)	< 0.001
Laboratory data					
WBC (10^3/μL)	8.1 ± 6.7	8 ± 4.2	8 ± 4.7	7.9 ± 3.5	0.084
Hb (g/dL)	12.5 ± 2.2	12.5 ± 2.2	12.5 ± 2.2	12.8 ± 2.1	< 0.001
MCV (fl)	89 ± 7.5	89 ± 7.3	88.9 ± 7.5	89.1 ± 7.3	0.444
PLT (10^3/μL)	230.6 ± 90	231.3 ± 90.9	230.3 ± 90.2	220.4 ± 83.2	< 0.001
Neut (%)	68.4 ± 14	68.2 ± 13.9	68.4 ± 14	68.4 ± 13.7	0.759
Lymph (%)	22.7 ± 11.6	22.7 ± 11.6	22.7 ± 11.7	23 ± 11.6	0.441
Mono (%)	6.2 ± 3.2	6.3 ± 3.2	6.2 ± 3.1	6.2 ± 3	0.025
Eos (%)	2 ± 2.4	2 ± 2.5	2 ± 2.4	2 ± 2.4	0.275
Baso (%)	0.5 ± 0.5	0.5 ± 0.5	0.4 ± 0.5	0.4 ± 0.4	0.09
Na (mmol/L)	137.9 ± 4.9	137.9 ± 4.8	138 ± 4.8	137.8 ± 4.6	0.055
K (mmol/L)	3.8 ± 0.5	3.8 ± 0.5	3.8 ± 0.5	3.9 ± 0.5	0.262
Cr (mg/dL)	1 ± 1.1	1 ± 1.1	1 ± 1.1	1 ± 0.9	0.016
eGFR (mL/min/1.73m^2)	88.8 ± 27.1	89 ± 27.3	88.6 ± 27.1	85.4 ± 25.8	< 0.001
AST (U/L)	32.4 ± 123.5	32.5 ± 114.1	31.2 ± 115.1	29.5 ± 70.6	0.247
ALT (U/L)	29.3 ± 100.7	30.3 ± 99.4	28.5 ± 89.5	27.2 ± 54.5	0.177
Alb (g/dL)	3.8 ± 0.7	3.8 ± 0.7	3.8 ± 0.7	3.8 ± 0.6	0.757
TB (mg/dL)	0.8 ± 0.5	0.8 ± 0.6	0.8 ± 0.5	0.8 ± 0.5	0.107
HbA1c (%)	6.4 ± 1.4	6.4 ± 1.4	6.5 ± 1.4	6.4 ± 1.4	0.545
TG (g/L)	119.7 ± 85	119.2 ± 78.3	119.8 ± 82.8	120.1 ± 77.5	0.893
TC (g/L)	167.4 ± 44.6	167.1 ± 44.8	167.5 ± 44.7	168.3 ± 46.8	0.564
LDL (mg/dL)	101 ± 36	100.9 ± 35.7	101 ± 36.2	101 ± 37.3	0.996
HDL (mg/dL)	47.2 ± 15.5	46.8 ± 15.6	47.1 ± 15.5	47.8 ± 15.3	0.004
CRP (mg/dL)	3.3 ± 5.2	3.3 ± 5.2	3.3 ± 5.3	2.8 ± 4.6	< 0.001
uPro (+)	11,178 (26.6%)	2802 (26.7%)	8515 (26.9%)	1188 (26.8%)	0.941

Abbreviation: *BMI*, body mass index; *AMI*, acute myocardial infarction; *CAD*, coronary artery disease; *HF*, heart failure; *Afib*, atrial fibrillation; *DM*, diabetes mellitus; *HTN*, hypertension; *CKD*, chronic kidney disease; *HLP*, hyperlipidemia; *COPD*, chronic obstructive pulmonary disease; *eGFR*, estimated glomerular filtration rate; *Cr*, creatinine; *BUN*, blood urea nitrogen; *Na*, sodium; *K*, potassium; *Hb*, hemoglobin; *MCV*, mean corpuscular volume; *WBC*, white blood cell count; *PLT*, platelet; *Neut*, neutrophil; *Lymph*, lymphocyte; *Mono*, monocyte; *Eos*, eosinophil; *Baso*, basophil; *AST*, aspartate aminotransferase; *ALT*, alanine aminotransferase; *Alb*, albumin; *TB*, total bilirubin; *TG*, triglyceride; *TC*, total cholesterol; *CRP*, C-reactive protein; *uPro*, urine protein (dipstick test)

