

# Multimodal Deep Learning for Integrating Chest Radiographs and Clinical Parameters: A Case for Transformers

Firas Khader, MSc • Gustav Müller-Franzes, MSc • Tianci Wang, BSc • Tianyu Han, MSc • Soroosh Tayebi Arasteh, MSc • Christoph Haarbuerger, PhD • Johannes Stegmaier, PhD • Keno Bressem, MD • Christiane Kuhl, MD • Sven Nebelung, MD • Jakob Nikolas Kather, MD\* • Daniel Truhn, MD, PhD\*

From the Department of Diagnostic and Interventional Radiology (F.K., G.M.F., T.W., S.T.A., C.K., S.N., D.T.) and Department of Medicine III (J.N.K.), University Hospital Aachen, Pauwelsstraße 30, 52074 Aachen, Germany; Physics of Molecular Imaging Systems, Institute of Experimental Molecular Imaging (T.H.), and Institute of Imaging and Computer Vision (J.S.), RWTH Aachen University, Aachen, Germany; Ocumed, Munich, Germany (C.H.); Department of Diagnostic and Interventional Radiology, Charité–Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany (K.B.); Else Kroener Fresenius Center for Digital Health, Medical Faculty Carl Gustav Carus, Technical University Dresden, Dresden, Germany (J.N.K.); Division of Pathology and Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK (J.N.K.); and Department of Medical Oncology, National Center for Tumor Diseases, University Hospital Heidelberg, Heidelberg, Germany (J.N.K.). Received April 7, 2023; revision requested June 9; revision received August 18; accepted August 28. **Address correspondence to F.K.** (email: [fkhader@ukaachen.de](mailto:fkhader@ukaachen.de)).

J.N.K. is supported by the German Federal Ministry of Health (DEEP LIVER, ZMVI1-2520DAT111) and Max-Eder-Programme of the German Cancer Aid (70113864). S.N. is supported by the Deutsche Forschungsgemeinschaft (DFG) (NE 2136/3-1).

\*J.N.K. and D.T. are co-senior authors.

Conflicts of interest are listed at the end of this article.

See also the editorial by Kitamura and Topol in this issue.

Radiology 2023; 309(1):e230806 • <https://doi.org/10.1148/radiol.230806> • Content codes: **CH** **AI**

**Background:** Clinicians consider both imaging and nonimaging data when diagnosing diseases; however, current machine learning approaches primarily consider data from a single modality.

**Purpose:** To develop a neural network architecture capable of integrating multimodal patient data and compare its performance to models incorporating a single modality for diagnosing up to 25 pathologic conditions.

**Materials and Methods:** In this retrospective study, imaging and nonimaging patient data were extracted from the Medical Information Mart for Intensive Care (MIMIC) database and an internal database comprised of chest radiographs and clinical parameters in patients in the intensive care unit (ICU) (January 2008 to December 2020). The MIMIC and internal data sets were each split into training ( $n = 33\,893$ ,  $n = 28\,809$ ), validation ( $n = 740$ ,  $n = 7203$ ), and test ( $n = 1909$ ,  $n = 9004$ ) sets. A novel transformer-based neural network architecture was trained to diagnose up to 25 conditions using nonimaging data alone, imaging data alone, or multimodal data. Diagnostic performance was assessed using area under the receiver operating characteristic curve (AUC) analysis.

**Results:** The MIMIC and internal data sets included 36 542 patients (mean age, 63 years  $\pm$  17 [SD]; 20 567 male patients) and 45 016 patients (mean age, 66 years  $\pm$  16; 27 577 male patients), respectively. The multimodal model showed improved diagnostic performance for all pathologic conditions. For the MIMIC data set, the mean AUC was 0.77 (95% CI: 0.77, 0.78) when both chest radiographs and clinical parameters were used, compared with 0.70 (95% CI: 0.69, 0.71;  $P < .001$ ) for only chest radiographs and 0.72 (95% CI: 0.72, 0.73;  $P < .001$ ) for only clinical parameters. These findings were confirmed on the internal data set.

**Conclusion:** A model trained on imaging and nonimaging data outperformed models trained on only one type of data for diagnosing multiple diseases in patients in an ICU setting.

© RSNA, 2023

Supplemental material is available for this article.

In medicine, the diagnosis of a disease is based on data from multiple sources. A clinician will base decisions on radiologic images, clinical data, patient history, laboratory findings, and information from many additional modalities. The human mind is capable of condensing all these inputs into a rational decision. Deep learning has long been proposed to have the ability to assist medical doctors in certain tasks and has already demonstrated equal or better performance than human experts (1). However, there is one crucial impediment that limits the general applicability of such models—these models are almost exclusively tailored to solve tasks with one type of data at a time, be it the diagnosis of pathologies on radiologic images (2,3) or the detection of genetic alterations on histopathologic images (4).

Building on this realization, models that are capable of combining both imaging and nonimaging data as inputs are needed to truly support physician decision-making (5). Unfortunately, prevailing deep learning architectures of the past are not suited to deal with large amounts of imaging and nonimaging data; convolutional neural networks (CNNs) make use of intrinsic biases that build on image properties, such as correlations between neighboring pixels; and the integration of nonimaging information is not straightforward (6).

Originally introduced for natural language tasks, transformer-based neural network architectures have recently been shown to be competitive with CNNs for image processing, while simultaneously being ideally suited

## Abbreviations

AUC = area under the receiver operating characteristic curve, CNN = convolutional neural network, ICU = intensive care unit, MIMIC = Medical Information Mart for Intensive Care

## Summary

A transformer-based artificial intelligence architecture was developed to integrate multimodal patient data and demonstrated improved diagnostic performance on two data sets of chest radiographic and clinical parametric data.

## Key Results

- A transformer-based model trained to diagnose up to 25 diseases using multimodal data from two retrospectively acquired data sets (training sets;  $n = 33\,893$ ,  $n = 28\,809$ ) comprising chest radiographs and clinical parameters showed improved diagnostic performance.
- For the publicly available Medical Information Mart for Intensive Care (MIMIC) data set, the mean area under the receiver operating characteristic curve was 0.77 when chest radiographs and clinical parameters were used, compared with 0.70 ( $P < .001$ ) when only chest radiographs and 0.72 ( $P < .001$ ) when only clinical parameters were used.
- The multimodal model provided a flexible neural network whose outputs are explainable and well aligned with radiologic image perception.

to combine imaging and nonimaging data (7). This largely input-agnostic property is enabled by the use of an attention mechanism, which assigns importance scores to different parts of the input data, regardless of whether these data are of an imaging or nonimaging nature. Moreover, visualization of these importance scores offers valuable insights into the decision-making process of the transformer model. Thus, their application in medicine is the next logical step (8,9).

However, transformers have one notable shortcoming; that is, their computational load scales quadratically with the number of inputs. Without remedy, this will limit progress in medical research. To address this, the aim of this study was to develop a transformer model specifically tailored to the medical context, whereby imaging data and a potentially large volume of nonimaging data specific to each patient should be efficiently processed in an explainable way. An additional aim was to assess the diagnostic capacities of the model using multimodal inputs from a public data set and an independent internal data set of patients in an intensive care unit (ICU) setting. The hypothesis was that the diagnostic performance of the transformer model would be superior when trained on imaging and nonimaging data (multimodal) rather than imaging or nonimaging data alone (unimodal).

