Combining chest X-rays and EHR data using machine learning to diagnose acute respiratory failure

Sarah Jabbour[1], David Fouhey[1], Ella Kazerooni[2], Jenna Wiens[1], Michael W Sjoding[3*]

Affiliations:

1. Computer Science and Engineering, University of Michigan, Ann Arbor, Michigan
2. Department of Radiology, University of Michigan Medical School, Ann Arbor, Michigan
3. Pulmonary and Critical Care Medicine, Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, Michigan

***Corresponding Author Information**:
Michael W. Sjoding
G027W Building 16 NCRC, 2800 Plymouth Road, SPC 2800, Ann Arbor, MI 48109
msjoding@umich.edu

Word count: 4219

Tables/Figures: 5

References: 49

Acknowledgements

This work was supported by NIH grants K01 HL136687 (MWS) and R01 LM013325 (JW, MWS), and MM Precision Health Award (DF, MWS).

ABSTRACT

When patients develop acute respiratory failure, accurately identifying the underlying etiology is essential for determining the best treatment, but it can be challenging to differentiate between common diagnoses in clinical practice. Machine learning models could improve medical diagnosis by augmenting clinical decision making and play a role in the diagnostic evaluation of patients with acute respiratory failure. While machine learning models have been developed to identify common findings on chest radiographs (e.g. pneumonia), augmenting these approaches by also analyzing clinically relevant data from the electronic health record (EHR) could aid in the diagnosis of acute respiratory failure. Machine learning models were trained to predict the cause of acute respiratory failure (pneumonia, heart failure, and/or COPD) using chest radiographs and EHR data from patients within an internal cohort using diagnoses based on physician chart review. Models were also tested on patients in an external cohort using discharge diagnosis codes. A model combining chest radiographs and EHR data outperformed models based on each modality alone for pneumonia and COPD. For pneumonia, the combined model AUROC was 0.79 (0.78-0.79), image model AUROC was 0.73 (0.72-0.75), and EHR model AUROC was 0.73 (0.70-0.76); for COPD, combined: 0.89 (0.83-0.91), image: 0.85 (0.77-0.89), and EHR: 0.80 (0.76-0.84); for heart failure, combined: 0.80 (0.77-0.84), image: 0.77 (0.71-0.81), and EHR: 0.80 (0.75-0.82). In the external cohort, performance was consistent for heart failure and COPD, but declined slightly for pneumonia. Overall, machine learning models combing chest radiographs and EHR data can accurately differentiate between common causes of acute respiratory failure. Further work is needed to determine whether these models could aid clinicians in the diagnosis of acute respiratory failure in clinical settings.

INTRODUCTION

Acute respiratory failure develops in over three million patients hospitalized in the US annually.[1] Pneumonia, heart failure, and/or chronic obstructive pulmonary disease (COPD) are three of the most common reasons for acute respiratory failure,[2] and these conditions are among the top reasons for hospitalization for in US.[3] Determining the underlying causes of acute respiratory failure is critically important for guiding treatment decisions, but can be clinically challenging, as initial testing such as brain natriuretic peptide levels or chest radiograph results can be non-specific or difficult to interpret.[4,5] This is especially true for older adults,[6-8] patients with comorbid illnesses,[9,10] or more severe disease.[11] Incorrect initial treatment may commonly occur, resulting in worse patient outcomes or treatment delays.[6,12,13] Artificial intelligence technologies have been proposed as a strategy for improving medical diagnosis by augmenting clinical decision making,[14] and could play a role in the diagnostic evaluation of patients with acute respiratory failure.

Convolutional neural networks (CNNs) are machine learning models that can be successfully trained to identify a wide range of relevant findings on medical images, including chest radiographs.[18-21] However, for many conditions such as acute respiratory failure, the underlying medical diagnosis is not determined solely on imaging findings. Patient symptoms, physical exam findings, laboratory results and radiologic imaging findings when available are used in combination to determine the underlying cause of acute respiratory failure. Therefore, machine learning models that synthesize information contained in chest radiographs and additional clinical data in the electronic health record (EHR) may be best suited to aid clinicians in the diagnosis of these patients. However, efforts to synthesize electronic health record and imaging data for machine learning applications in healthcare have been limited to date.

Toward this end, we developed a machine learning model combing chest radiographs and clinical data from the EHR to identify pneumonia, heart failure, and COPD in patients hospitalized with acute respiratory failure. We envisioned that such a model could ultimately be used by bedside clinicians to assist in the diagnostic work-up of patients with acute respiratory failure, helping them to synthesize multi-modal data and derive estimates of the likelihood of these common conditions. We hypothesized imaging and clinical data would provide complementary information, resulting in a more accurate model that better replicates the diagnostic process. We also validated the models at an external center to determine whether combining these data improves the generalizability of the models.

We found that a model combining chest radiographs and clinical data was better able to identify pneumonia and COPD as the underlying cause of acute respiratory failure compared to models based on each data modality alone. For identifying pneumonia, the combined model's AUROC was 0.79 (range: 0.78-0.79), while an image model had an AUROC of 0.73 (range: 0.72-0.75), and an EHR model had an AUROC of 0.73 (range: 0.70-0.76). For identifying COPD, the combined model's AUROC was 0.89 (0.83-0.91), while an image model AUROC was 0.85 (0.77-0.89), and EHR model AUROC was 0.80 (0.76-0.84). For heart failure, the combined model AUROC of 0.80 (0.77-0.84) was the same as the EHR model AUROC of 0.80 (0.75-0.82), while the image only model performed slightly worse, with an AUROC of 0.77 (0.71-0.81), In the external cohort, performance was consistent for heart failure and COPD, but declined slightly for pneumonia. Overall, these results suggest that machine learning models combing chest radiograph and EHR data can accurately differentiate between common causes of acute respiratory failure and have the potential for use at the beside to support the diagnostic evaluation of these patients.

RESULTS

Study population

Models were trained using an internal cohort of patients admitted to an academic medical center in the upper Midwest (Michigan Medicine, MM) in 2017-2018 who developed acute respiratory failure (ARF) during the hospitalization. Models were externally validated on patients admitted to an academic medical center in the northeast (Beth Israel Deaconess Medical Center, BIDMC) in 2014-2016, with data available from the MIMIC-IV[22] and MIMIC-CXR[23-25] datasets. While there are many datasets containing chest radiographs,[19,20,26-29] to the best of our knowledge, MIMIC is the only large, publicly available dataset with a similar patient population that contains both chest radiographs and clinical data that can be mapped between UM and BIDMC. In both cohorts, ARF was defined as patients who required significant respiratory support (high flow nasal cannula, noninvasive mechanical ventilation, or invasive mechanical ventilation) and had a chest radiograph performed. We excluded patients who were admitted after routine surgery or a surgical related problem (see Supplement for additional details). The time of ARF diagnosis was defined as when patients first received significant respiratory support.

The internal MM cohort included 1,618 patients, with a median age of 63 years (IQR: 52-72), and 666 (41%) were female. Demographics of the external cohort were similar, although there was a higher percentage of patients with other or unknown race (**Table 1**).

Determining the cause of acute respiratory failure

For patients in the MM cohort, one or more physicians independently reviewed the entirety of each patient's hospitalization to determine whether patients had pneumonia, heart failure, and/or COPD. Physician chart review was not possible for the external BIDMC cohort because clinical notes are unavailable in MIMIC-IV. Thus, we evaluated model performance in the external cohort based on International Classification of Disease (ICD)-10 discharge diagnosis

codes (see Supplement **Table 1**). To evaluate model generalizability, we compared the model's ability to determine causes of ARF based on ICD codes in the MM and BIDMC cohorts.

In the internal cohort, there were 508 (31%) patients with pneumonia as the underlying cause of ARF, 363 (22%) with heart failure and 137 (9%) with COPD based on chart review. More than one of these diagnoses were present in 194 (11%) patients. The prevalence of pneumonia, heart failure, and COPD was lower in the BIDMC cohort than the MM cohort based on diagnosis codes (**Table 1**).

Model training and evaluation

We trained models to determine the likelihood that pneumonia, heart failure, and/or COPD was an underlying cause of ARF based on clinical data from the EHR (EHR model), chest radiographs (image model), or both data types (combined model). Additional technical details and figures illustrating model architectures are provided in the Methods and Supplement (Supplement **Figure 1)**.

The internal cohort was randomly split five times into train (60%), validation (20%) and test (20%) sets. Partitions were made at the patient level such that in each random split, data from the same patient were only in one of the train, validation, and held-out test sets. We evaluated the value of combining chest radiographs and EHR data by comparing the combined model to the EHR and image models in terms of the individual and macro average AUROC, sensitivity, and specificity for pneumonia, heart failure, and COPD when applied to the internal MM cohort test sets. Then, all models were applied to the full BIDMC cohort which served as an external validation. Finally, we calculated AUPR, macro average AUPR, positive predictive value (PPV), and (NPV) in further analyses in the Supplement (Supplement **Tables 5** & **7**).