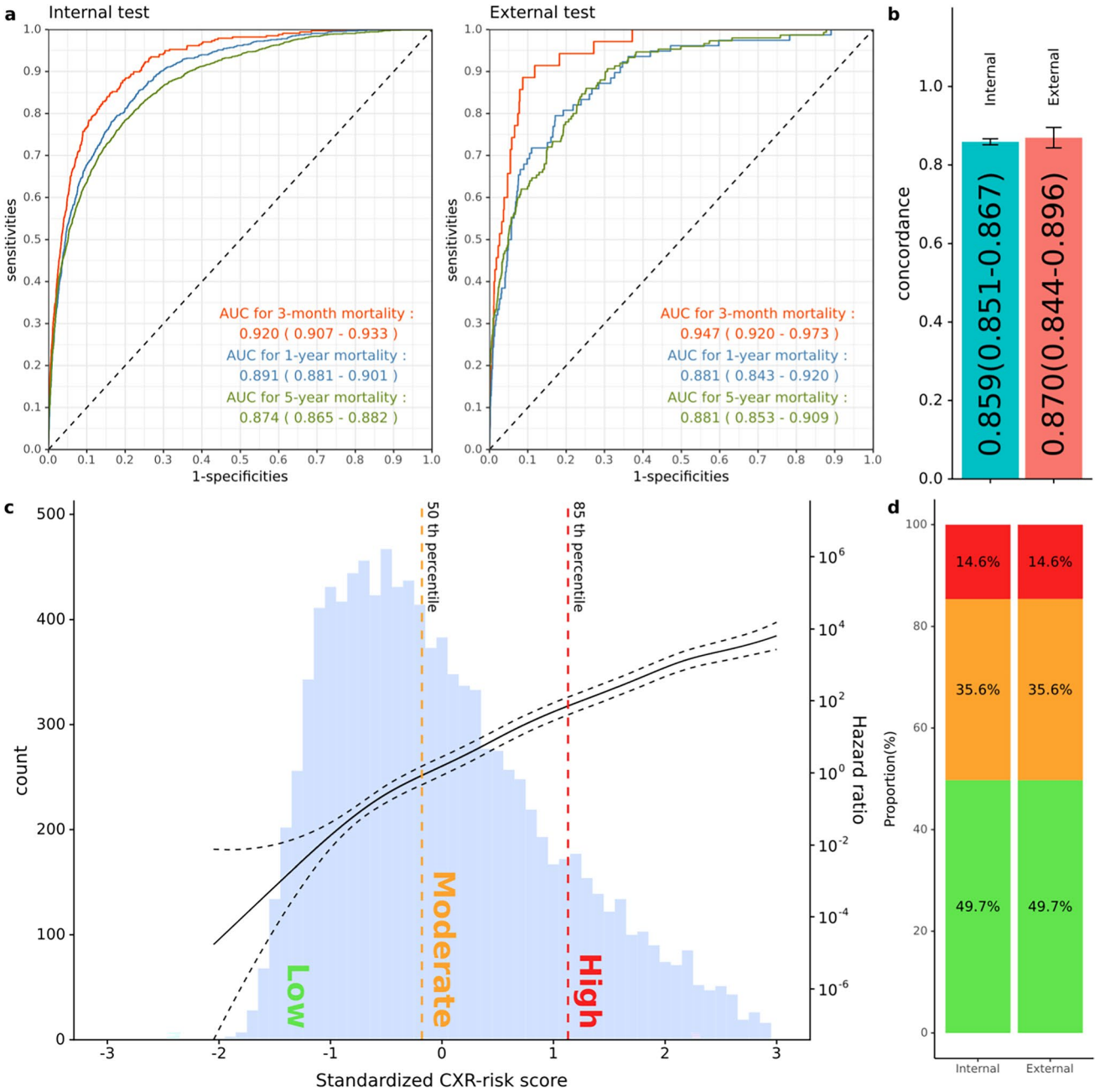


Fig. 2 Performance of CXR-risk score in mortality analysis and risk stratification. **a** Receiver operating characteristic curve and area under curve with 95% confidence interval of 1-year, 5-year, and overall mortality of CXR-risk score in internal and external test sets. **b** Concordance index analysis of CXR-risk score in all-cause mortality in internal and external test sets. **c** Distribution of population according to standardized CXR-risk score in validation set and the curve of corresponding hazard ratio. Operating point to distinguish low-, moderate-, and high-risk group were selected at where hazard ration was 50th and 85th percentile. **d** According to the operating point defined previously, the patients in internal and external validation sets were classified as low-, moderate-, and high-risk groups for following analyses

In the four models developed in our study, the "Lab-risk" yielded the worst performance among all models, which may be due to relatively large variations throughout the time and hard to predict long-term outcome with single cross-sectional data. The variants chosen by the model, such as albumin and hemoglobin level, were relatively stable in outpatient setting and considered indicative of general conditions rather than specific disease. Hypoalbuminemia had