## Materials and Methods

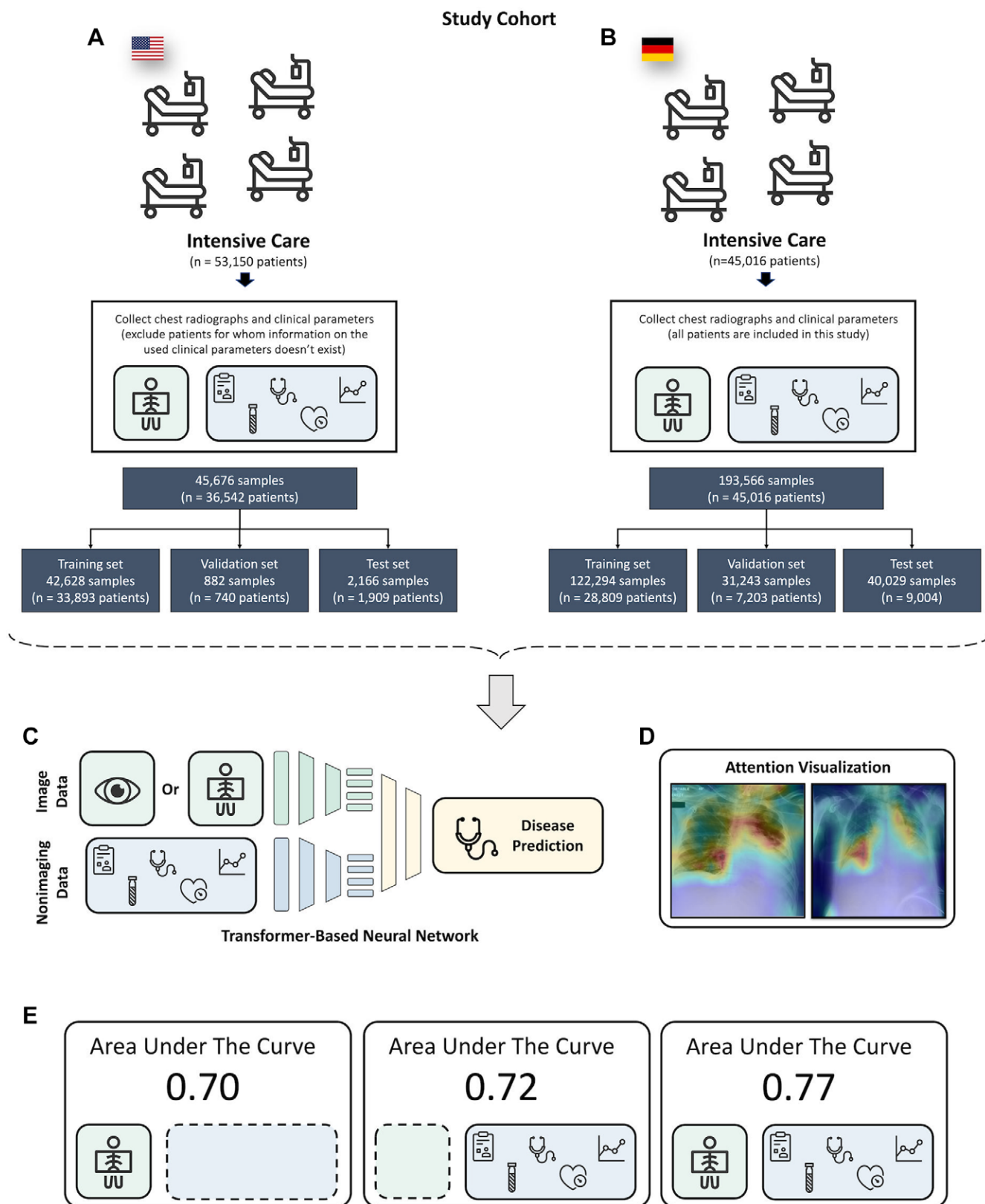
### Ethics Statement

All experiments were conducted retrospectively, and local ethical committee approval was granted (EK 028/19) and informed consent was waived. For the external Medical Information Mart for Intensive Care (MIMIC) data set, patient-specific identifiers were removed in compliance with the Health Insurance Portability and Accountability Act.

## Study Patients and Data Sets

To enable replication of the results of this retrospective study and to foster research in this direction, this model was evaluated primarily on data from the publicly available MIMIC database (10,11). This database is comprised of retrospectively collected imaging and nonimaging data in 53 150 patients admitted to an ICU at the Beth Israel Deaconess Medical Center from January 2008 to December 2019. Following the work of Hayat et al (12), imaging and nonimaging information was extracted from the MIMIC-IV (10) and MIMIC-CXR-JPG (11) databases for which either information on approximately 15 clinical parameters alone or information on these parameters combined with imaging information in the form of chest radiographs were available. The clinical parameters included systolic, diastolic, and mean blood pressure; respiratory rate; motor, verbal, and eye-opening commands per the Glasgow Coma Scale; oxygen inspiration; heart rate; body temperature, weight, and height; acidic value of blood serum; blood glucose level; and blood oxygen level. The chest radiograph was paired with the clinical parameters and laboratory parameters from the same ICU stay. This resulted in a subset of 45 676 samples in 36 542 patients (Fig 1, Table 1). Data from all 36 542 patients have been previously reported (12). The prior article dealt with the development of a CNN and recurrent neural network–based architecture to combine multimodal data, whereas this study deals with the development of a transformer-based architecture. Moreover, the approach by Hayat et al (12) was followed, and the available International Classification of Diseases (ICD)–9 and ICD-10 codes for 25 superordinate disease categories (see Table 2 for a comprehensive list) were grouped based on the Clinical Classifications Software (Agency for Healthcare Research and Quality) (13), which is a commonly used clinical classification framework.

Additionally, the model was evaluated on an in-house data set from 45 016 patients who were admitted to the ICU of a tertiary academic medical center (University Hospital Aachen, Aachen, Germany) from January 2009 to December 2020 (14). Data from all patients were used in the current study. In addition to imaging data (ie, chest radiographs), this data set also contained time-series data of laboratory tests, including C-reactive protein (CRP) levels, leukocyte count, procalcitonin (PCT) levels, and brain natriuretic peptide (BNP) levels. These values were included if available within a 20-day window before the acquisition of the chest radiograph, and, in total, data for 34 595 (CRP), 40 267 (leukocyte count), 23 084 (PCT), and 9771 (BNP) patients were available. Images were paired with all laboratory data that preceded the images by up to 20 days. The imaging data in this data set were generated during routine clinical reporting. In total, 98 modality-versed radiologists used an itemized template for structured reporting on the presence and severity of pleural effusion (left and right), atelectasis (left and right), pulmonary opacities (left and right), pulmonary congestion, and cardiomegaly. Images for which no disease was found were assigned a binarized target value of 0, while the remaining labels indicated the presence of a disease and thus were assigned the value of one. Data from 45 016 patients in this data set have been previously reported (14).



**Figure 1:** Diagram shows an overview of the study. **(A–E)** Imaging and nonimaging information were extracted from the publicly available Medical Information Mart for Intensive Care data set **(A)** and an internal data set of chest radiographic and accompanying clinical parametric data **(B)**. The data sets were split into training, validation, and test sets, and a transformer-based neural network architecture **(C)** was trained to predict the diagnosis of up to 25 different pathologic conditions. First, the attention mechanism in the transformer architecture **(D)** was leveraged to provide insight into the decision-making process of the neural network, and it was shown that the predictive performance of the neural network **(E)** increased for all three data sets when both imaging and nonimaging inputs (area under the receiver operating characteristic curve [AUC], 0.77) were provided compared with either imaging (AUC, 0.70) or nonimaging (AUC, 0.72) inputs alone.

**Table 1: Characteristics of the Study Data Sets**

| Data Set     | Data Source                      | Sample Size*     | Age (y) <sup>†</sup> | Sex    |        | Nonimaging Parameters                                                                                              | Imaging Data      |
|--------------|----------------------------------|------------------|----------------------|--------|--------|--------------------------------------------------------------------------------------------------------------------|-------------------|
|              |                                  |                  |                      | Male   | Female |                                                                                                                    |                   |
| MIMIC        | MIMIC-IV and MIMIC-CXR databases | 45 676 (36 542)  | 63 ± 17              | 20 567 | 15 975 | 15 clinical parameters (eg, blood pressure, heart rate, Glasgow Coma Scale components)                             | Chest radiographs |
| In-house ICU | Internal data set                | 193 566 (45 016) | 66 ± 16              | 27 577 | 17 439 | Time-series laboratory values (eg, C-reactive protein, leukocyte, procalcitonin, brain natriuretic peptide levels) | Chest radiographs |

Note.—ICU = intensive care unit, MIMIC = Medical Information Mart for Intensive Care.