Model performance on the internal cohort

On the MM internal test set, the combined model was more accurate than the image and EHR models in terms of AUROC, sensitivity, and specificity (**Figure 1, Table 2,** Supplement **Table 4**) for pneumonia and COPD, and performed similarly to the EHR model for heart failure. Overall, the combined model demonstrated a higher macro average AUROC (AUROC = 0.82, range: 0.80-0.85) compared to the image (AUROC = 0.77, range: 0.76-0.82) and EHR models (AUROC = 0.77, range: 0.76-0.80).

Among specific diagnoses the combined model consistently outperformed the other two models in diagnosing pneumonia and COPD and performed similarly to the EHR model in diagnosing heart failure, while the relative performance of the image and EHR models varied based on the diagnosis (**Figure 1, Table 2,** Supplement **Tables 4**). The combined model's sensitivity and specificity for diagnosing pneumonia was 70% (range: 70-76) and 71% (range: 67-78) respectively, for heart failure was 70% (range: 64-80) and 74% (68-75), and for COPD was 82% (81-85) and 84% (76-88). The image model demonstrated higher discrimination than the EHR model for diagnosing COPD, the EHR model outperformed the image model for diagnosing heart failure, and both models performed similarly when diagnosing pneumonia.

We also evaluated model performance with respect to diagnosis codes to better compare to performance in the external cohort. Model diagnostic accuracy measured using AUROC dropped moderately for pneumonia, and most significantly for COPD (**Figure 1,** Supplement **Table 4**), while increasing slightly for heart failure. Similar performance drops were seen in terms of diagnostic sensitivity with minimal changes observed for specificity (**Table 2,** Supplement **Table 6**). The performance drop across diagnoses aligned with the positive predictive value of diagnosis codes with respect to diagnoses based on chart review, with COPD having lowest positive predictive value (Supplement **Table 3**).

Model performance in the external cohort

In the external validation cohort, the combined model was consistently more accurate than other models in terms of AUROC (**Figure 1**, Supplement **Table 4**). The image model consistently outperformed the EHR model for all three diagnoses. Overall, the combined model demonstrated higher macro average AUROC (AUROC = 0.74, range: 0.73-0.74) compared to the EHR model (AUROC = 0.69, range: 0.66-0.69) and image model (AUROC = 0.72, range: 0.71-0.72) based on diagnoses codes. Between centers, there was a minimal change in the combined model AUROC performance for heart failure and COPD, with median AUROCs decreasing from 0.83 to 0.82 for heart failure and increasing from 0.73 to 0.75 for COPD, suggesting transferability. However, the decline for pneumonia was more substantial (0.74 to 0.66). Model sensitivity, specificity, PPV, NPV and AUPR are reported in the supplement (Supplement **Tables 5, 6,** & **7**).

Understanding model decisions

Since large capacity models are known to pick up on spurious features,[33] we performed a feature importance analysis to understand how models used chest radiograph and EHR data to make predictions. For chest radiographs, heatmaps were generated to understand which regions of the chest radiograph influenced the model prediction.[34] Both the image and combined models focused on appropriate areas on the chest radiograph when correctly diagnosing heart failure and pneumonia, including the lungs and heart (**Figure 2)**. Interestingly, when correctly diagnosing COPD, models seemed to focus on the trachea. In cases where models made incorrect diagnoses, they still appeared to focus on appropriate areas, such as the lungs for pneumonia, the lung and trachea for COPD, and the heart for heart failure (Supplement **Figure 2**).

To understand which EHR features were important in model decisions, we measured permutation importance. The EHR and combined models were influenced by similar clinical features with some deviations (**Table 3**). In most cases, important clinical features identified by the model aligned with the clinical understanding of diagnosis. For pneumonia, the oxygen saturation, procalcitonin level and troponin were important variables. For heart failure, brain natriuretic peptide (BNP), troponin and patient age were important variables. In contrast, variables identified as important for identifying COPD were less closely aligned with clinical understand of diagnosis for COPD, such as mean corpuscular hemoglobin (MCHC) or magnesium. For example, the combined model found that lower values of mean corpuscular hemoglobin concentration (correlation = -0.17) and higher values of magnesium (correlation = 0.05) were associated with COPD.

DISCUSSION

We developed and validated machine learning models combining chest radiographs and clinical data to determine the underlying etiology of patients with ARF. Overall, the models combining chest radiographs and clinical data led to better discriminative performance on both internal and external validation cohorts compared to models analyzing each data alone based on their macro-average AUROC and for the individual diagnoses of pneumonia and COPD. Given the diagnostic challenges of determining the underlying etiology of ARF, machine learning models may have the potential to aid clinicians in the diagnosis of these patients.

Most studies of machine learning applied to chest radiographs have used a radiologist interpretation of chest radiology studies to train models.[20,21] However, for medical conditions including pneumonia, heart failure, or COPD, a clinical diagnosis is not determined solely based on chest radiographic findings. The underlying diagnosis is based on a combination of

concordant clinical symptoms (e.g., productive cough), physical examination findings, laboratory results, and radiologic imaging findings when available. Our models more closely resemble clinical practice, as chest radiographs and other clinical data were combined, and were also trained using diagnoses determined by physicians who reviewed the entirety of each patient's hospitalization, rather than just chest radiographs.

Improving clinical diagnosis has been identified as important for improving healthcare quality,[35] and machine learning could support the diagnostic process in several ways. First, clinicians may overly focus on certain clinical data (e.g., BNP value when diagnosing heart failure) or may be prone to other cognitive errors.[36] However, models may provide more consistent estimates of disease probabilities based on the same data (though models may be prone to other errors as further discussed below). Second, models may identify features not typically considered by clinicians. For example, when diagnosing COPD, our models frequently focused on the tracheal region, whereas clinical references do not emphasize radiology findings.[37] Yet, tracheal narrowing (i.e., "saber-sheath" trachea) can be a marker of severe air-flow obstruction,[38] so training clinicians to look for this feature might also be useful. Radiologists may only apply criteria for reporting a saber-sheath trachea in severe cases, with milder transverse narrowing on front chest radiographs not considered specific enough for a diagnosis of COPD.

Importantly, the machine learning models presented in this paper are not envisioned to replace clinicians, but to provide additional information and decisions of diagnosis similar to diagnostic tests which could result in shorter time to diagnosis and treatment and add to diagnostic confidence. Clinicians have access to important diagnostic data such as subjective patient complaints or physical exam findings that are not readily available as model inputs. Thus, collaborations between clinicians and models, where clinicians consider model results in the full context of the patient's hospitalization, might be more optimal. Models may also use shortcuts,[33]

i.e., taking advantage of spurious correlations in the training data that might not hold across populations. For example, we noted that the model focused on the presence of pacemakers for heart failure (similar to Seah *et al.*[39]), but this may lead to poor performance in heart failure subpopulations without pacemakers, or overestimate the probability of heart failure when the rest of the data would suggest an alternative diagnosis. Clinicians might be able to recognize when the model is taking a shortcut and discount the model's information in such settings. Similarly, since there are no established EHR markers for COPD, the clinical variables identified as important in the permutation importance for COPD might not align with clinical intuition. Our model might be using potentially important EHR markers but could also be learning noise in the data and further investigation would be needed for confirmation.

The potential use and benefit of this model is likely in multiple scenarios. Our model identifies many important features that are likely already obvious to a bedside clinician. However, such a model is also able to consider many more features that a clinician can at once. Thus, our model is potentially useful in even easy cases where a clinician might not be able to consider all information at once. Additionally, our model may be greatly beneficial in difficult cases. Clinicians may make diagnostic and treatment errors in up to 30% of patient with acute dyspnea, particularly among patients with acute respiratory failure who cannot breathe on their own without support.[6,40-43] While our model has the potential to improve clinical care in these settings, such a model needs to be integrated carefully into clinical workflows to support the diagnostic process. However, work studying the implementation of models combining chest radiographs and EHR data is yet to be done and is important and necessary future work.

Our study has limitations. We made many modeling choices during the EHR and image data preprocessing but and make our code available so others can investigate other alternative approaches. For example, we ignore the temporal ordering of the EHR data (i.e., using only the

most recent, rather than all measurements), which may miss some relevant diagnostic information or trends in directionality of variables over time. Moreover, we only considered a limited set of EHR inputs that transfer across institutions and would be unlikely to leak labels. Additional data, such as comorbidity data, could help improve the EHR and combined models. However, we are encouraged by the fact that even with a limited set of data, we see improvements in the combined model compared to the EHR and image models in diagnosing pneumonia and COPD. We also used a simple architecture that concatenated EHR and image features, which may prevent the network from using the EHR data as guidance when extracting features from the chest radiographs earlier in the network. However, introducing EHR data at the beginning of the network requires retraining the large DenseNet-121 network, which is likely infeasible given the limited training data in the current study. While pretraining was used to enhance model performance, this does not rule out the possibility of negative transfer.[44] More pretraining data specific to the diagnostic task could improve performance as well as model pretraining that includes both structured clinical and imaging data. Finally, we used diagnosis codes as a proxy for the underlying cause of ARF in the external cohort since we did not have access to clinical notes for these patients to conduct physician chart review. However, diagnosis codes are often only moderately aligned with the actual clinical diagnosis.[45] Despite the potential differences in coding practices between institutions, model performance did not significantly drop for heart failure and COPD. Ultimately, prospective testing of such models can determine their true performance and ability to support clinicians in the diagnostic process.