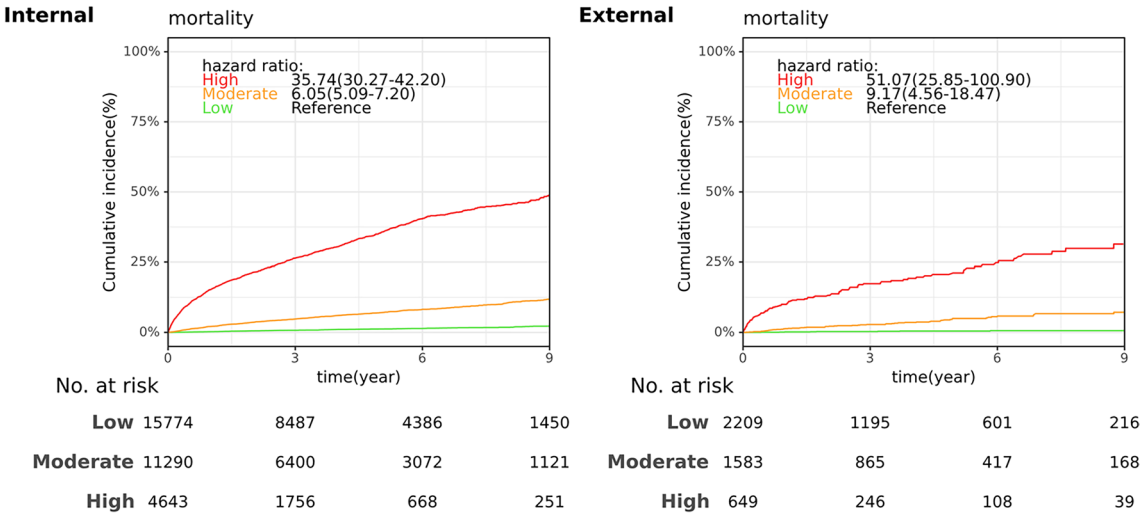


Fig. 3 Kaplan–Meier curves for each risk stratification on all-cause mortality. The analyses are conducted both in internal and external test sets. The table shows the of number at-risk population in each risk stratification

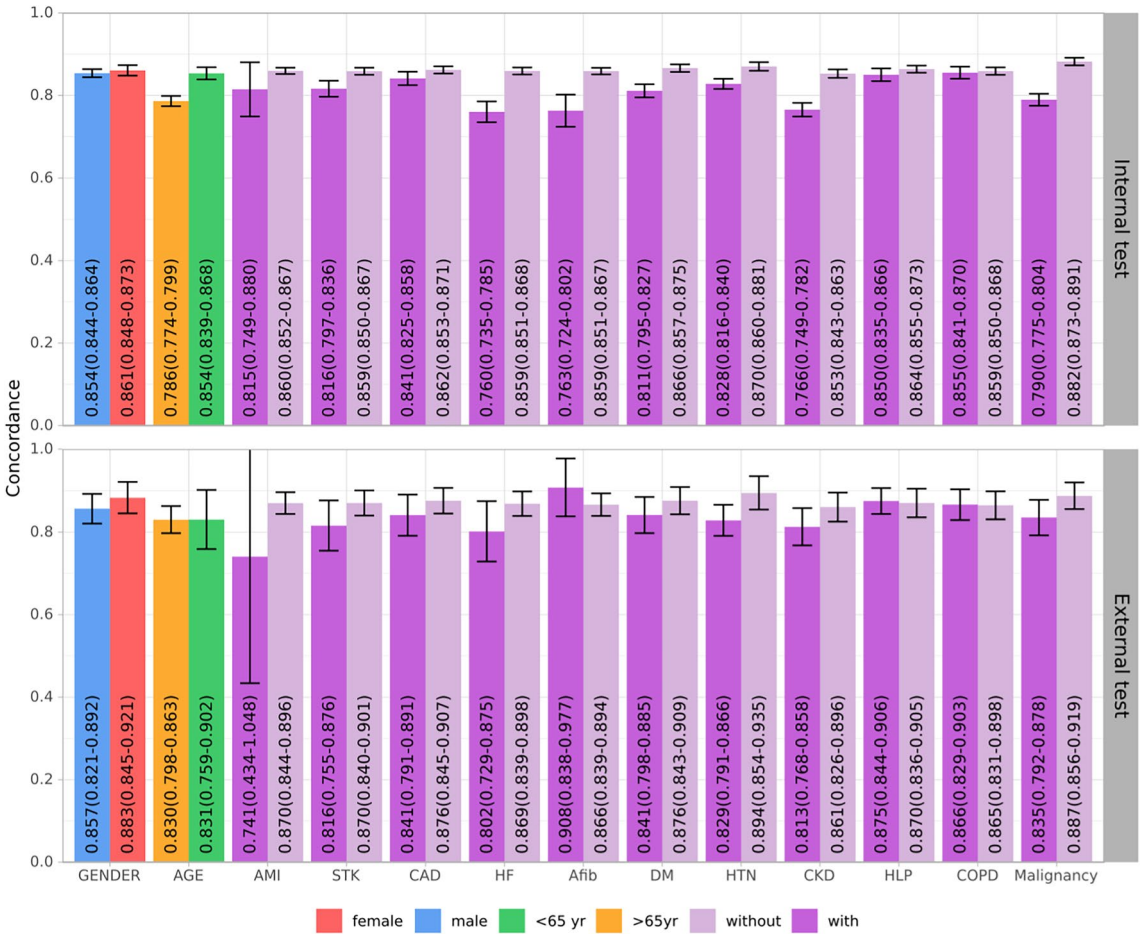


Fig. 4 Stratified analysis of concordance index in different subgroup on long-term all-cause mortality. Abbreviations: BMI, body mass index; AMI, acute myocardial infarction; CAD, coronary artery disease; HF, heart failure; Afib, atrial fibrillation; DM, diabetes mellitus; HTN, hypertension; CKD, chronic kidney disease; HLP, hyperlipidemia; COPD, chronic obstructive pulmonary disease

been proved to be associated with the mortality from multiple causes, including but not limited to sepsis, heart, kidney, stoke, and malignancy [31–33]. A similar effect was found in hemoglobin, where the presence of anemia was linked to an increased risk of hospitalization and mortality [34], particularly from sepsis [35], cardiac, and malignant disease [36].