\* Numbers in parentheses are numbers of patients.

<sup>†</sup> Data are means ± SDs.

The prior article dealt with the development of a CNN trained to provide clinical support for nonradiologist physicians using only imaging data, whereas in this study, additional laboratory values were used to train a transformer-based neural network on multimodal data.

### Imaging Protocols

The internal data set consisted of chest radiographs obtained using 18 mobile radiography machines (Mobilett Mira; Siemens Healthineers). These imaging systems used conventional film systems until 2016, after which they transitioned to digital flat-panel detectors. All radiographs were acquired using automatic exposure control and exclusively in the anteroposterior projection. Correspondingly, images from the external MIMIC data set were acquired in the anteroposterior projection.

### Data Preprocessing and Neural Network Design

For a fair evaluation of the models, following the approach detailed by Hayat et al (12), the MIMIC data set was randomly divided into a training set of 42 628 samples (33 893 patients), a validation set of 882 samples (740 patients) to select the most optimized model, and a holdout test set of 2166 samples (1909 patients) to evaluate the model on unseen data. Similarly, the internal data set comprising 193 566 samples (45 016 patients) was randomly split into a training set of 122 294 samples (28 809 patients), validation set of 31 243 samples (7203 patients), and holdout test set of 40 029 samples (9004 patients) (Fig 1). Special care was taken to ensure that each patient appears only in a single set. Images were normalized to the range from 0 to 255, contrast-enhanced using a histogram equalization, resized to 384 × 384 pixels, and *z*-normalized to match the ImageNet (15) data set statistics, enabling potential use of pretrained models.

The neural network architecture (Fig 2) is based on the transformer model as follows (16). Images are tokenized and fed through a Vision Transformer (7) backbone to extract relevant features from the imaging data. Building on the Perceiver model (17), nonimaging data are incorporated through the use of the cross-attention mechanism (16), enabling scalability and flexibility in handling variable input sizes. A final transformer

encoder block is then used for cross-modality information fusion, and a multilayer perceptron is used to generate the outputs of the multilabel classification. See Appendix S1, Tables S2 and S3, and Figure 2B and 2C for further details.

### Data Availability

The MIMIC data set, including imaging and nonimaging data, is publicly available via PhysioNet (<https://physionet.org/content/mimiciv/1.0/>) (18). The internal ICU data set is private due to data protection issues but will be shared by the authors upon the submission of a research proposal and given the consent of the data protection officer and the ethical board.

### Code Availability

The code used to train the model described herein is publicly available on GitHub (<https://github.com/FirasGit/lsmr>).

### Statistical Analysis

Statistical analyses were conducted by F.K. and D.T. using Python (version 3.8; <https://www.python.org/>), along with the NumPy and SciPy libraries. The statistical spread was determined using bootstrapping with 1000 redraws, with replacement from the test set for each measure. The Youden criterion was used to determine a threshold for sensitivity, specificity, and positive predictive value calculations, which involves finding the threshold that maximizes the sum of sensitivity and specificity. To compute *P* values for the individual diseases, the DeLong test (19) was used, which was specifically developed for testing area under the receiver operating characteristic curve (AUC) scores. To estimate *P* values of the mean AUC scores, we computed the pairwise differences between the bootstrapped AUC scores for each model with identical redraws and calculated the fraction of differences with values less than 0. A particular significance level was not chosen to avoid dichotomization of the results as either significant or not significant (20) and to obviate the need to compensate for multiple hypothesis testing. Data are presented as means ± SDs and AUCs with 95% CIs. The debate on minimum sample sizes is ongoing, and at least 200 patients are considered necessary for classification tasks (21). In this study, as many patients as possible were

included (ie, 36 542 and 45 016 patients), thereby obviating the need to perform sample size estimations.

## Results

### Patient Characteristics

In this study, two data sets (Table 1, Fig 1) were used to evaluate the proposed neural network architecture. The MIMIC data set contains data from 53 150 patients, of which 16 608 patients were excluded as they did not have measurements for any of the 15 clinical parameters used in this study; thus, data

from 36 542 patients (mean age, 63 years  $\pm$  17 [SD]; 20 567 male patients) were used in this study. The internal data set contains data from 45 016 patients (mean age, 66 years  $\pm$  16; 27 577 male patients).

### Performance of Multimodal Transformer for Diagnosis of Multiple Diseases

The model was trained and evaluated on the publicly available data of 36 542 patients who received treatment in an ICU (10,11). Chest radiographs and accompanying clinical data were employed as inputs to the model, and the model was

**Table 2: Predictive Performance of the Model When Trained on the Publicly Available MIMIC Data Set**

| Pathologic Condition                                       | Chest Radiographs Only AUC | CPs Only AUC         | CPs Plus Chest Radiographs AUC | P Value* | P Value† | P Value‡ |
|------------------------------------------------------------|----------------------------|----------------------|--------------------------------|----------|----------|----------|
| Acute and unspecified kidney failure                       | 0.70<br>(0.68, 0.73)       | 0.74<br>(0.72, 77)   | 0.77<br>(0.75, 0.79)           | <.001    | <.001    | .004     |
| Acute cerebrovascular disease                              | 0.70<br>(0.66, 0.75)       | 0.89<br>(0.86, 0.91) | 0.91<br>(0.88, 0.93)           | .006     | <.001    | <.001    |
| Acute myocardial infarction                                | 0.68<br>(0.65, 0.72)       | 0.71<br>(0.68, 0.75) | 0.76<br>(0.73, 0.80)           | <.001    | <.001    | .24      |
| Cardiac dysrhythmias                                       | 0.69<br>(0.66, 0.71)       | 0.64<br>(0.61, 0.66) | 0.71<br>(0.70, 0.74)           | <.001    | <.001    | .001     |
| Chronic kidney disease                                     | 0.71<br>(0.69, 0.74)       | 0.70<br>(0.68, 0.73) | 0.78<br>(0.76, 0.80)           | <.001    | <.001    | .54      |
| Chronic obstructive pulmonary disease and bronchiectasis   | 0.75<br>(0.72, 0.78)       | 0.67<br>(0.63, 0.69) | 0.76<br>(0.76, 0.80)           | <.001    | .65      | <.001    |
| Complications of surgical procedures or medical care       | 0.65<br>(0.62, 0.68)       | 0.70<br>(0.67, 0.72) | 0.72<br>(0.69, 0.75)           | .02      | <.001    | .003     |
| Conduction disorders                                       | 0.83<br>(0.80, 0.87)       | 0.67<br>(0.63, 0.70) | 0.85<br>(0.82, 0.88)           | <.001    | .09      | <.001    |
| Congestive heart failure (nonhypertensive)                 | 0.79<br>(0.77, 0.81)       | 0.71<br>(0.69, 0.73) | 0.82<br>(0.81, 0.84)           | <.001    | <.001    | <.001    |
| Coronary atherosclerosis and other heart disease           | 0.76<br>(0.74, 0.78)       | 0.73<br>(0.71, 0.75) | 0.79<br>(0.77, 0.81)           | <.001    | <.001    | .03      |
| Diabetes mellitus with complications                       | 0.64<br>(0.61, 0.67)       | 0.87<br>(0.84, 0.89) | 0.87<br>(0.84, 0.89)           | .93      | <.001    | <.001    |
| Diabetes mellitus without complications                    | 0.63<br>(0.61, 0.66)       | 0.73<br>(0.71, 0.76) | 0.73<br>(0.71, 0.76)           | .74      | <.001    | <.001    |
| Disorders of lipid metabolism                              | 0.68<br>(0.66, 0.70)       | 0.67<br>(0.64, 0.69) | 0.72<br>(0.67, 0.71)           | <.001    | <.001    | .34      |
| Essential hypertension                                     | 0.64<br>(0.61, 0.66)       | 0.65<br>(0.62, 0.67) | 0.69<br>(0.67, 0.71)           | <.001    | <.001    | .58      |
| Fluid and electrolyte disorders                            | 0.66<br>(0.63, 0.68)       | 0.71<br>(0.69, 0.73) | 0.73<br>(0.71, 0.75)           | .04      | <.001    | <.001    |
| Gastrointestinal hemorrhage                                | 0.63<br>(0.59, 0.68)       | 0.74<br>(0.70, 0.78) | 0.75<br>(0.72, 0.79)           | .42      | <.001    | <.001    |
| Hypertension with complications and secondary hypertension | 0.71<br>(0.69, 0.74)       | 0.68<br>(0.65, 0.71) | 0.77<br>(0.74, 0.79)           | <.001    | <.001    | .07      |
| Other liver diseases                                       | 0.70<br>(0.67, 0.73)       | 0.68<br>(0.65, 0.71) | 0.74<br>(0.71, 0.77)           | <.001    | <.001    | .29      |
| Other lower respiratory disease                            | 0.61<br>(0.57, 0.65)       | 0.59<br>(0.54, 0.62) | 0.61<br>(0.58, 0.65)           | .07      | .81      | .22      |
| Other upper respiratory disease                            | 0.60<br>(0.54, 0.65)       | 0.72<br>(0.68, 0.77) | 0.72<br>(0.66, 0.77)           | .70      | <.001    | <.001    |