In summary, models leveraging both chest radiographs and EHR data can better predict the underlying causes of ARF (pneumonia, heart failure, and/or COPD) and generalize better to another institution compared to using only radiographic or EHR data alone. These findings highlight the potential of machine learning to aid in the clinical diagnoses of pneumonia, heart

failure, and COPD. Combined with the expertise of clinicians, such models could improve the diagnostic accuracy of clinicians in this challenging clinical problem.

METHODS

This study was approved by the Institutional Review Board with a waiver of informed consent among study patients. The study followed the Transparent Reporting of Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRI-POD) reporting guidelines. All code for the analysis is made publicly available on GitHub: https://github.com/MLD3/Combining-chest-X-rays-and-EHR-data-ARF.

Determining the cause of acute respiratory failure

To determine the underlying cause of ARF in the MM cohort, one or more physicians independently reviewed the entirety of each patient's hospitalization, including the patient's medical history, laboratory, echocardiogram, chest imaging results, and response to specific treatments. Patients could be assigned to multiple diagnoses if physicians designated multiple causes of ARF, as previous research suggests that multiple concurrent etiologies may be possible.[46,47] Thus, each physician provided independent estimates of the likelihood that each of the three diagnoses (pneumonia, heart failure, and COPD) was a primary reason for patient's respiratory failure on a scale of 1-4, with 1 being very likely and 4 being unlikely. Scores were averaged across physicians and patients were assigned the diagnosis if the score was less than 2.5, since 2.5 is the midpoint of 1 and 4. Physician reviewers were board certified in internal medicine and had completed at least one year of pulmonary fellowship training.

Because clinical notes are unavailable in MIMIC-IV for BIDMC patients, we evaluated model performance in the external cohort based on International Classification of Disease (ICD)-10

discharge diagnosis codes (see Supplement **Table 1**). If ICD-10 codes for pneumonia, heart failure, or COPD were present, the patient was assigned the diagnosis as the etiology of ARF.

Chest radiograph and EHR data extraction and processing

We obtained chest radiographs nearest to the time of ARF onset (i.e., before or after ARF) in the form of digital imaging and communications in medicine (DICOM) files. Each patient had a corresponding study, containing one or most chest radiographs taken at the same time. Images were preprocessed and downsized to 512 x 512 pixels, as further described in the Supplement. EHR data included vital signs, laboratory measurements, and demographic data for which a mapping existed between MM and BIDMC (Supplement **Table 2)**. If ARF developed more than 24 hours after admission, we extracted data up until the time of ARF. If ARF developed during the first 24 hours of admission, we extracted 24 hours of data to ensure sufficient data for modeling, including data available after the chest radiograph was performed. Although EHR data collected after the time of the chest radiograph was used, we avoid temporal information leakage by excluding variables related to patient treatment, such as medications. Comorbidity data in the context of diagnosing ARF is typically useful for clinicians when making a diagnosis, but we exclude such data from our analysis as comorbidities are not particularly straightforward to accurately identify using EHR data. This is especially true when we are mapping such data between two institutions. However, in practice, such data could be included and has the potential to improve model performance further. In the case of multiple observations for the same variable, the most recent observation to the time of ARF diagnosis was used. Missing data was explicitly encoded as missing, as missingness has prognostic importance. For example, the presence or absence of a laboratory value (e.g., procalcitonin) might indicate the level of suspicion a physician might have for a particular diagnosis (e.g., pneumonia). We analyze the correlation between missingness and each diagnosis in the Supplement (Supplement **Table 8**). We used FIDDLE, an open-source preprocessing pipeline that

transforms structured EHR data into features suitable for machine learning models.[48] After preprocessing, the EHR data were represented by 326 binary features (further described in the Supplement).

Model architectures

EHR model: We trained a L2-regularize logistic regression and two-layer neural network (1 hidden layer, size = 100) with a sigmoid activation to estimate the probability of each diagnosis based on EHR data inputs.

Image model: A CNN with a DenseNet-121[30] architecture was used to estimate the probability of each diagnosis based on the chest radiograph input. The model was first pre-trained using chest radiographs from the publicly available CheXpert[20] and MIMIC-CXR-DICOM[21] datasets (excluding patients in the BIDMC validation cohort) to identify common radiographic findings annotated in radiology reports. Then the last layer of the model was fine-tuned to determine ARF diagnoses.[31]

Combined model: Chest radiographs were first passed through the pre-trained DenseNet-121 CNN to extract image features. EHR inputs were either passed through a neural network hidden layer or directly concatenated with the extracted image-based features. The concatenation was passed through an output layer with a sigmoid activation to estimate the probability of each diagnosis. Like the image model, parameters of the DenseNet-121 were frozen after pre-training.

Model evaluation

We evaluated the value of combining chest radiographs and EHR data by comparing the combined model to the EHR and image models in terms of the individual and macro average

AUROC for pneumonia, heart failure, and COPD when applied to the internal MM cohort test sets. The median and range of results on the internal cohort test sets are reported across the five splits. We then applied each of the five models trained on MM to the external BIDMC cohort, calculating performance based on diagnosis codes. We also calculated the sensitivity and specificity of the models, selecting a point on the ROC curve where the difference between sensitivity and specificity is minimized based on the MM validation set.[32] To evaluate model generalizability, we compared the model's ability to determine causes of ARF based on ICD codes in the MM and BIDMC cohorts. Finally, we calculated AUPR, positive predictive value (PPV), and (NPV) in further analyses in the Supplement (Supplement Tables **5** & **7**).

We also measured the diagnostic performance of ICD-10 codes for identifying the underlying cause of ARF based on chart review in the MM cohort in terms of sensitivity, specificity, and positive predictive value (Supplement **Table 3**).

Feature Importance

We performed a feature importance analysis to understand how models used chest radiograph and EHR data to make predictions. For chest radiographs, heatmaps were generated to understand which regions of the chest radiograph influenced the model prediction.[34] To highlight the most important regions in each image, heatmaps were normalized on a per-image basis. We qualitatively reviewed all heatmaps and identified high level patterns. Results are shown in **Figure 2** and **Figure 2** for three randomly selected patients from the group where both the image and combined models either correctly classified or incorrectly classified the diagnosis and were most confident in their predictions (i.e., those patients whose predictions were in the top 85th percentile of predictions in the test set).

To understand which EHR features were important in model decisions, we measured permutation importance. We grouped highly correlated variables together (Pearson's correlation > 0.6). Features were ranked from most to least important based on the drop in AUROC when these features were randomly shuffled across examples in the test set.[49] We averaged feature rankings across all five test sets and report the five highest ranked features for each diagnosis. Correlation between clinical variables and diagnoses was measured to determine the direction of the association.

References

1. Kempker, J.A., *et al.* The Epidemiology of Respiratory Failure in the United States 2002-2017: A Serial Cross-Sectional Study. *Crit Care Explor* **2**, e0128 (2020).
2. Stefan, M.S., *et al.* Epidemiology and outcomes of acute respiratory failure in the United States, 2001 to 2009: a national survey. *Journal of hospital medicine* **8**, 76-82 (2013).
3. HCUP Fast Stats. Healthcare Cost and Utilization Project (HCUP). April 2021. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/faststats/national/inpatientcommondiagnoses.jsp?year1=2018.
4. Roberts, E., *et al.* The diagnostic accuracy of the natriuretic peptides in heart failure: systematic review and diagnostic meta-analysis in the acute care setting. *BMJ* **350**, h910 (2015).
5. Hagaman, J.T., Rouan, G.W., Shipley, R.T. & Panos, R.J. Admission chest radiograph lacks sensitivity in the diagnosis of community-acquired pneumonia. *The American journal of the medical sciences* **337**, 236-240 (2009).
6. Ray, P., *et al.* Acute respiratory failure in the elderly: etiology, emergency diagnosis and prognosis. *Critical care (London, England)* **10**, R82 (2006).
7. Lien, C.T., Gillespie, N.D., Struthers, A.D. & McMurdo, M.E. Heart failure in frail elderly patients: diagnostic difficulties, co-morbidities, polypharmacy and treatment dilemmas. *European journal of heart failure* **4**, 91-98 (2002).
8. Metlay, J.P., *et al.* Influence of age on symptoms at presentation in patients with community-acquired pneumonia. *Archives of internal medicine* **157**, 1453-1459 (1997).
9. Daniels, L.B., *et al.* How obesity affects the cut-points for B-type natriuretic peptide in the diagnosis of acute heart failure. Results from the Breathing Not Properly Multinational Study. *Am Heart J* **151**, 999-1005 (2006).
10. Takase, H. & Dohi, Y. Kidney function crucially affects B-type natriuretic peptide (BNP), N-terminal proBNP and their relationship. *European journal of clinical investigation* **44**, 303-308 (2014).
11. Levitt, J.E., *et al.* Diagnostic utility of B-type natriuretic peptide in critically ill patients with pulmonary edema: a prospective cohort study. *Critical care (London, England)* **12**, R3 (2008).
12. Zwaan, L., Thijs, A., Wagner, C., van der Wal, G. & Timmermans, D.R. Relating faults in diagnostic reasoning with diagnostic errors and patient harm. *Academic medicine : journal of the Association of American Medical Colleges* **87**, 149-156 (2012).
13. Matsue, Y., *et al.* Time-to-Furosemide Treatment and Mortality in Patients Hospitalized With Acute Heart Failure. *Journal of the American College of Cardiology* **69**, 3042-3051 (2017).
14. National Academies of Sciences, E., and Medicine. *Improving diagnosis in health care*, (The National Academies Press, Washington, DC, 2015).
15. Liu, H., *et al.* Development and Validation of a Deep Learning System to Detect Glaucomatous Optic Neuropathy Using Fundus Photographs. *JAMA ophthalmology* **137**, 1353-1360 (2019).