The "Clinical-risk" presented the closest performance to "CXR-risk." This result was consistent with the data, which AUC of DLM 0.75 vs. clinical risk 0.76, reported by Lu et al. [17] and pointed out the considerable information carried by demographic data and medical history themselves when compared with either lab data or image report alone.

A radiologist-based CXR scoring system had been developed for mortality risk stratification in various diseases [37–39]. Most of the scoring systems relied on the structural abnormalities such as infiltration of interstitial or alveolar, degree of fibrosis, and vascular calcification. These features stood for the severity of specific disease and, unsurprisingly, were correlated with mortality. In our study, the "Report-risk," developed through specific image findings, also showed its ability in mortality prediction. However, its performance was inferior to our CXR-risk score, probably due to its binary coding of image features and lack of detail information.

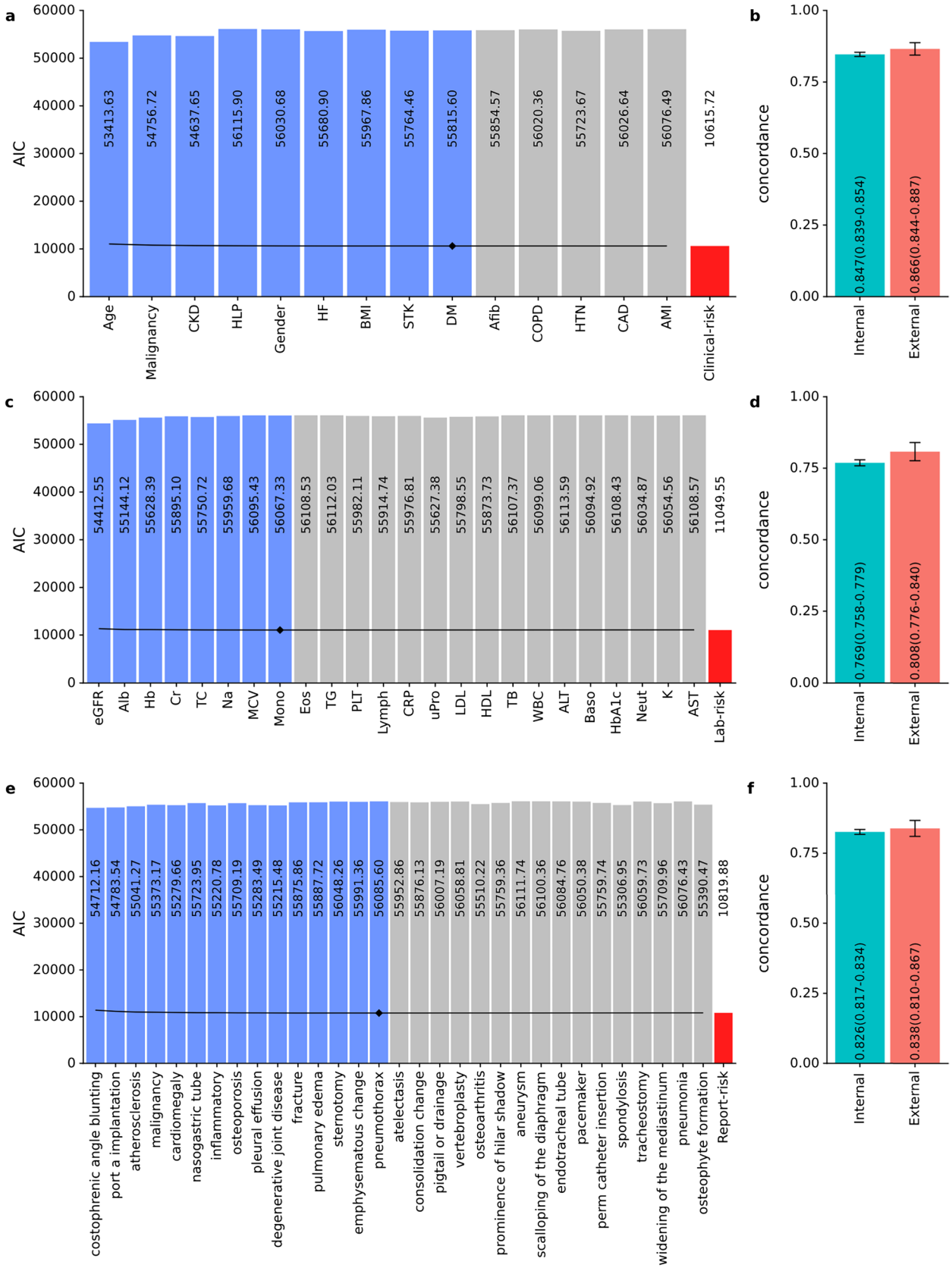
Our study is not the first one to apply deep learning in mortality prediction with CXR. A study conducted by Lu et al. [17] reported that addition of DLM (Inception v4) risk score to model established through clinical risk factors or radiographic findings improved the performance, AUC from 0.58 and 0.76 to 0.74 and 0.78 for 12-year mortality. There was no comparison between combined model and CXR-risk, whose AUC was about 0.75. Another study conducted by Kim et al. [40] used DenseNet121 to predict 30-day mortality in patients with community-acquired pneumonia (CAP) and revealed AUCs about 0.8 in two external test cohorts, while the value of CURB-65 is 0.73. Also, they combined the DLM prediction and CURB-65 together, like our study, but failed to achieve significant improvements in model capability. The research mentioned above suggested that DLM-assisted CXR-risk score performed better than usual clinical factors. The varying results of combined models across studies may stem from differences in study populations and the specific prediction targets. For instance, in Kim's study focusing on CAP patients, CXR alone provided sufficient information, rendering additional benefits from combining CURB-65 negligible. In contrast, our study encompassed a general population with broader and higher information requirements, where the CXR-risk score derived substantial benefit from integration with clinical models. Furthermore, while Kim's DLM and CURB-65 models shared largely overlapping domain knowledge related to pneumonia, our study's models encompassed non-overlapping domains such as medical history, laboratory examinations, and imaging studies. Therefore, it is reasonable to infer that the integrated model derives benefits from the expanded knowledge base obtained through the inclusion of diverse models.