**Table 2 (continues)**

**Table 2 (continued): Predictive Performance of the Model When Trained on the Publicly Available MIMIC Data Set**

| Pathologic Condition                                                           | Chest Radiographs Only AUC | CPs Only AUC         | CPs Plus Chest Radiographs AUC | <i>P</i> Value* | <i>P</i> Value† | <i>P</i> Value‡ |
|--------------------------------------------------------------------------------|----------------------------|----------------------|--------------------------------|-----------------|-----------------|-----------------|
| Pleurisy, pneumothorax, pulmonary collapse                                     | 0.76<br>(0.73, 0.79)       | 0.63<br>(0.59, 0.66) | 0.76<br>(0.72, 0.79)           | <.001           | .80             | <.001           |
| Pneumonia (except that caused by tuberculosis or sexually transmitted disease) | 0.75<br>(0.72, 0.77)       | 0.77<br>(0.75, 0.79) | 0.81<br>(0.79, 0.83)           | <.001           | <.001           | .07             |
| Respiratory failure, insufficiency, arrest (adult)                             | 0.76<br>(0.74, 0.78)       | 0.82<br>(0.80, 0.84) | 0.86<br>(0.84, 0.87)           | <.001           | <.001           | <.001           |
| Septicemia (except in labor)                                                   | 0.73<br>(0.70, 0.75)       | 0.79<br>(0.77, 0.82) | 0.81<br>(0.78, 0.83)           | .15             | <.001           | <.001           |
| Shock                                                                          | 0.74<br>(0.71, 0.77)       | 0.85<br>(0.83, 0.87) | 0.85<br>(0.83, 0.87)           | .62             | <.001           | <.001           |
| Mean overall                                                                   | 0.70<br>(0.69, 0.71)       | 0.72<br>(0.72, 0.73) | 0.77<br>(0.77, 0.78)           | <.001           | <.001           | <.001           |

Note.—Data in parentheses are 95% CIs. Three different neural networks that relied on nonimaging information (CPs) alone, imaging information (chest radiographs) alone, or both (CPs plus chest radiographs) were trained to predict the presence of 25 pathologic conditions. AUC = area under the receiver operating characteristic curve, CP = clinical parameter, MIMIC = Medical Information Mart for Intensive Care.

\* CPs plus chest radiographs versus CPs only.

† CPs plus chest radiographs versus chest radiographs only.

‡ CPs only versus chest radiographs only.

allowed to predict a comprehensive set of 25 pathologic conditions. Consistently, the AUC was higher when both imaging and nonimaging data were employed than when either imaging or nonimaging data alone were used (Table 2, Fig S1). The mean AUC was 0.77 (95% CI: 0.77, 0.78) when both chest radiographs and clinical parameters were used, compared with 0.70 (95% CI: 0.69, 0.71;  $P < .001$ ) when only chest radiographs and 0.72 (95% CI: 0.72, 0.73;  $P < .001$ ) when only clinical parameters were used. Similar trends were seen for the sensitivity (clinical parameters plus chest radiographs: 70% [95% CI: 69, 71]; clinical parameters: 69% [95% CI: 68, 70]; chest radiographs: 66% [95% CI: 65, 67]), specificity (clinical parameters plus chest radiographs: 72% [95% CI: 72, 73]; clinical parameters: 65% [95% CI: 64, 65]; chest radiographs: 65% [95% CI: 65, 66]), and positive predictive value (clinical parameters plus chest radiographs: 40% [95% CI: 40, 41]; clinical parameters: 35% [95% CI: 34, 35]; chest radiographs: 34% [95% CI: 34, 35]). More importantly, the performance of the multimodal transformer is comparable with other state-of-the-art approaches (eg, MedFuse [12], which demonstrated an AUC of 0.770 [95% CI: 0.745, 0.795] for the multimodal case), while requiring no extensive hyperparameter tuning. See Table 3 for a detailed comparison with the previous performance achieved with CNNs. The model was further evaluated on an additional task using an independent data set, which was comprehensive radiologic diagnosis of chest radiographs based on imaging data and accompanying laboratory data (Table S1, Fig S2) (14). The mean AUC was 0.84 (95% CI: 0.83, 0.84) when chest radiographs and clinical parameters were used, compared with 0.83 (95% CI: 0.82, 0.83;  $P < .001$ ) when only chest radiographs

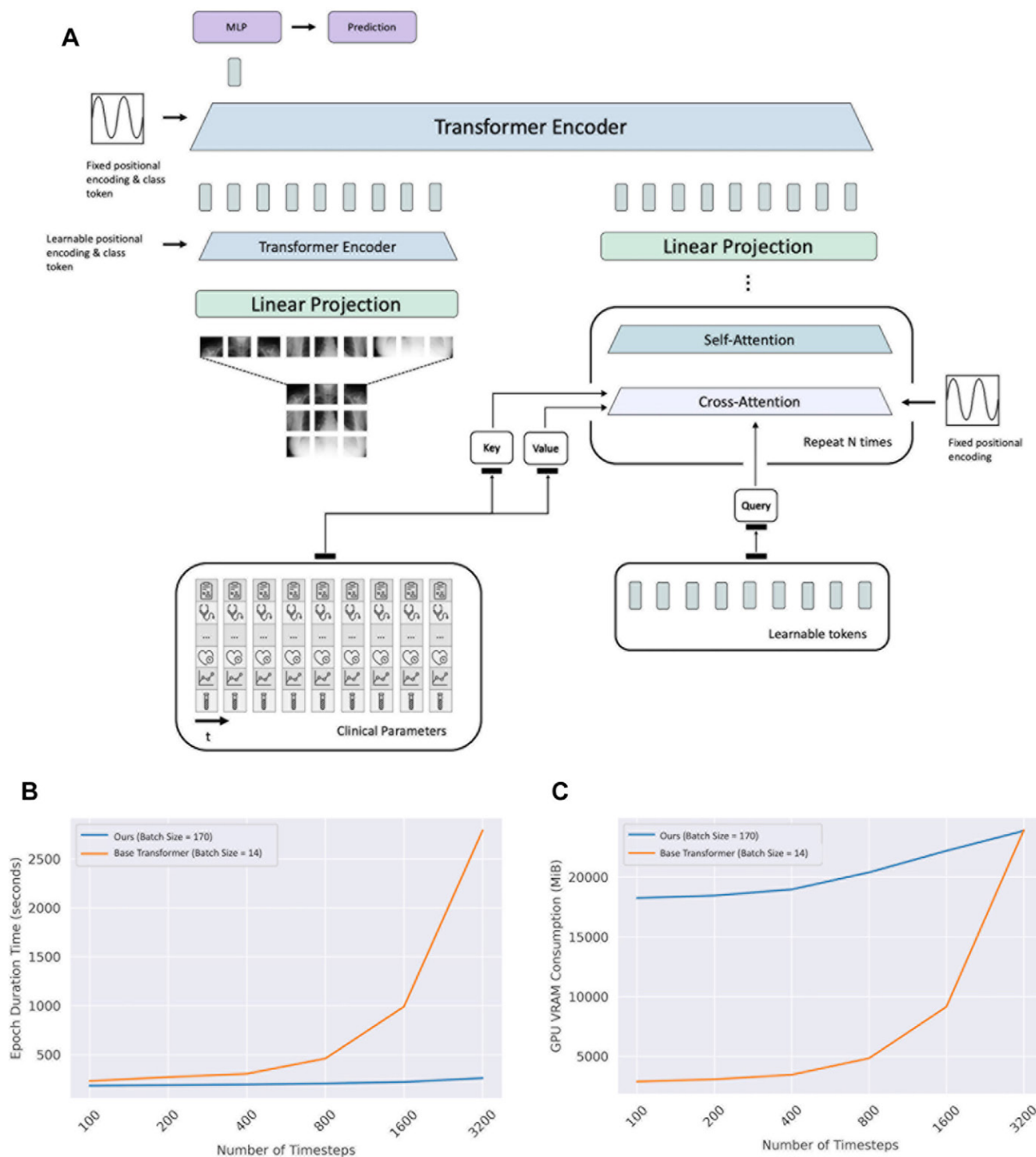
and 0.67 (95% CI: 0.66, 0.67;  $P < .001$ ) when only clinical parameters were used (Fig 3, Table S1). Again, similar trends were also seen for the sensitivity (clinical parameters plus chest radiographs: 77% [95% CI: 77, 77]; clinical parameters: 73% [95% CI: 73, 73]; chest radiographs: 76% [95% CI: 76, 76]), specificity (clinical parameters plus chest radiographs: 74% [95% CI: 73, 73]; clinical parameters: 52% [95% CI: 52, 52]; chest radiographs: 73% [95% CI: 73, 73]), and positive predictive value (clinical parameters plus chest radiographs: 71% [95% CI: 71, 71]; clinical parameters: 56% [95% CI: 56, 56]; chest radiographs: 70% [95% CI: 69, 70]).