16. Pasa, F., Golkov, V., Pfeiffer, F., Cremers, D. & Pfeiffer, D. Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. *Sci Rep* **9**, 6268 (2019).
17. Esteva, A.*, et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115-118 (2017).
18. Sjoding, M.W.*, et al.* Deep learning to detect acute respiratory distress syndrome on chest radiographs: a retrospective study with external validation. *Lancet Digit Health* (2021).
19. Wang, X.*, et al.* Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. 2097-2106.
20. Irvin, J.*, et al.* Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. Vol. 33 590-597.
21. Johnson, A.E.W.*, et al.* MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042* (2019).
22. Johnson, A.*, et al.* MIMIC-IV. (PhysioNet, 2020).
23. Johnson, A.E.W.*, et al.* MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data* **6**, 317 (2019).
24. Johnson, A.E.W.P., Tom J and Berkowitz, Seth Greenbaum, Nathaniel R Lungren, Matthew P Deng, Chih-ying Mark, Roger G Horng, Steven. MIMIC-CXR Database. (PhysioNet, 2019).
25. Goldberger, A.L.*, et al.* PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation* **101**, e215-e220 (2000).
26. Jaeger, S.*, et al.* Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery* **4**, 475 (2014).
27. Candemir, S.*, et al.* Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE transactions on medical imaging* **33**, 577-590 (2013).
28. Demner-Fushman, D.*, et al.* Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association* **23**, 304-310 (2016).
29. Bustos, A., Pertusa, A., Salinas, J.-M. & de la Iglesia-Vayá, M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis* **66**, 101797 (2020).
30. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K.Q. Densely connected convolutional networks. 4700-4708.
31. Pan, S.J. & Yang, Q. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* **22**, 1345-1359 (2010).
32. Youden, W.J. Index for rating diagnostic tests. *Cancer* **3**, 32-35 (1950).
33. Geirhos, R.*, et al.* Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2**, 665-673 (2020).
34. Selvaraju, R.R.*, et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. 618-626.

35. National Academies of Sciences, E.a.M. *Improving diagnosis in health care*, (National Academies Press, 2015).
36. Croskerry, P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad Med* **78**, 775-780 (2003).
37. Rabe, K.F.*, et al.* Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *American journal of respiratory and critical care medicine* **176**, 532-555 (2007).
38. Ciccarese, F.*, et al.* Saber-sheath trachea as a marker of severe airflow obstruction in chronic obstructive pulmonary disease. *La radiologia medica* **119**, 90-96 (2014).
39. Seah, J.C.Y., Tang, J.S.N., Kitchen, A., Gaillard, F. & Dixon, A.F. Chest radiographs in congestive heart failure: visualizing neural network learning. *Radiology* **290**, 514-522 (2019).
40. Zwaan, L., Thijs, A., Wagner, C., van der Wal, G. & Timmermans, D.R.M. Relating faults in diagnostic reasoning with diagnostic errors and patient harm. *Academic Medicine* **87**, 149-156 (2012).
41. Dharmarajan, K.*, et al.* Treatment for multiple acute cardiopulmonary conditions in older adults hospitalized with pneumonia, chronic obstructive pulmonary disease, or heart failure. *Journal of the American Geriatrics Society* **64**, 1574-1582 (2016).
42. Matsue, Y.*, et al.* Time-to-furosemide treatment and mortality in patients hospitalized with acute heart failure. *Journal of the American College of Cardiology* **69**, 3042-3051 (2017).
43. Seymour, C.W.*, et al.* Delays from first medical contact to antibiotic administration for sepsis. *Critical care medicine* **45**, 759 (2017).
44. Wang, Z., Dai, Z., Póczos, B. & Carbonell, J. Characterizing and avoiding negative transfer. 11293-11302.
45. O'Malley, K.J.*, et al.* Measuring diagnoses: ICD code accuracy. *Health services research* **40**, 1620-1639 (2005).
46. Corrales-Medina, V.F.*, et al.* Cardiac complications in patients with community-acquired pneumonia: incidence, timing, risk factors, and association with short-term mortality. *Circulation* **125**, 773-781 (2012).
47. Wells, J.M.*, et al.* Pulmonary Arterial Enlargement and Acute Exacerbations of COPD. *New England Journal of Medicine* **367**, 913-921 (2012).
48. Tang, S.*, et al.* Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. *Journal of the American Medical Informatics Association* **27**, 1921-1934 (2020).
49. Altmann, A., Toloşi, L., Sander, O. & Lengauer, T. Permutation importance: a corrected feature importance measure. *Bioinformatics* **26**, 1340-1347 (2010).

Table 1. Characteristics of the internal and external cohorts.

Characteristic	MM internal cohort (n=1618)	BIDMC external cohort (n=1774)
Age, median (IQR)	63 (52-72)	63 (48-75)
Gender		
Male	952 (59)	1020 (57)
Female	666 (41)	754 (43)
Race		
White	1364 (84)	904 (51)
Black	159 (10)	151 (9)
Other/Unknown	95 (6)	719 (41)
Acute Respiratory failure etiology		
Pneumonia	508 (31)	NA
Heart Failure	363 (22)	NA
COPD	137 (9)	NA
Pneumonia & Heart Failure	82 (5)	NA
Pneumonia & COPD	64 (4)	NA
COPD & Heart Failure	35 (2)	NA
All 3	13 (1)	
Diagnosis codes		
Pneumonia	661 (41)	322 (18)
Heart Failure	490 (30)	204 (11)
COPD	423 (26)	70 (4)
Pneumonia & Heart Failure	217 (13)	103 (0.06)
Pneumonia & COPD	196 (12)	46 (3)
COPD & Heart Failure	195 (12)	29 (2)
All 3	90 (6)	21 (1)

Acute respiratory failure etiology was determined based on retrospective chart review performed by one or more physicians.
Diagnosis codes are the International Classification of Disease-10 diagnosis codes assigned to the hospitalization.
Abbreviations: NA: not available; IQR: interquartile range; COPD: chronic obstructive pulmonary disease;

Table 2. Model sensitivity and specificity for detecting the underlying etiology of acute respiratory failure on patients in the internal MM cohort.

	Combined model		Image Model		EHR model	
Diagnosis	**Sensitivity % (range)**	**Specificity % (range)**	**Sensitivity % (range)**	**Specificity % (range)**	**Sensitivity % (range)**	**Specificity % (range)**
Pneumonia	70 (70-76)	71 (67-78)	64 (64-68)	70 (67-71)	64 (59-69)	71 (65-73)
Heart Failure	70 (64-80)	74 (68-75)	65 (61-85)	69 (66-76)	74 (61-78)	71 (68-75)
COPD	82 (81-85)	84 (76-88)	81 (72-89)	81 (69-81)	71 (62-84)	77 (66-80)

The median and range of model sensitivity and specificity are reported when applied to the held-out test sets for each of the 5 splits of the internal cohort. Sensitivity and specificity were calculated at the point on the ROC curve where, based on the validation sets, the difference between sensitivity and specificity is minimized.

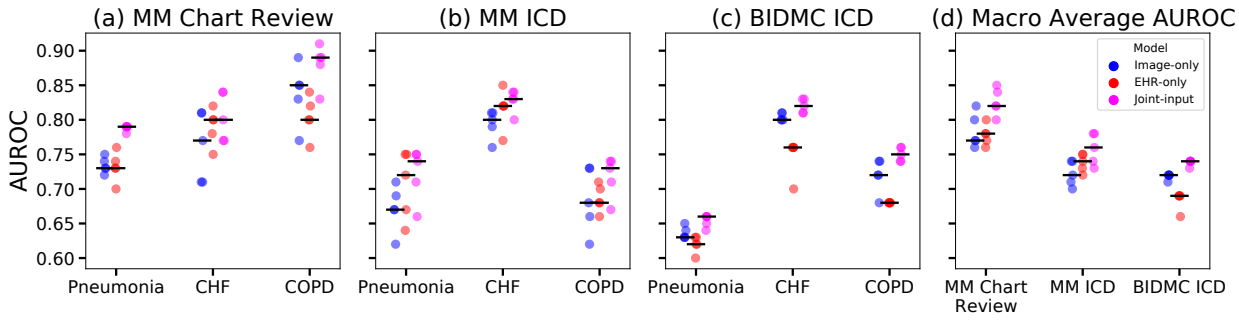
Table 3. Top five important clinical features used by the EHR and combined models to identify etiologies of acute respiratory failure.