Although our "CXR-risk" score is not specific to any disease, it serves as a reflection of each patient's general condition. It (CXR-risk score) is similar to what performance status, evaluated by either Eastern Cooperative Oncology Group (ECOG) performance status [41] or Karnofsky Performance Status [42], dose and performance status has been proved to be correlate with the survival in cancer patients [43] through large population data and significantly influenced the cancer-associated treatment guidelines. The CXR-risk score, which integrates information through a deep learning process, could function similarly to performance status and provide valuable guidance for clinical decision-making.

Beyond current physical conditions, the "high-risk" population identified by "CXR-risk" in our study exhibited more than a tenfold risk of developing newly onset heart failure, atrial fibrillation, and CKD, which is similar to previous research in a DLM-enabled electrocardiogram [44]. As mortality is the ultimate outcome of all diseases, DLMs trained on mortality labels also have the potential to predict other diseases simultaneously. With the assistance of proposed VAI-based "CXR-risk" score, the first-line professionals can adjust the intensity and the goals of clinical management to achieve a favorable outcome in patients with multimorbidity.

Given the advantages of CXR-risk score, our system can find application in various clinical conditions. In the outpatient department, the CXR-risk score, along with clinical information, can help the physicians in adjusting the therapeutic strategies and evaluating the interval change during short visiting. Individuals with a high mortality risk can be referred to further detail examinations. In the location with limited healthcare resource, the CXR-risk score can offer preliminary information to the physicians during the initial encounter. Patients identified as high-risk of mortality through CXR-risk score are advised to undergo comprehensive physical examination, including laboratory tests and advance image studies. A previous systematic review [45] had shown that general health checks provided no benefit to unselected adults, but the study from "Lifetime Transition Period Health Screening" in South Korea [46], focusing on people older than 66 years, demonstrated benefit in reducing stroke, myocardial infarction, and death. This finding suggested that health checkups are more likely to be beneficial for high-risk populations.

With the help of a foundation model, the VAI significantly outperformed clinical models. However, the VAI and clinical model are not developed to repel each other. As discussed before, the domain knowledge learned by VAI and clinical models could be either the same or different. Models with


a
AIC
60000
50000
40000
30000
20000
10000
0
53413.63
54756.72
54637.65
56115.90
56030.68
55680.90
55967.86
55764.46
55815.60
55854.57
56020.36
55723.67
56026.64
56076.49
10615.72
Age
Malignancy
CKD
HLP
Gender
HF
BMI
STK
DM
Afib
COPD
HTN
CAD
AMI
Clinical-risk
b
concordance
1.00
0.75
0.50
0.25
0.00
0.847(0.839-0.854)
0.866(0.844-0.887)
Internal
External
c
AIC
54412.55
55144.12
55628.39
55895.10
55750.72
55959.68
56095.43
56067.33
56108.53
56112.03
55982.11
55914.74
55976.81
55627.38
55798.55
55873.73
56107.37
56099.06
56113.59
56094.92
56108.43
56034.87
56054.56
56108.57
11049.55
eGFR
Alb
Hb
Cr
TC
Na
MCV
Mono
Eos
TG
PLT
Lymph
CRP
uPro
LDL
HDL
TB
WBC
ALT
Baso
HbA1c
Neut
K
AST
Lab-risk
d
concordance
0.769(0.758-0.779)
0.808(0.776-0.840)
Internal
External
e
AIC
54712.16
54783.54
55041.27
55373.17
55279.66
55723.95
55220.78
55709.19
55283.49
55215.48
55875.86
55887.72
56048.26
55991.36
56085.60
55952.86
55876.13
56007.19
56058.81
55510.22
55759.36
56111.74
56100.36
56084.76
56050.38
55759.74
55306.95
56059.73
55709.96
56076.43
55390.47
10819.88
costophrenic angle blunting
port a implantation
atherosclerosis
malignancy
cardiomegaly
nasogastric tube
inflammatory
osteoporosis
pleural effusion
degenerative joint disease
fracture
pulmonary edema
sternotomy
emphysematous change
pneumothorax
atelectasis
consolidation change
pigtail or drainage
vertebroplasty
osteoarthritis
prominence of hilar shadow
aneurysm
scalloping of the diaphragm
endotracheal tube
pacemaker
perm catheter insertion
spondylosis
tracheostomy
widening of the mediastinum
pneumonia
osteophyte formation
Report-risk
f
concordance
0.826(0.817-0.834)
0.838(0.810-0.867)
Internal
External