### Multimodal Transformer Performance When Data Were Missing

The proposed transformer architecture worked when data were missing and resembled human reasoning in the sense that its performance continuously declined when increasing amounts of clinically relevant data were missing. Patient data from the test set were fed to the trained transformer with some of the input parameters randomly omitted. The performance in terms of the mean AUC (15 parameters at 0.77 [95% CI: 0.76, 0.77] vs one parameter at 0.73 [95% CI: 0.73, 0.74]) continuously declined when increasing amounts of data were omitted, in agreement with expectations. Similar trends were observed for sensitivity, specificity, and positive predictive value (Fig 4A).

### Multimodal Transformer Agreement with Clinical Reasoning

When measuring the percentage decrease in mutual information between the model's prediction and the ground truth to



**Figure 2:** (A) Schematic shows the model architecture, whereby images are first split into nonoverlapping patches and subsequently fed through a transformer encoder. To account for scalability with regard to the number of nonimaging parameters, a fixed set of 64 learnable tokens serves as the neural network working memory, and cross-attention is employed to feed the clinical information to this working memory. This keeps the network scalable with respect to the number of input tokens (ie, clinical parameters). The output tokens of both modality-specific neural networks are then merged in a final transformer encoder, such that information from both modalities is fused. (B) Line graph shows the epoch duration time for models trained on the same graphics processing unit (Quadro RTX 6000; NVIDIA). To ensure a comparable usage of the graphics processing unit video random-access memory, different batch sizes were employed, allowing for a batch size of 170 for the proposed model (blue) and a batch size of 14 for the base transformer approach (orange). Compared with the conventional setting, in which the imaging and nonimaging (time-series) data are fed directly into a common transformer encoder block for information fusion, the model used in the current study results in shorter training times. (C) Line graph shows graphics processing unit (GPU) video random-access memory (VRAM) consumption as a function of the number of input parameters. The findings indicate that the employed approach (blue) scales much more efficiently than the base transformer approach (orange) for an increasing number of input parameters and, therefore, allows for larger batch sizes during training. Here, the batch size used for each model was based on the maximal possible batch size (in terms of video random-access memory consumption of the graphics processing unit) when training the model with 3200 timesteps. MiB = mebibyte, MLP = multilayer perceptron.

**Table 3: Comparison of the Model to Other State-of-the-Art Architectures Evaluated on the Publicly Available MIMIC Data Set**

| Method                                     | Mean AUC             | Comments                                                                                                                                                                                                                                                                                                                                                   |
|--------------------------------------------|----------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Early (12)                                 | 0.739 (0.712, 0.766) | Two feature extractors are used for training on the nonimaging and imaging data separately. The features are then extracted and concatenated, such that a classification layer can be used to merge the inputs.                                                                                                                                            |
| Joint (12)                                 | 0.754 (0.727, 0.780) | Imaging and nonimaging data are trained end to end in a common neural network architecture that consists of two separate feature extraction blocks that are later fused by a final classification head.                                                                                                                                                    |
| MMTM (12,31)                               | 0.734 (0.707, 0.761) | To merge the features of the imaging and nonimaging feature extraction backbone, an MMTM is used.                                                                                                                                                                                                                                                          |
| DAFT (12,32)                               | 0.737 (0.710, 0.764) | Features from the imaging and nonimaging feature extraction backbone are merged by a DAFT block that scales and shifts the feature maps for information fusion.                                                                                                                                                                                            |
| Unified (12,33)                            | 0.765 (0.742, 0.794) | To merge the features from paired imaging and nonimaging inputs, two feature extractors are employed that, in every training iteration, are first trained separately on unpaired samples and then, in a second step, used to extract the features of the paired inputs. The features are concatenated and merged with a final classification head.         |
| MedFuse (12)                               | 0.756 (0.729, 0.782) | Similar to previous models, two separate feature extractors are trained for the imaging and nonimaging inputs. However, unlike the classification heads used before, LSTM (34) layers are used to merge the features of both extractors in a sequential manner.                                                                                            |
| MedFuse OPTIMAL (12)                       | 0.770 (0.745, 0.795) | This model is identical to the previous MedFuse model, except that instead of using all unpaired samples, the authors selectively used only a fraction of the unpaired data set that led to the best performance.                                                                                                                                          |
| Multimodal model used in the current study | 0.772 (0.766, 0.777) | This model uses the transformer architecture to extract features from imaging (chest radiographs) and nonimaging (clinical parameters) information. To keep the model scalable to a large number of nonimaging input tokens, the cross-attention module is leveraged and the features from both modalities are fused in a final transformer encoder block. |

Note.—Data in parentheses are 95% CIs. AUC analysis was used for the comparison, and the mean for all available pathologic conditions and 95% CIs were computed. AUC = area under the receiver operating characteristic curve, DAFT = Dynamic Affine Feature Map Transform, LSTM = long short-term memory, MIMIC = Medical Information Mart for Intensive Care, MMTM = Multimodal Transfer Module.

uncover relationships between the available data and the diagnostic performance for specific clinical parameters (see Appendix S1 for details on this methodology), the results of this analysis were found to largely agree with clinical reasoning. Clinical parameters that are relevant to a specific patient state, such as blood pressure for shock or glucose concentration for diabetes, led to the greatest information loss when omitted (Fig 4B). A comprehensive overview of all clinical conditions is provided in Appendix S1 and Figure S3.