Diagnosis	EHR model	Combined model
Pneumonia	Oxygen saturation or PaO_2	Oxygen saturation or PaO_2
	Procalcitonin	Procalcitonin
	Plateau pressure*	Troponin-I
	Troponin-I	Absolute lymphocyte count
	Absolute lymphocyte count	Plateau pressure*
Heart Failure	BUN or Creatinine	BUN or Creatinine
	BNP	Age
	Troponin-I	Troponin-I
	Age	BNP
	Tidal Volume*	Tidal Volume*
COPD	MCHC	MCHC
	Oxygen saturation or PaO_2	Magnesium
	Lymphocytes % or Neutrophils %	Total bilirubin
	Age	BUN or Creatinine
	Bicarbonate	Alanine aminotransferase or Asparate aminotransferase

Top features identified by permutation importance. Highly correlated features (>0.6) were grouped together during the permutation importance analysis and reported together (e.g. BUN or Creatinine).
*Plateau pressure and tidal volume measured during invasive mechanical ventilation.
Abbreviations: PaO_2: Partial pressure of oxygen; MCHC: mean corpuscular hemoglobin concentration; BUN: Blood urea nitrogen; BNP: Brain natriuretic peptide.

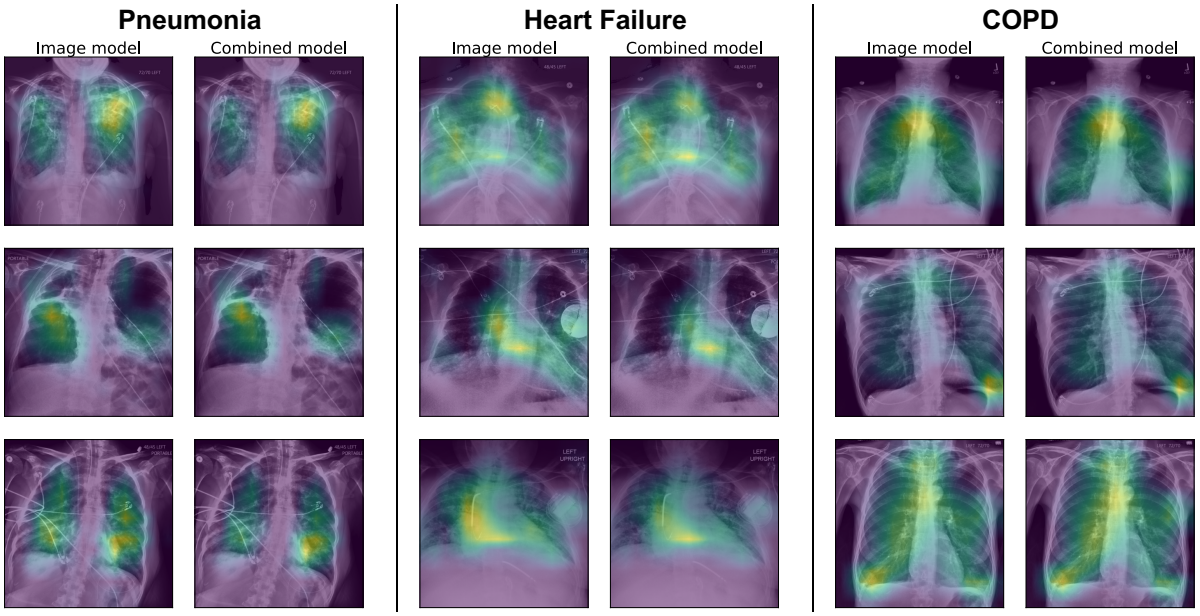
Figures

Figure 1. Performance of the image, EHR, and combined models on the internal and external cohorts.



Model performance evaluated based on the area under the receiver operator characteristic curve (AUROC). Black horizontal lines indicate median performance for each model. When the models were evaluated using diagnosis based on chart reviews in the internal cohort (a), the combined model outperforms the image and EHR models on most data splits in terms of AUROC for identifying pneumonia and COPD, and better for one of the five data splits for diagnosing heart failure. Model performance decreased for pneumonia and COPD when evaluated based on discharge diagnosis codes (b). Model performance on the external cohort was based on discharge diagnosis codes (c) and was similar to the internal cohort (b) with the exception of pneumonia. The combined model consistently outperformed the other models across cohorts when evaluated using macro average AUROC which combines model performance across all three diagnoses. Abbreviations: COPD: chronic obstructive pulmonary disease.

Figure 2. Example chest radiograph heatmaps in patients where the model correctly diagnosed pneumonia, heart failure or COPD with high probability.



Chest radiographs are shown for patients that the model correctly classified as positive for each disease with high probability. The overlying heatmap generated by Grad-CAM highlights the regions that the model focuses on when estimating the likely diagnosis (blue: low contribution, yellow: high contribution). For both the image and combined model, the model looks at the lung regions for pneumonia and COPD and the heart region for heart failure. Heatmaps were normalized on individual images to highlight the most important areas of each image, therefore heatmap values should not be compared across images. Image processing was performed, including histogram equalization to increase contrast in the original images, and then images were resized to 512x512 pixels.

SUPPLEMENT

Combining chest X-rays and EHR data using machine learning to diagnose acute respiratory failure

Methods

Cohort selection

Michigan Medicine (MM) internal cohort: Patients were included if developed acute respiratory failure during hospitalizations in 2016 and 2017, which was defined as the need for high flow nasal cannula, endotracheal tube, or bipap mask based on respiratory flowsheet documentation during the first 7 days of their hospitalization. Patients were excluded if they were admitted to the neurologic or cardiovascular vascular ICU after a surgical procedure.

Beth Israel Deaconess Medical Center (BIDMC) external cohort: Patients were included if they who received supplemental oxygen in the form of high flow nasal cannula, endotracheal tube, or bipap mask and had linked chest radiographic images in the MIMIC-CXR dataset and clinical data in the MIMIC-IV dataset. Patients who were admitted for a surgical related problem were excluded; specifically, if a patient received oxygen support while admitted under a surgical unit (CSURG, NSURG, ORTHO, SURG, TSURG, or VSURG) or within 24 hours after leaving a surgical unit, they were excluded from the cohort.

Additional details of data preprocessing

After obtaining chest radiographic images in the form of digital imaging and communications in medicine (DICOM) files, global histogram equalization was first applied to the images to increase contrast in the original images. Then, images were resized, while preserving their aspect ratio, such that their smaller axis was 512 pixels. We randomly cropped the training images to $512x512$ and augmented with random in-plane rotations up to 15 degrees. Validation and test images were center cropped to $512x512$.

To process the EHR data, we used FIDDLE, an open-source preprocessing pipeline that transforms structured EHR data into feature vectors suitable for machine learning models.[1] FIDDLE maps all variables (e.g., temperature = 37℃) into five binary features (e.g., [0,1,0,0,0]) corresponding to ranges of values (e.g., [35-36, 37-38, 39-40, 40-41, 41-42]), and accounts for missingness by setting all values in the feature vector to zero. After preprocessing, the EHR data were represented by 326 binary features.