◂**Fig. 5** AIC analysis of each demographic and laboratory covariate on all-cause mortality in validation set and the development of Clinical-, Lab-, and Report-risk score. **a**, **c**, **e** AIC analysis of each covariate acquired from training set (blue and gray bar). The risk score was developed through repeated process that selected covariate was added into the model to achieve minimal AIC in training set. The detail process of development is described in Online Resource 2. The candidate models acquired in previous process were then validated in validation set and the concordance values were presented as black curve with 95% confidant interval (dashed line). After completing of the process, the models having minimal AIC value were used and labeled as the "Clinical," "Lab," and "CXR report" risk score (red). The variables used in the final model were labeled as blue and those not used were gray. **b**, **d**, **f** Concordance index analysis of the "Clinical," "Lab," and "CXR report" risk score in internal and external test sets

different domain knowledge can work together like human's brain area and enable the combined model to provide accurate prediction in high information-requiring condition. Disposing the clinical risk score is unwise due to its wide application in daily practice and the advantage gained from combining it to a VAI model. Considering the rapid growth of a large, pre-trained model in the community during these years, we can add not only image but also any non-numeric data, including but not limited to text, video, and audio, into current risk models through corresponding DLM. Through the application of adequate pre-trained model, there is no need to collect large amounts of training data and set up the training equipment, like a high-level GPU and workstation.

Several limitations should be considered in this study. First, the single-center, retrospective design of the study introduces the risk of bias in various aspects. The nature of a single-center study may cause selection bias, and caution should be exercised when applying the results to a non-selective population. The study population was confined to patients from the outpatient department and physical examination center who were relatively stable in their clinical conditions. The application of our CXR-risk score to more critically ill patients, such as those in the emergency department, had not yet been validated. Additionally, missing values, which were addressed through effective imputation, still influence the results since the actual data is unknown. Secondly, the morbidities included in our study primarily pertained to the cardiovascular and pulmonary systems. Other diseases, such as liver cirrhosis or connective tissue disease, which are encompassed in the Charlson Comorbidity Index, were not considered in our analysis. Thirdly, the method of model construction was selected based on minimizing computational load. Therefore, alternative algorithms such as pre-trained or naïve models, the use of fine-tuning, and different stepwise selection methods (forward, backward, or bidirectional) were not tested and analyzed in our study. Fourthly, the prediction of VAI is neither disease nor time-specific, and future effort is needed to provide more specific clinical suggestions. And we did not explore how the VAI interpreted and where it focused on the CXR in our study. Consequently, the explainability is

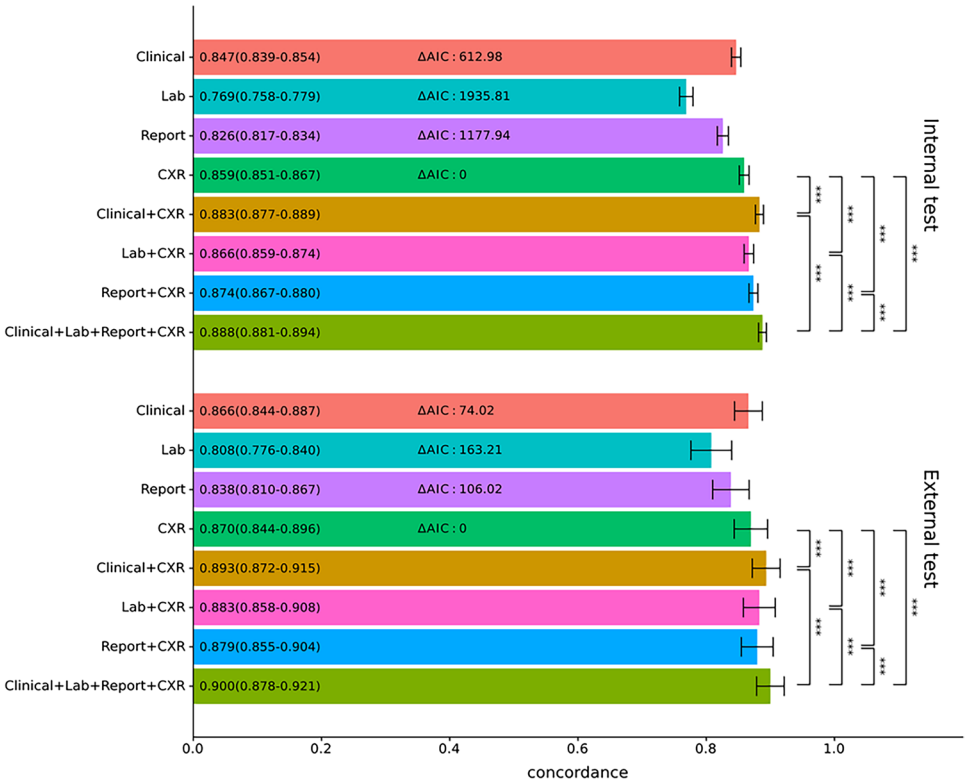


Fig. 6 Concordance index and AIC analysis of different combination model of CXR, Lab, and CXR report—risk score on all-cause mortality. The performance of three models, "CXR," "Report," "Clinical," and "Lab," and their combinations in internal and external test sets. *p*-value, < 0.05 *, < 0.01 **, < 0.001 ***; ns, non-significant

limited and the relationship between image patterns and the risk of future mortality remains unclear.