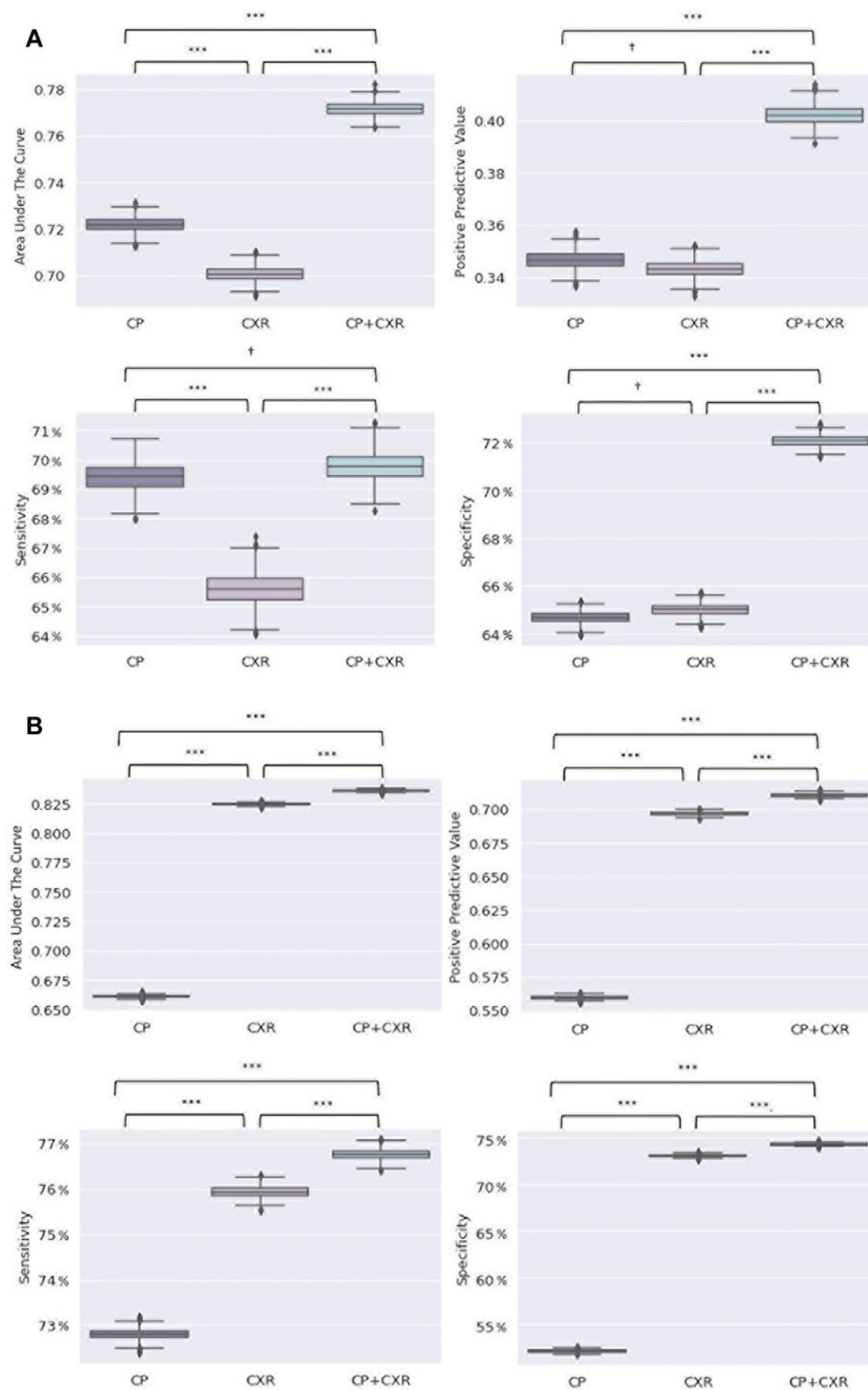
### Multimodal Transformer Focus on Pathologic Image Regions

By making use of the inherent attention mechanism, maps can be generated that showcase where the transformer model focuses its attention, including which subregions of the radiographs get the highest weighting factors for the final diagnosis (see Appendix S1 for more details on the method used). Figure 5 illustrates these attention maps by using three representative examples from each data set. Consistently, the attention maps exhibit their highest values in image regions that are indicative of the pathologies.

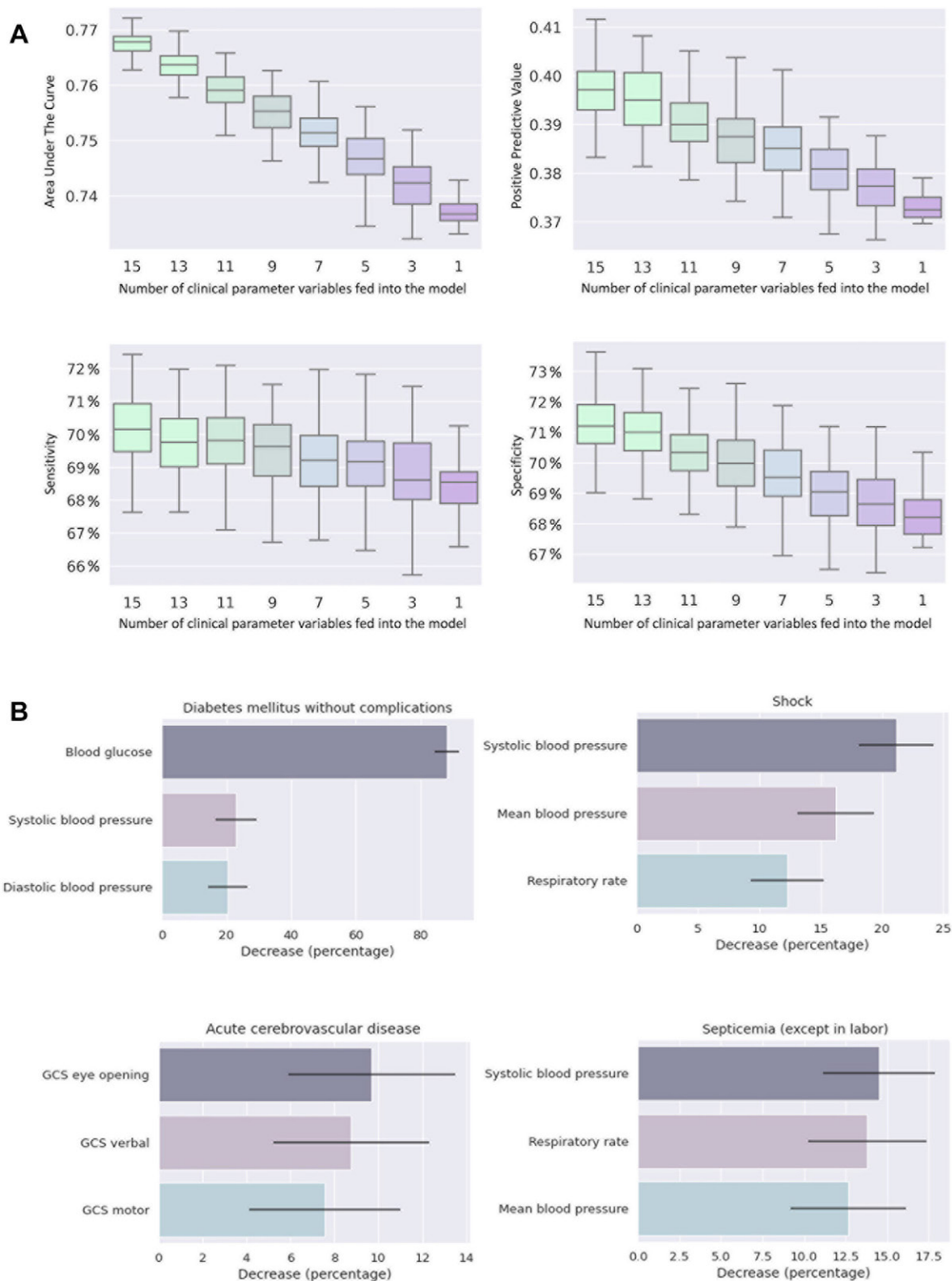
## Discussion

In recent years, there has been a surge of applications of deep learning models to solve medical problems (14,22–24); however, these models typically use data from only one modality (eg, imaging data).