Model training details and hyperparameter tuning

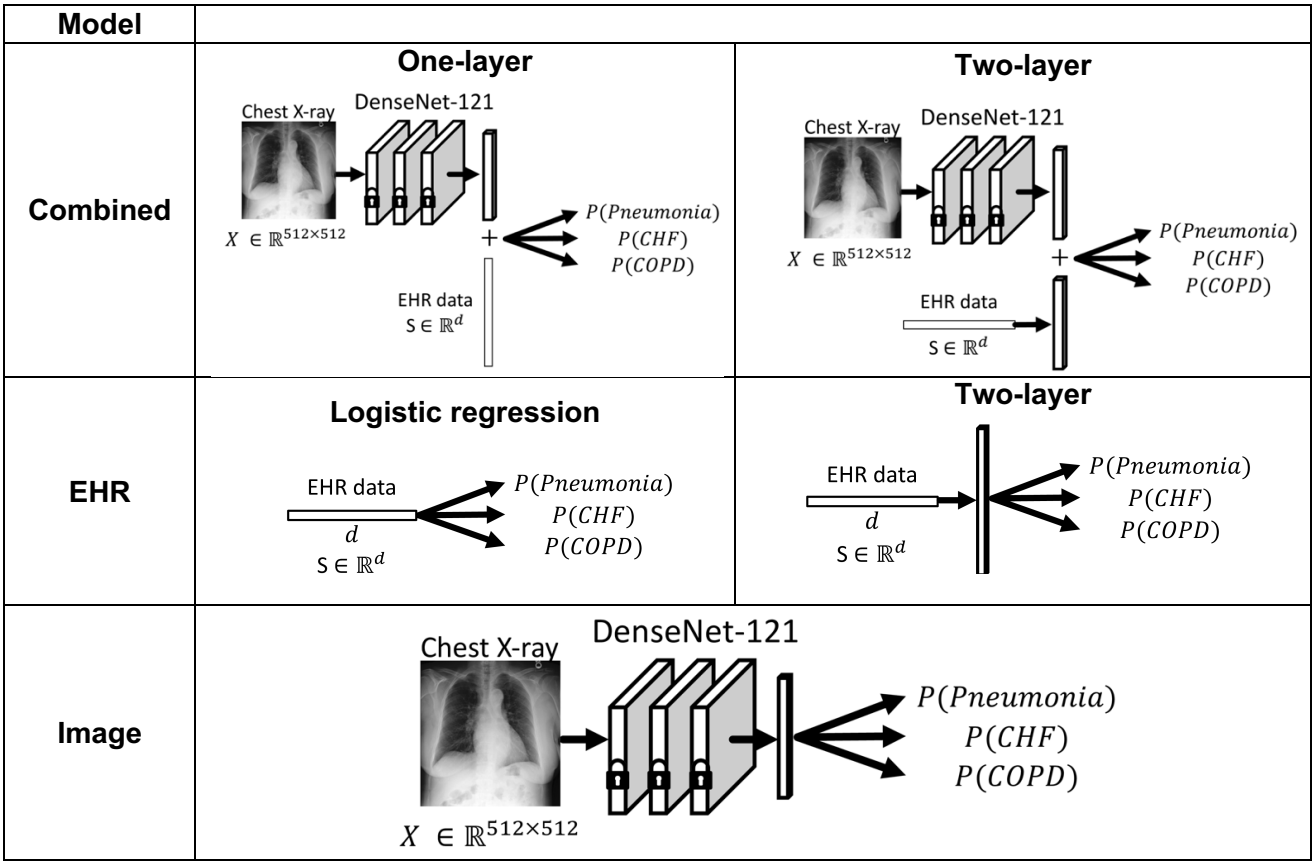
In all cases, model parameters were learned using stochastic gradient descent with momentum to minimize cross-entropy loss based on the chart review diagnostic labels. Since some patients had multiple chest radiographs taken at the same time, models were applied to all chest radiographs and the predictions were averaged. Final models and hyperparameters including the learning rate, momentum, and weight decay were selected based on validation AUROC performance, and a patience of 5 was used for early-stopping. We swept the learning rate from [10^{-4}, 10^{-3}, 10^{-2}, 10^{-1},1, 3], momentum from 0.8 and 0.9, and weight decay from [10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}]. A batch size of 32 was used throughout. For the EHR model, two architectures were swept: one and two-layer (hidden units: 100, ReLu activation) neural networks with sigmoid output activation. Similarly for the combined model, two architectures were swept: one where EHR data were concatenated with Image-based features, and one with a hidden layer (hidden units: 100, ReLu activation) after EHR data before concatenation with Image-based features.

Model initialization

We initialized the DenseNet-121 model first using pre-trained weights on CheXpert[2] and then MIMIC-CXR[3] chest radiographs that were excluded from the external validation cohort. Histogram equalization was first applied to the images to increase contrast in the original images, and then images were resized such that their smaller axis was 512 pixels while preserving their aspect ratio. This allowed for cropping the images horizontally or vertically to a square image as input to the model. We trained the model to predict text-mined radiology report labels and optimized the sum of the *masked* binary cross-entropy loss across labels, masking the loss for labels with a missing value. Following Irvin *et al.,* we used Adam with default parameters of $\beta_1 = 0.9$ and $\beta_2 = 0.999$, learning rate of 10^{-4}, and a batch size of 16.[2] We trained for 3 epochs with 3 different random initializations, saving checkpoints every 4,800 batches. We first trained on the CheXpert data and selected the checkpoint that performed the best on a the CheXpert validation set of size 200, measured by average area under the receiver operator characteristic curve (AUROC) across all 14 labels. Then, we trained on the MIMIC-CXR data and again selected the checkpoint that performed the best on a randomly sampled validation set of size 5000.

Model Architecture

Figure 1. Model architectures.



The image model consisted of a DenseNet-121 with a final fully connected layer and sigmoid output activation for each diagnosis. The EHR model consists of a one- or two-layer neural network with sigmoid output activation for each diagnosis. The combined model merged the design of the image and EHR models by (i) passing chest radiographs through the frozen DenseNet-121 to extract image-based features, (ii) concatenating these features with EHR-based features, and (iii) passing these features through the output layer followed by a sigmoid output activation for each diagnosis. lock: frozen parameters. d: dimension of input EHR data. X: chest radiograph input.

Table 1. International Disease classification 10 codes for pneumonia, heart failure, and COPD.

Diagnosis	ICD Code	Description
Pneumonia	J69.0	Pneumonitis due to inhalation of food and vomit
	A48.1	Legionnaires' disease
	J09.X1	Influenza due to identified novel influenza A virus with pneumonia
	J10.00	Influenza due to other identified influenza virus with unspecified type of pneumonia
	J10.01	Influenza due to other identified influenza virus with the same other identified influenza virus pneumonia
	J10.08	Influenza due to other identified influenza virus with other specified pneumonia
	J11.00	Influenza due to unidentified influenza virus with unspecified type of pneumonia
	J11.08	Influenza due to unidentified influenza virus with specified pneumonia
	J12.0	Adenoviral pneumonia
	J12.1	Respiratory syncytial virus pneumonia
	J12.2	Parainfluenza virus pneumonia
	J12.3	Human metapneumovirus pneumonia
	J12.81	Pneumonia due to SARS-associated coronavirus
	J12.89	Other viral pneumonia
	J12.9	Viral pneumonia, unspecified
	J13	Pneumonia due to Streptococcus pneumoniae
	J14	Pneumonia due to Hemophilus influenzae
	J15.0	Pneumonia due to Klebsiella pneumoniae
	J15.1	Pneumonia due to Pseudomonas
	J15.20	Pneumonia due to staphylococcus, unspecified
	J15.211	Pneumonia due to Methicillin susceptible Staphylococcus aureus
	J15.212	Pneumonia due to Methicillin resistant Staphylococcus aureus
	J15.29	Pneumonia due to other staphylococcus
	J15.3	Pneumonia due to streptococcus, group B
	J15.4	Pneumonia due to other streptococci
	J15.5	Pneumonia due to Escherichia coli
	J15.6	Pneumonia due to other Gram-negative bacteria
	J15.7	Pneumonia due to Mycoplasma pneumoniae
	J15.8	Pneumonia due to other specified bacteria
	J15.9	Unspecified bacterial pneumonia
	J16.0	Chlamydial pneumonia
	J16.8	Pneumonia due to other specified infectious organisms
	J18.0	Bronchopneumonia, unspecified organism
	J18.1	Lobar pneumonia, unspecified organism
	J18.8	Other pneumonia, unspecified organism

Diagnosis	ICD Code	Description
Pneumonia	J18.9	Pneumonia, unspecified organism
Heart Failure	I11.0	Hypertensive heart disease with heart failure
	I13.0	Hypertensive heart and chronic kidney disease with heart failure and stage 1 through stage 4 chronic kidney disease
	I13.2	Hypertensive heart and chronic kidney disease with heart failure and with stage 5 chronic kidney disease, or end stage renal disease
	I50.1	Left ventricular failure, unspecified
	I50.20	Unspecified systolic (congestive) heart failure
	I50.21	Acute systolic (congestive) heart failure
	I50.22	Chronic systolic (congestive) heart failure
	I50.23	Acute diastolic (congestive) heart failure
	I50.30	Unspecified diastolic (congestive) heart failure
	I50.31	Acute diastolic (congestive) heart failure
	I50.32	Chronic diastolic (congestive) heart failure
	I50.33	Acute on chronic diastolic (congestive) heart failure
	I50.40	Unspecified combined systolic (congestive) and diastolic (congestive) heart failure
	I50.41	Acute combined systolic (congestive) and diastolic (congestive) heart failure
	I50.42	Chronic combined systolic (congestive) and diastolic (congestive) heart failure
	I50.43	Acute on chronic combined systolic (congestive) and diastolic (congestive) heart failure
	I50.810	Right heart failure, unspecified
	I50.811	Acute right heart failure
	I50.812	Chronic right heart failure
	I50.813	Acute on chronic right heart failure
	I50.814	Right heart failure due to left heart failure
	I50.82	Biventricular heart failure
	I50.83	High output heart failure
	I50.84	End stage heart failure
	I50.89	Other heart failure
	I50.9	Heart failure, unspecified
COPD	J41.0	Simple chronic bronchitis
	J41.1	Mucopurulent chronic bronchitis
	J41.8	Mixed simple and mucopurulent chronic bronchitis
	J42	Unspecified chronic bronchitis
	J43.0	Unilateral pulmonary emphysema [MacLeod's syndrome]
	J43.1	Panlobular emphysema
	J43.2	Centrilobular emphysema

J43.8	Other emphysema
J43.9	Emphysema, unspecified
J44.0	Chronic obstructive pulmonary disease with acute lower respiratory infection
J44.1	Chronic obstructive pulmonary disease with (acute) exacerbation
J44.9	Chronic obstructive pulmonary disease, unspecified

Abbreviations: COPD: chronic obstructive pulmonary disease.