Conclusion

In conclusion, the interpretation of a single CXR image by foundation VAI offers a means of risk stratification and prognosis prediction in sub-healthy groups which outperformed usual clinical indicators from demographics, medical history, laboratory, and radiographic exam. The combination of VAI-based model and clinical models can provide a more reliable result than each alone. Also, we provided a novel and convenient method, which allows non-numeric resources, radiologic image in this study, to be quantified and plugged into currently used risk tools, to provide an accurate assessment of clinical outcomes and valuable support in decision-making.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10278-024-01247-y.

Author Contribution Yu-Cheng Chen, Wen-Hui Fang, Chin-Sheng Lin, Yung-Tsai Lee, and Chin Lin contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Dung-Jang Tsai, Chih-Wei Hsian, Cheng-Kuang Chang, Kai-Hsiung Ko, and Guo-Shu Huang. The first draft of the manuscript was written by Yu-Cheng Chen and Chin Lin. Yu-Cheng Chen and Chin Lin provided a deep learning model for stratifying high risks of mortality. Yu-Cheng Chen analyzed the data. Wen-Hui Fang, Chin-Sheng Lin, and Yung-Tsai Lee revised the manuscript for important intellectual content. Yung-Tsai Lee and Chin Lin take final responsibility for this article and provided the final approval of the version to be published.

Funding This study was supported by funding from the Medical Affairs Bureau, Taiwan (MND-MAB-C13-112050, and MND-MAB-C07-113021 to Chin Lin), and the Cheng Hsin General Hospital, Taiwan (CHNDMC-112–05 and CHNDMC-113–11205 to C. Lin).

Data Availability The data that support the findings of this study are available on request from the corresponding author, Dr. Chin Lin, upon reasonable request.

Declarations

Ethics Approval This study was ethically approved by the institutional review board (IRB NO. C202105150) in Tri-Service General Hospital, Taipei, Taiwan.

Competing Interests The authors declare no competing interests.

References

1. World Health Organization: World report on ageing and health: World Health Organization, 2015
2. Pati S, et al.: Prevalence and outcomes of multimorbidity in South Asia: a systematic review. BMJ Open 5:e007235, 2015
3. Chen H, Cheng M, Zhuang Y, Broad JB: Multimorbidity among middle-aged and older persons in urban China: Prevalence, characteristics and health service utilization. Geriatrics & gerontology international 18:1447-1452, 2018
4. Xu X, Mishra GD, Jones M: Evidence on multimorbidity from definition to intervention: an overview of systematic reviews. Ageing research reviews 37:53-68, 2017
5. Glynn LG, et al.: The prevalence of multimorbidity in primary care and its effect on health care utilization and cost. Family practice 28:516-523, 2011
6. Hubbell FA, Greenfield S, Tyler JL, Chetty K, Wyle FA: The impact of routine admission chest x-ray films on patient care. New England Journal of Medicine 312:209-213, 1985
7. Delrue L, Gosselin R, Ilsen B, Landeghem AV, Mey Jd, Duyck P: Difficulties in the interpretation of chest radiography: Springer, 2011
8. Wang Y, Xu L, Zou Q, Lin C: prPred-DRLF: Plant R protein predictor using deep representation learning features. Proteomics 22:e2100161, 2022
9. Jumper J, et al.: Highly accurate protein structure prediction with AlphaFold. Nature 596:583-589, 2021
10. Wysocka M, Wysocki O, Zufferey M, Landers D, Freitas A: A systematic review of biologically-informed deep learning models for cancer: fundamental trends for encoding and interpreting oncology data. BMC Bioinformatics 24:198, 2023
11. Feng C, et al.: Prediction of early hematoma expansion of spontaneous intracerebral hemorrhage based on deep learning radiomics features of noncontrast computed tomography. Eur Radiol 34:2908-2920, 2024
12. Chan HP, Samala RK, Hadjiiski LM, Zhou C: Deep Learning in Medical Image Analysis. Adv Exp Med Biol 1213:3-21, 2020
13. Niu S, et al.: Enhancing healthcare decision support through explainable AI models for risk prediction. Decision Support Systems 181:114228, 2024
14. Çallı E, Sogancioglu E, van Ginneken B, van Leeuwen KG, Murphy K: Deep learning for chest X-ray analysis: A survey. Medical Image Analysis 72:102125, 2021
15. Rajpurkar P, et al.: Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS medicine 15:e1002686, 2018
16. Anis S, et al.: An overview of deep learning approaches in chest radiograph. IEEE Access 8:182347-182354, 2020
17. Lu MT, Ivanov A, Mayrhofer T, Hosny A, Aerts HJ, Hoffmann U: Deep learning to assess long-term mortality from chest radiographs. JAMA network open 2:e197416-e197416, 2019
18. Kwon YJF, et al.: Combining Initial Radiographs and Clinical Variables Improves Deep Learning Prognostication in Patients with COVID-19 from the Emergency Department. Radiol Artif Intell 3:e200098, 2021
19. Knaus WA, Draper EA, Wagner DP, Zimmerman JE: APACHE II: a severity of disease classification system. Critical care medicine 13:818-829, 1985
20. Headley J, Theriault R, Smith TL: Independent validation of APACHE II severity of illness score for predicting mortality in patients with breast cancer admitted to the intensive care unit. Cancer 70:497-503, 1992
21. Charlson ME, Pompei P, Ales KL, MacKenzie CR: A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal of chronic diseases 40:373-383, 1987
22. Quan H, et al.: Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. American journal of epidemiology 173:676-682, 2011
23. Moor M, et al.: Foundation models for generalist medical artificial intelligence. Nature 616:259-265, 2023