Transformer models have been proposed as an ideal candidate for assessing multimodal data, as they were first developed on nonimaging data (25,26) and have now proven to be competitive with CNNs developed on imaging data (27). In our study, we developed a scalable fully transformer-based approach for multimodal prediction based on medical imaging and nonimaging data. Our model demonstrates improved performance when presented with multimodal data, can handle missing data, and allows insight into the network's decision-making process. Most importantly, building on the Perceiver architecture (17), our model is scalable and can be applied to data sets in which both the number of patients and the data per patient are extensive. When trained jointly on data from chest radiographs and clinical parameters from the publicly available MIMIC database, the mean AUC was consistently higher (0.77 [95% CI: 0.77, 0.78])



**Figure 3: (A, B)** Box plots show the predictive performance of the trained neural networks in terms of the area under the receiver operating characteristic curve (AUC), positive predictive value, sensitivity, and specificity for the Medical Information Mart for Intensive Care data set **(A)** and an internal data set of chest radiographs and clinical parameters **(B)**. For both data sets, three models were trained on only the nonimaging information (CP), only the imaging information (CXR) as input, and the multimodal model exhibited a higher AUC than its unimodal counterparts. Box plots show the result of 1000 bootstrapping runs, with boxes indicating the IQR between the first and third quartiles and whiskers extending to  $\pm 1.5 \times$  IQR. The center line denotes the median. † =  $P > .05$ , \*\*\* =  $P < .001$ .

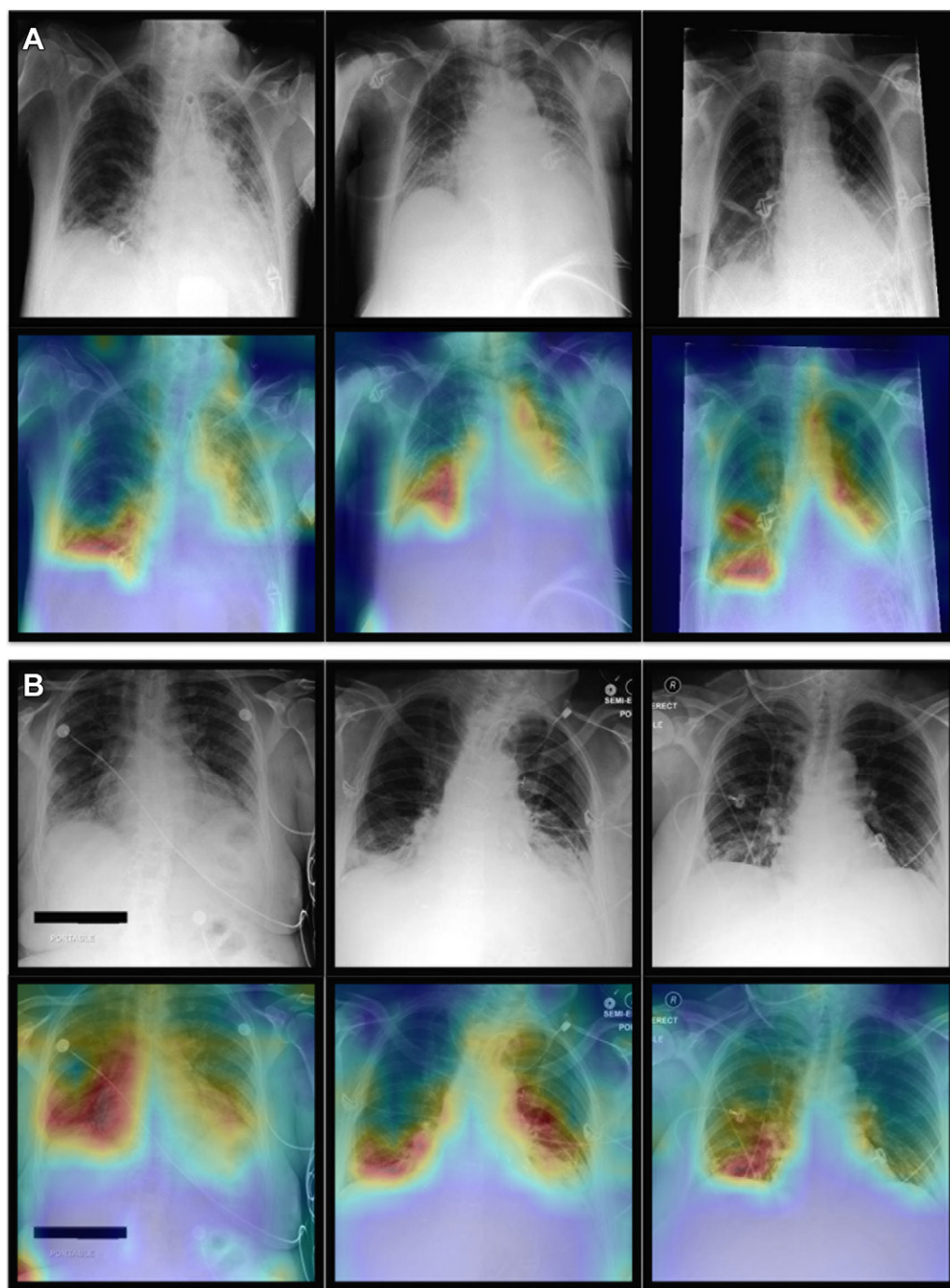


**Figure 4:** (A) Box plots show performance in terms of the area under the receiver operating characteristic curve, positive predictive value, sensitivity, and specificity of the neural network trained on the Medical Information Mart for Intensive Care data set when a number of clinical parameters (nonimaging information) were omitted. Performance continuously decreased with an increasing number of omitted clinical parameters. Boxes indicate the IQR between the first and third quartiles and whiskers extend to  $\pm 1.5 \times$  IQR, while the center line denotes the median. (B) Horizontal bar graphs show clinical parameters that most affected the performance of the neural network for diagnosis of diabetes (without complications), shock, acute cerebrovascular disease, and septicemia. To gain an understanding of the clinical parameters that most affected neural network performance for a specific pathologic condition, the percentage decrease in mutual information between the predicted distribution over all samples and their ground truth labels when a specific clinical parameter was left out during inference was determined. Error bars denote SDs. GCS = Glasgow Coma Scale.

compared with that of the models trained on either imaging (0.70 [95% CI: 0.69, 0.71],  $P < .001$ ) or nonimaging (0.72 [95% CI: 0.72, 0.73],  $P < .001$ ) data alone.

Previous research groups have invested considerable effort to process contextual (nonimaging) and imaging data. Huang et al (28) surveyed the literature and identified three principal fusion strategies, which are early, joint, and late fusion. Early fusion concatenates the multimodal features at the input level; joint fusion employs separate feature extractors for each modality, subsequently joining the learned feature representations; and late fusion aggregates predictions of separate models at the decision level. Our approach can best be described as joint fusion because it makes use of feature extractors before combining the modalities. As the backbone, we made use of the well-established Vision Transformer model. By design, exchangeable backbones are the centerpiece of our transformer-centered ensemble and, if needed, this backbone may be exchanged against more advanced transformers once more performant future models are available.

With this backbone, our model achieved results that are comparable with other state-of-the-art approaches (12) while simultaneously being scalable, providing insight into the decision-making process and being robust in the sense that the model can be applied when data are missing. These properties are indispensable for the application to clinical routine where missing data and long time series of data are very common (29).



**Figure 5:** Representative radiographs (top), acquired in anteroposterior projection in the supine position, and corresponding attention maps (bottom). **(A)** Images show main diagnostic findings of the internal data set in a 49-year-old male patient with congestion, pneumonic infiltrates, and effusion (left); a 64-year-old male patient with congestion, pneumonic infiltrates, and effusion (middle); and a 69-year-old female patient with effusion (right). **(B)** Images show main diagnostic findings of the Medical Information Mart for Intensive Care data set in a 79-year-old male patient with cardiomegaly and pneumonic infiltrates in the right lower lung (left); a 58-year-old female patient with bilateral atelectasis and effusion in the lower lungs (middle); and a 48-year-old female patient with pneumonic infiltrates in the lower right lung (right). Note that the attention maps consistently focus on the most relevant image regions (eg, pneumonic opacities are indicated by opaque image regions of the lung).

For both investigated data sets of radiographs with accompanying nonimaging data, we found a consistent increase in diagnostic performance when nonimaging clinical data were used along with imaging data. This is in line with other studies that have used deep learning models to combine different

data modalities, such as histopathology and CT (30), and it is expected that multimodal models that can combine a vast array of modalities will be dominating the future artificial intelligence landscape (29). However, diagnostic performance may not inevitably benefit from integrating both imaging and nonimaging data. For instance, diabetes is predominantly diagnosed without imaging, while pneumothorax diagnosis relies primarily on imaging.