Table 2. Feature mapping from Michigan Medicine to Beth Israel Deaconess

	MM Feature Name	MIMIC-IV/BIDMC Feature Name
Vital Signs	Diastolic blood pressure	Arterial Blood Pressure diastolic
	Diastolic blood pressure	ART BP Diastolic
	Diastolic blood pressure	Non Invasive Blood Pressure diastolic
	Diastolic blood pressure	Manual Blood Pressure Diastolic Left
	Fraction inspired oxygen	Inspired O2 Fraction
	Heart Rate	Heart Rate
	Height	Height (cm)
	Mean blood pressure	Arterial Blood Pressure mean
	Mean blood pressure	ART BP mean
	Mean blood pressure	IABP Mean
	Mean blood pressure	Non Invasive Blood Pressure mean
	Pulse oximetry	O2 saturation pulse oxymetry
	Peak inspiratory pressure	Peak Insp. Pressure
	Positive end-expiratory pressure Set	PEEP set
	Respiratory rate	Respiratory Rate
	Respiratory rate	Respiratory Rate (spontaneous)
	Respiratory rate	Spont RR
	Respiratory rate	Respiratory Rate (Total)
	Respiratory Rate (Set)	Respiratory Rate (Set)
	Systolic blood pressure	Arterial Blood Pressure systolic
	Systolic blood pressure	ART BP Systolic
	Systolic blood pressure	Non Invasive Blood Pressure systolic
	Systolic blood pressure	Manual Blood Pressure Systolic Left
	Systolic blood pressure	Manual Blood Pressure Systolic Right
	Temperature (C)	Temperature Celsius
	Weight	Admission Weight (Kg)
	Plateau Pressure	Plateau Pressure
	Tidal Volume Observed	Tidal Volume (observed)
	Tidal Volume Set	Tidal Volume (set)
	Tidal Volume Spontaneous	Tidal Volume (spontaneous)
Lab Results	Alanine aminotransferase	ALANINE AMINOTRANSFERASE (ALT)
	Albumin	ALBUMIN
	Alkaline phosphate	ALKALINE PHOSPHATASE
	Asparate aminotransferase	ASPARATE AMINOTRANSFERASE (AST)
	Basophils	BASOPHILS %
	Bicarbonate	BICARBONATE
	Bilirubin (conjugated)	BILIRUBIN, DIRECT
	Bilirubin (total)	BILIRUBIN, TOTAL
	Bilirubin (unconjugated)	BILIRUBIN, INDIRECT
	Blood urea nitrogen	UREA NITROGEN
	Calcium (total)	CALCIUM, TOTAL
	Calcium ionized	FREE CALCIUM
	Chloride	CHLORIDE
	Cholesterol (total)	CHOLESTEROL, TOTAL
	Cholesterol (HDL)	CHOLESTEROL, HDL
	Creatinine	CREATININE
	Eosinophils (blood)	EOSINOPHILS %

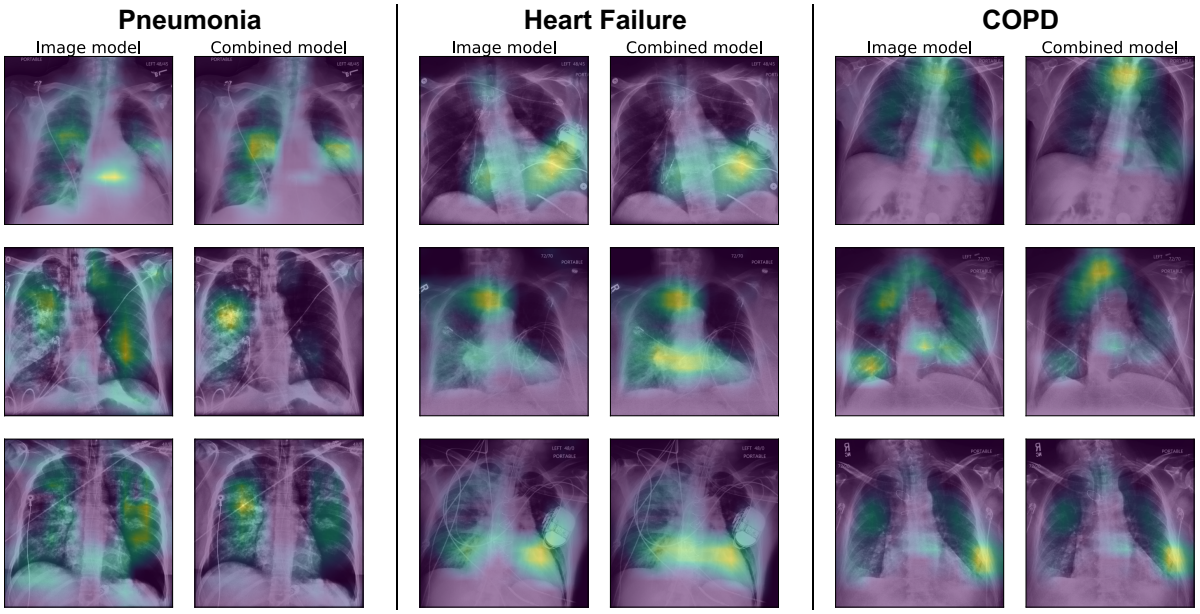
	Glucose	GLUCOSE
	Hematocrit	HEMATOCRIT
	Hemoglobin	HEMOGLOBIN
	Lactate	LACTATE
	Lactate dehydrogenase	LACTATE DEHYDROGENASE (LD)
	Lymphocytes	LYMPHOCYTES
	Lymphocytes (absolute)	ABSOLUTE LYMPHOCYTE COUNT
	Magnesium	MAGNESIUM
	Mean corpuscular hemoglobin	MCH
	Mean corpuscular hemoglobin concentration	MCHC
	Mean corpuscular volume	MCV
	Monocytes	MONOCYTES
	Neutrophils	NEUTROPHILS
	partial pressure of oxygen	PO2
	Partial pressure of carbon dioxide	PCO2
	Oxygen saturation	OXYGEN SATURATION
	Partial thromboplastin time	PTT
	pH	PH
	Phosphate	PHOSPHATE
	Platelet Count	PLATELET COUNT
	Potassium	POTASSIUM
	Prothrombin time	INR(PT)
	Red blood cell count	RED BLOOD CELLS
	Sodium	SODIUM
	Troponin-I	TROPONIN T
	White blood cell count	WHITE BLOOD CELLS
	Fibrinogen	FIBRINOGEN, FUNCTIONAL
	BNP	NTproBNP
	Procalcitonin	---
Static Variables	Age	anchor_age
	Gender	gender
	Race	ethnicity

Table 3. Accuracy of discharge diagnosis codes for identifying the etiology of acute respiratory failure.

Diagnosis	Sensitivity	Specificity	Positive Predictive Value
Pneumonia	0.86	0.73	0.60
Heart Failure	0.80	0.79	0.46
COPD	0.96	0.80	0.29

Accuracy of the discharge diagnosis codes is based on retrospective chart review.

Figure 2. Example chest radiograph heatmaps in patients where the model incorrectly diagnoses pneumonia, heart failure or COPD.



Chest radiographs are shown for patients that the model incorrectly classified as positive for each disease with high probability. The overlying heatmap generated by Grad-CAM highlights the regions that the model focus on when estimating the likely diagnosis (blue: low contribution, yellow: high contribution). For both the image and combined models, the model looks at the lung regions for pneumonia and COPD and the heart region for heart failure. Heatmaps are normalized on individual images to highlight the most important areas of each image, therefore heatmap values should not be compared across images. Image processing was performed, including histogram equalization to increase contrast in the original images, and then images were resized to 512x512 pixels.

Table 4. Performance of image, EHR and combined models on the internal MM held-out test set and external validation cohort in terms of AUROC.