24. Tiu E, Talius E, Patel P, Langlotz CP, Ng AY, Rajpurkar P: Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nat Biomed Eng 6:1399-1406, 2022
25. Raghunath S, et al.: Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. Nat Med 26:886-891, 2020
26. Buuren Sv, Groothuis-Oudshoorn K: mice: Multivariate Imputation by Chained Equations inR. Journal of Statistical Software 45, 2011
27. Azur MJ, Stuart EA, Frangakis C, Leaf PJ: Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 20:40-49, 2011
28. Yamashita T, Yamashita K, Kamimura R: A Stepwise AIC Method for Variable Selection in Linear Regression. Communications in Statistics - Theory and Methods 36:2395-2403, 2007
29. Burnham KP, Anderson DR: Multimodel Inference. Sociological Methods & Research 33:261-304, 2016
30. Lewis F, Butler A, Gilbert L: A unified approach to model selection using the likelihood ratio test. Methods in Ecology and Evolution 2:155-162, 2010
31. Akirov A, Masri-Iraqi H, Atamna A, Shimon I: Low Albumin Levels Are Associated with Mortality Risk in Hospitalized Patients. Am J Med 130:1465 e1411–1465 e1419, 2017
32. Goldwasser P, Feldman J: Association of serum albumin and mortality risk. Journal of Clinical Epidemiology 50:693-703, 1997
33. Kendall H, Abreu E, Cheng AL: Serum Albumin Trend Is a Predictor of Mortality in ICU Patients With Sepsis. Biol Res Nurs 21:237-244, 2019
34. Culleton BF, Manns BJ, Zhang J, Tonelli M, Klarenbach S, Hemmelgarn BR: Impact of anemia on hospitalization and mortality in older adults. Blood 107:3841-3846, 2006
35. Jiang Y, et al.: Inflammatory anemia-associated parameters are related to 28-day mortality in patients with sepsis admitted to the ICU: a preliminary observational study. Annals of intensive care 9:1-11, 2019
36. Gupta K, et al.: Anemia, Mortality, and Hospitalizations in Heart Failure With a Preserved Ejection Fraction (from the TOPCAT Trial). Am J Cardiol 125:1347-1354, 2020
37. Borghesi A, Maroldi R: COVID-19 outbreak in Italy: experimental chest X-ray scoring system for quantifying and monitoring disease progression. Radiol Med 125:509-513, 2020
38. Abdelmalek JA, Stark P, Walther CP, Ix JH, Rifkin DE: Associations between coronary calcification on chest radiographs and mortality in hemodialysis patients. Am J Kidney Dis 60:990-997, 2012
39. Kirkil G, Lower EE, Baughman RP: Predictors of Mortality in Pulmonary Sarcoidosis. Chest 153:105-113, 2018
40. Kim C, et al.: A Deep Learning Model Using Chest Radiographs for Prediction of 30-Day Mortality in Patients With Community-Acquired Pneumonia: Development and External Validation. AJR Am J Roentgenol 221:586-598, 2023
41. Oken MM, et al.: Toxicity and response criteria of the Eastern Cooperative Oncology Group. American Journal of Clinical Oncology 5:649-656, 1982
42. Karnofsky D, Burchenal J: Evaluation of chemotherpeutic agents. NY, Columbia University, New York 19, 1949
43. Peus D, Newcomb N, Hofer S: Appraisal of the Karnofsky Performance Status and proposal of a simple algorithmic system for its evaluation. BMC Med Inform Decis Mak 13:72, 2013
44. Tsai DJ, et al.: Mortality risk prediction of the electrocardiogram as an informative indicator of cardiovascular diseases. Digit Health 9:20552076231187247, 2023
45. Krogsboll LT, Jorgensen KJ, Gotzsche PC: General health checks in adults for reducing morbidity and mortality from disease. Cochrane Database Syst Rev 1:CD009009, 2019
46. Cho JH, Han KD, Jung HY, Bond A: National health screening may reduce cardiovascular morbidity and mortality among the elderly. Public Health 187:172-176, 2020