This study had several limitations. First, the examples demonstrated herein make use of two-dimensional imaging data; however, a substantial amount of medical imaging data are three-dimensional, and whether the presented paradigms hold up with three-dimensional data should be demonstrated once such data sets become available. Second, we tested our neural network in the context of supervised learning, which requires the presence of labels for each patient and limited the range of data that can be employed to train our architecture. Third, the domain transferability of our model could not be tested due to the lack of suitable data sets that have concordant labels and concordant data available for training. Fourth, we did not perform a comparison to other model architectures regarding training times, as this was beyond the scope of the present study. Future studies comparing such models would require implementing these models head to head using identical training, validation, and test settings.

In conclusion, this study has demonstrated that a transformer model trained on large-scale imaging and nonimaging data outperformed models trained on unimodal data, although future studies should investigate other imaging scenarios to reliably confirm the generalizability. With the advent of transformer architectures and the growing interest in multimodal deep learning models, we expect large-scale data sets that include different modalities from radiography to MRI, anatomies from head to toe, and various conditions to become publicly available. This will constitute an ideal application and testing ground for the transformer models presented in this study.

**Acknowledgment:** We thank the MIMIC consortium for providing the data sets used in our study.

**Author contributions:** Guarantors of integrity of entire study, F.K., D.T.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, F.K., G.M.F., T.W., T.H., C.H., S.N., D.T.; clinical studies, F.K., C.K., S.N.; experimental studies, F.K., C.H., D.T.; statistical analysis, F.K., J.N.K., D.T.; and manuscript editing, F.K., S.T.A., J.S., K.B., C.K., S.N., J.N.K., D.T.

**Disclosures of conflicts of interest:** F.K. No relevant relationships. G.M.F. No relevant relationships. T.W. No relevant relationships. T.H. No relevant relationships. S.T.A. No relevant relationships. C.H. No relevant relationships. J.S. No relevant relationships. K.B. Grant from the European Union (101079894) and Wilhem-Sander Foundation; advisor for the LifeChamps project. C.K. No relevant relationships. S.N. No relevant relationships. J.N.K. Consultant for Owkin, DoMore Diagnostics, and Panakeia; lecture honoraria from Bayer, Eisai, MSD, BMS, Roche, and Fresenius; advisory board for MSD. D.T. No relevant relationships.

## References

- Shmatko A, Ghaffari Laleh N, Gerstung M, Kather JN. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nat Can* 2022;3(9):1026–1038.
- Truhn D, Schrading S, Hauburger C, Schneider H, Merhof D, Kuhl C. Radiomic versus Convolutional Neural Networks Analysis for Classification of Contrast-enhancing Lesions at Multiparametric Breast MRI. *Radiology* 2019;290(2):290–297.
- Han T, Kather JN, Pedersoli F, et al. Image Prediction of Disease Progression for Osteoarthritis by Style-Based Manifold Extrapolation. *Nat Mach Intell* 2022;4(11):1029–1039.
- Schrammen PL, Ghaffari Laleh N, Echle A, et al. Weakly supervised annotation-free cancer detection and prediction of genotype in routine histopathology. *J Pathol* 2022;256(1):50–60.
- Lipkova J, Chen RJ, Chen B, et al. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell* 2022;40(10):1095–1110.
- D'Ascoli S, Touvron H, Leavitt ML, Morcos AS, Biroli G, Sagun L. ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. In: Proceedings of the 38th International Conference on Machine Learning 2286–2296 (PMLR, 2021).
- Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv 2010.11929 [preprint]. <https://arxiv.org/abs/2010.11929>. Posted October 22, 2020. Accessed August 26, 2022.
- Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. In: Crimi A, Bakas S, eds. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, 2022; 272–284.
- Park S, Kim G, Oh Y, et al. Self-evolving vision transformer for chest X-ray diagnosis through knowledge distillation. *Nat Commun* 2022;13(1):3848.
- Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV (version 1.0). PhysioNet. <https://doi.org/10.13026/S6N6-XD98>. Published 2021. Accessed August 29, 2022.
- Johnson AEW, Pollard TJ, Greenbaum NR, et al. MIMIC-CXR-JPG, a Large Publicly Available Database of Labeled Chest Radiographs. arXiv 1901.07042 [preprint]. <https://arxiv.org/abs/1901.07042>. Posted January 21, 2019. Accessed August 26, 2022.
- Hayat N, Geras KJ, Shamout FE. MedFuse: Multi-Modal Fusion With Clinical Time-Series Data and Chest X-Ray Images. arXiv 2207.07027 [preprint]. <https://arxiv.org/abs/2207.07027>. Posted July 14, 2022. Accessed July 14, 2022.
- Agency for Healthcare Research and Quality. Clinical Classifications Software (CCS) for ICD-9-CM fact sheet. Agency for Healthcare Research and Quality, 2012.
- Khader F, Han T, Müller-Franzes G, et al. Artificial Intelligence for Clinical Interpretation of Bedside Chest Radiographs. *Radiology* 2023;307(1):e220510.
- Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: A Large-Scale Hierarchical Image Database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009; 248–255.
- Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). Curran Associates, 2017; 5998–6008.
- Jagde A, Gimeno F, Brock A, Zisserman A, Vinyals O, Carreira J. Perceiver: General Perception with Iterative Attention. arXiv 2103.03206 [preprint]. <https://arxiv.org/abs/2103.03206>. Posted March 4, 2021. Accessed April 11, 2022.
- Goldberger AL, Amaral LA, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 2000;101(23):E215–E220.
- DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44(3):837–845.
- Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature* 2019;567(7748):305–307.
- Tanguay W, Acar P, Fine B, et al. Assessment of Radiology Artificial Intelligence Software: A Validation and Evaluation Framework. *Can Assoc Radiol J* 2023;74(2):326–333.
- Khader F, Hauburger C, Kirr JC, et al. Elevating Fundoscopic Evaluation to Expert Level - Automatic Glaucoma Detection Using Data from the Airog Challenge. In: 2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC) 2022; 1–4.
- De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018;24(9):1342–1350.
- Liu Y, Jain A, Eng C, et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med* 2020;26(6):900–908.
- Brown TB, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. arXiv 2005.14165 [preprint]. <https://arxiv.org/abs/2005.14165>. Posted May 28, 2020. Accessed September 12, 2022.

26. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv 1810.04805 [preprint]. <https://arxiv.org/abs/1810.04805>. Posted October 11, 2018. Accessed September 12, 2022.
27. Ghaffari Laleh N, Truhn D, Veldhuizen GP, et al. Adversarial Attacks and Adversarial Robustness in Computational Pathology. bioRxiv 2022.03.15.484515. <https://doi.org/10.1101/2022.03.15.484515>. Posted March 18, 2022. Accessed August 26, 2022.
28. Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ Digit Med* 2020;3:136.
29. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. *Nat Med* 2022;28(9):1773–1784.
30. Boehm KM, Aherne EA, Ellenson L, et al. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nat Can* 2022;3(6):723–733.
31. Jozse HRV, Shaban A, Iuzzolino ML, Koishida K. MMTM: Multimodal Transfer Module for CNN Fusion. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020; 13289–13299.
32. Pölsterl S, Wolf TN, Wachinger C. Combining 3D Image and Tabular Data via the Dynamic Affine Feature Map Transform. In: de Bruijne M, Cattin PC, Cotin S, et al, eds. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. MICCAI 2021. *Lecture Notes in Computer Science*, vol 12905. Springer, 2021; 688–698.
33. Hayat N, Geras KJ, Shamout FE. Towards Dynamic Multi-Modal Phenotyping Using Chest Radiographs and Physiological Data. arXiv 2111.02710 [preprint]. <https://arxiv.org/abs/2111.02710>. Posted November 4, 2021. Accessed August 31, 2022.
34. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;9(8):1735–1780.