Cohort and Model	Pneumonia	Heart Failure	COPD	Macro-Average AUROC
MM chart review (n, % pos)	**n=324 33% (32-36)**	**n=324 22% (20-23)**	**n=324 7% (5-8)**	**--**
Image	0.73 (0.72-0.75)	0.77 (0.71-0.81)	0.85 (0.77-0.89)	0.77 (0.76-0.82)
EHR	0.73 (0.70-0.76)	0.80 (0.75-0.82)	0.80 (0.76-0.84)	0.77 (0.76-0.80)
Combined	0.79 (0.78-0.79)	0.80 (0.77-0.84)	0.89 (0.83-0.91)	0.82 (0.80-0.85)
MM diagnosis codes (n-% pos)	**n=324 44% (38-47)**	**n=324 30% (24-34)**	**n=324 26% (24-27)**	**--**
Image	0.67 (0.62-0.71)	0.80 (0.76-0.81)	0.68 (0.62-0.73)	0.71 (0.69-0.74)
EHR	0.72 (0.64-0.75)	0.82 (0.77-0.85)	0.68 (0.66-0.71)	0.73 (0.72-0.75)
Combined	0.74 (0.66-0.75)	0.83 (0.80-0.84)	0.73 (0.67-0.74)	0.76 (0.73-0.78)
BIDMC diagnosis codes (n-% pos)	**n=1774 19%**	**n=1774 13%**	**n=1774 9%**	**--**
Image	0.63 (0.63-0.65)	0.80 (0.80-0.81)	0.72 (0.68-0.74)	0.72 (0.71-0.72)
EHR	0.62 (0.60-0.63)	0.76 (0.70-0.76)	0.68 (0.68-0.68)	0.69 (0.66-0.69)
Combined	0.66 (0.64-0.66)	0.82 (0.81-0.83)	0.75 (0.74-0.76)	0.74 (0.73-0.74)

Performance as determined based on the area under the receiver operator characteristic curve (AUROC). The internal cohort was randomly split five times into train (60%), validation (20%) and test (20%) sets. The median AUROC and AUROC range are reported for models trained on each split. The resulting five models were applied to the external cohort and the median AUROC and AUROC range are reported for models.

Table 5. Performance of image, EHR and combined models on the internal and external cohorts in terms of AUPR.

Cohort and Model	Pneumonia	Heart Failure	COPD	Macro-Average AUPR
MM chart review (n, % pos)	**n=324 33% (32-36)**	**n=324 22% (20-23)**	**n=324 7% (5-8)**	**--**
Image	0.58 (0.58-0.64)	0.49 (0.44-0.57)	0.56 (0.32-0.58)	0.53 (0.46-0.57)
EHR	0.60 (0.57-0.65)	0.55 (0.44-0.58)	0.42 (0.37-0.44)	0.51 (0.49-0.53)
Combined	0.67 (0.64-0.72)	0.56 (0.53-0.64)	0.71 (0.42-0.74)	0.64 (0.55-0.67)
MM diagnosis codes (n, % pos)	**n=324 44% (38-47)**	**n=324 30% (24-34)**	**n=324 26% (24-27)**	**--**
Image	0.63 (0.53-0.64)	0.66 (0.57-0.67)	0.49 (0.37-0.53)	0.56 (0.55-0.61)
EHR	0.62 (0.54-0.68)	0.71 (0.56-0.71)	0.46 (0.40-0.51)	0.59 (0.55-0.62)
Combined	0.67 (0.55-0.70)	0.72 (0.65-0.74)	0.56 (0.47-0.58)	0.62 (0.60-0.67)
BIDMC diagnosis codes (n, % pos)	**n=1774 19%**	**n=1774 13%**	**n=1774 9%**	**--**
Image	0.29 (0.27-0.29)	0.40 (0.38-0.41)	0.22 (0.16-0.23)	0.30 (0.28-0.31)
EHR	0.26 (0.24-0.27)	0.30 (0.24-0.34)	0.20 (0.20-0.20)	0.25 (0.23-0.27)
Combined	0.29 (0.27-0.29)	0.41 (0.39-0.43)	0.25 (0.24-0.26)	0.31 (0.31-0.32)

Performance as determined based on the area under the precision-recall curve (AUPR). The internal cohort was randomly split five times into train (60%), validation (20%) and test (20%) sets. The median AUPR and AUPR range are reported for models trained on each split. The resulting five models were applied to the external cohort and the median AUPR and AUPR range are reported for models.

Table 6. Sensitivity and specificity of the combined, image, and EHR models based on discharge diagnosis codes.

Internal cohort (n = 324)	Combined model		Image model		EHR model	
Diagnosis	**Sensitivity % (range)**	**Specificity % (range)**	**Sensitivity % (range)**	**Specificity % (range)**	**Sensitivity % (range)**	**Specificity % (range)**
Pneumonia	63 (53-65)	69 (67-75)	57 (52-59)	69 (66-71)	54 (52-64)	69 (69-72)
Heart failure	66 (64-75)	77 (74-81)	67 (56-77)	77 (70-78)	71 (67-75)	74 (74-79)
COPD	44 (43-49)	87 (78-91)	49 (42-50)	82 (72-85)	47 (43-60)	80 (71-83)

External cohort (n = 1774)	Combined model		Image model		EHR model	
Diagnosis	**Sensitivity % (range)**	**Specificity % (range)**	**Sensitivity % (range)**	**Specificity % (range)**	**Sensitivity % (range)**	**Specificity % (range)**
Pneumonia	39 (30-44)	81 (77-85)	47 (43-49)	74 (73-75)	38 (31-41)	76 (76-81)
Heart failure	76 (66-78)	71 (70-80)	75 (69-84)	69 (63-76)	71 (69-74)	68 (58-71)
COPD	64 (62-73)	75 (71-76)	68 (66-74)	62 (58-72)	42 (41-68)	81 (55-83)

Sensitivity and specificity are calculated at the point on the ROC curve where, based on the Michigan Medicine (MM) validation set, the difference between sensitivity and specificity is minimized. Abbreviations: COPD: chronic obstructive pulmonary disease; MM: Michigan Medicine; BIDMC: Beth Israel Deaconess Medical Center.

Table 7. Performance of image, EHR and combined models on the internal MM held-out test set and external validation cohort in terms of positive predictive value and negative predictive value.

	Combined model		Image Model		EHR model	
MM chart review (n = 324)	**PPV % (range)**	**NPV % (range)**	**PPV % (range)**	**NPV % (range)**	**PPV % (range)**	**NPV % (range)**
Pneumonia	54 (52-64)	84 (82-86)	49 (48-54)	80 (78-81)	50 (48-54)	81 (75-82)
Heart Failure	42 (39-43)	91 (87-93)	38 (38-40)	90 (85-94)	41 (37-43)	91 (88-92)
COPD	33 (18-36)	98 (98-99)	27 (13-30)	98 (97-99)	21 (10-23)	97 (96-98)

MM diagnosis codes (n = 324)	**Combined model**		**Image model**		**EHR model**	
Diagnosis	**PPV % (range)**	**NPV % (range)**	**PPV % (range)**	**NPV % (range)**	**PPV % (range)**	**NPV % (range)**
Pneumonia	64 (49-66)	70 (65-73)	61 (48-62)	66 (63-71)	61 (52-63)	70 (62-71)
Heart failure	57 (50-64)	85 (82-89)	53 (46-61)	82 (82-91)	59 (49-60)	85 (85-89)
COPD	54 (41-62)	83 (80-83)	47 (33-52)	81 (79-83)	46 (38-46)	81 (81-83)

BIDMC diagnosis codes (n = 1774)	**Combined model**		**Image model**		**EHR model**	
Diagnosis	**PPV % (range)**	**NPV % (range)**	**PPV % (range)**	**NPV % (range)**	**PPV % (range)**	**NPV % (range)**
Pneumonia	31 (30-32)	85 (84-86)	29 (28-29)	86 (85-86)	27 (25-28)	84 (84-85)
Heart failure	29 (27-34)	95 (94-96)	27 (25-30)	95 (94-96)	25 (21-26)	94 (94-94)
COPD	20 (20-20)	96 (95-96)	16 (14-18)	96 (95-96)	18 (13-19)	93 (93-95)

Positive predictive value, and negative predictive value are calculated at the point on the ROC curve where, based on the Michigan Medicine (MM) validation set, the difference between sensitivity and specificity is minimized. Abbreviations: PPV: Positive predictive value; NPV: negative predictive value; COPD: chronic obstructive pulmonary disease; MM: Michigan Medicine; BIDMC: Beth Israel Deaconess Medical Center.

Table 8. Top 5 correlations between diagnoses and the presence of EHR features.

Diagnosis	EHR Feature Name	Correlation
Pneumonia	Procalcitonin	0.28
	BNP	0.19
	Absolute lymphocyte count	0.17
	Monocytes	0.17
	Basophils	0.17
Heart Failure	BNP	0.30
	Tidal Volume Observed	-0.26
	Tidal Volume Set	-0.26
	Plateau Pressure	-0.25
	Troponin-I	-0.23
COPD	Tidal Volume Set	-0.19
	BNP	0.18
	Tidal Volume Observed	-0.18
	Plateau Pressure	-0.17
	Respiratory Rate (Set)	-0.17

For each patient, each EHR feature out of the 68 total UM features was labelled as missing (0) or not (1). The correlation between missingness and each diagnosis was then measured. The presence of certain measures correlates with the prognostic importance of such measures in clinical practice. For example, the presence of procalcitonin correlated with a patient having pneumonia, and the presence of BNP correlated with a patient having heart failure. Although there aren't established EHR markers for COPD, we observe that the presence of tidal volume set has a negative correlation with COPD.

Supplement References

1. Tang S, Davarmanesh P, Song Y, Koutra D, Sjoding MW, Wiens J. Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. *Journal of the American Medical Informatics Association*. 2020;27(12):1921-1934.
2. Irvin J, Rajpurkar P, Ko M, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. 590-597.
3. Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*. 2019/12/12 2019;6(1):317. doi:10.1038/s41597-019-0322-0