Utilizing artificial intelligence and medical experts to identify predictors for common diagnoses in dyspneic adults: A cross-sectional study of consecutive emergency department patients from Southern Sweden

Tolestam Heyman, Ellen; Ashfaq, Awais; Ekelund, Ulf; Ohlsson, Mattias; Björk, Jonas; Schubert, Alexander Marcel; Lingman, Markus; Khoshnood, Ardavan M.

Published in:
International Journal of Medical Informatics

DOI:
10.1016/j.ijmedinf.2025.105969

2025

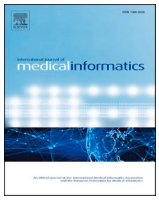
Document Version:
Publisher's PDF, also known as Version of record

International Journal of Medical Informatics 202 (2025) 105969

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Utilizing artificial intelligence and medical experts to identify predictors for common diagnoses in dyspneic adults: A cross-sectional study of consecutive emergency department patients from Southern Sweden

Ellen T. Heyman [a,b,*], Awais Ashfaq [c,d], Ulf Ekelund [b,e], Mattias Ohlsson [f,d], Jonas Björk [g,h], Alexander Marcel Schubert [i,j], Markus Lingman [c,d,k,1], Ardavan M. Khoshnood [l,m,1]

[a] Department of Emergency Medicine, Halland Hospital, Region Halland, Sweden
[b] Emergency Medicine, Department of Clinical Sciences Lund, Lund University, Lund, Sweden
[c] Halland Hospital, Region Halland, Sweden
[d] Center for Applied Intelligent Systems Research (CAISR), Halmstad University, Halmstad, Sweden
[e] Skåne University Hospital, Lund, Sweden
[f] Centre for Environmental and Climate Science, Lund University, Lund, Sweden
[g] Department of Laboratory Medicine, Division of Occupational and Environmental Medicine, Lund University, Lund, Sweden
[h] Clinical Studies Sweden, Forum South, Skåne University Hospital, Lund, Sweden
[i] Department of Computational Precision Health, University of California, Berkeley, USA
[j] Department of Computational Precision Health, University of California, San Francisco, USA
[k] Department of Molecular and Clinical Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
[l] Emergency Medicine, Department of Clinical Sciences Malmö, Lund University, Lund, Sweden
[m] Skåne University Hospital, Malmö, Sweden

ARTICLE INFO

Keywords:
Emergency medicine
Emergency department
Diagnostics
Artificial intelligence
Machine learning

ABSTRACT

Objective: Half of all adult emergency department (ED) visits with a complaint of dyspnea involve acute heart failure (AHF), exacerbation of chronic obstructive pulmonary disease (eCOPD), or pneumonia, which are often misdiagnosed. We aimed to create an artificial intelligence (AI) diagnostic decision support tool to detect patients with AHF, eCOPD, and pneumonia among dyspneic adults at the beginning of their ED visit.

Methods: In this cross-sectional study, we included all ED visits of patients 18 years or older with dyspnea at two regional Swedish EDs 07/01/2017–12/31/2019. In-hospital or ED discharge notes were used as outcome labels, with a subset manually reviewed by experts. We analyzed data from a complete regional healthcare system, along with socioeconomic factors, using Hierarchical Attention Networks. Each patient displayed a unique set of variables important for diagnosing dyspnea. All patients' unique variable sets were aggregated into a variable list. The top 100, 50, and 20 variables were tested in a simpler CatBoost model. Finally, performance was compared after adding medical expertise to the AI model.

Results: We included 10,869 visits, with 15.1% having AHF, 13.6% eCOPD, and 13.1% pneumonia. The median number of variables per unique ED visit was 187 (IQR 111–307). Aggregating the unique sets of variables resulted in a cohort list of 2,064 variables. The median micro AUROC was 87.8% (2.5–97.5 percentile; 86.4–89.3%). Age, ECGs, previous diagnoses, and medication were considered important by the AI model, while sex, vital signs, and socioeconomic factors were deemed almost non-predictive. Using the top 20 AI-selected variables, the AUROC was 86.6% (85.1–88.1%). Adding human medical expertise did not significantly change the AUROC.

Glossary: AHF, Acute Heart Failure; CDSS, Clinical Decision Support Systems; COPD, Chronic Obstructive Pulmonary Disease; ECG, Electrocardiogram; eCOPD, Exacerbation of Chronic Obstructive Pulmonary Disease; ED, Emergency Department; EHR, Electronic Health Record; GRU, Gated Recurrent Units; HAN, Hierarchical Attention Network; NLP, Natural Language Processing; RETTS, Rapid Emergency Triage and Treatment System.

* Corresponding author at: Department of Emergency Medicine, Halland Hospital, Region Halland, Sweden.
E-mail addresses: ellen.tolestam-heyman@regionhalland.se (E.T. Heyman), awais.ashfaq@regionhalland.se (A. Ashfaq), ulf.ekelund@med.lu.se (U. Ekelund), mattias.ohlsson@hh.se (M. Ohlsson), jonas.bjork@med.lu.se (J. Björk), alexander_schubert@berkeley.edu (A.M. Schubert), markus.lingman@regionhalland.se (M. Lingman), ardavan.khoshnood@med.lu.se (A.M. Khoshnood).

[1] Shared last authorship.

https://doi.org/10.1016/j.ijmedinf.2025.105969
Received 11 July 2024; Received in revised form 15 February 2025; Accepted 14 May 2025
Available online 22 May 2025

Conclusion: Based on the analysis of a high-dimensional dataset, we designed a lightweight 20-variable machine learning model that can early and effectively diagnose AHF, eCOPD, and pneumonia among ED patients with dyspnea.

1. Introduction

Dyspnea, a sensation of breathing difficulty, is a time-critical and high-mortality complaint in the emergency department (ED) [1–3]. Early treatment increases the likelihood of ED discharge to home, shortens ED and in-hospital stays, and reduces readmission and mortality [4–8], while erroneous treatment may increase morbidity and mortality [9–12].

About half of adult ED visits for dyspnea are due to acute heart failure (AHF), exacerbation of chronic obstructive pulmonary disease (eCOPD), or pneumonia, [13,14] conditions that are challenging to differentiate due to overlapping symptoms, risk factors, and triggers [15–18].

At ED discharge, after physical examination, blood tests, and other assessments, the ED discharge diagnosis is concordant with the subsequent in-hospital discharge diagnosis in only 50–70 % of AHF, eCOPD, and pneumonia cases [19–23]. Older dyspneic patients receive erroneous ED treatment in 36 %, 46 %, and 32 % of cases of AHF, eCOPD, and pneumonia, respectively [20].

To our knowledge, no published research—whether machine learning-based or otherwise—has aimed to simultaneously identify AHF, eCOPD, and pneumonia in dyspneic patients using text-based analysis. Previous machine learning studies in dyspneic ED patients have primarily focused on individual diagnoses:

For AHF, one study used L1-Regularized Logistic Regression on 31 expert-selected variables and unstructured text, achieving an area under the receiver operating characteristic curve (AUROC) of 99 %. Key predictors included the words "diuresis," "BNP," and "CHF" from text, alongside NT-proBNP lab results and inpatient diuretic use [24]. *Another study developed a hybrid AI-based clinical decision support system (CDSS) combining expert-designed decision trees and machine learning. The most predictive variables—three cardiac measurements—were extracted from ultrasonic heart exam free text reports. The CDSS performed comparably to cardiologists and outperformed non-cardiologists* [25].

For COPD, Gradient Boosting models predicted exacerbation risk with an AUROC of 83 %. SHAP values identified COPD Assessment Test scores and wheezing as top predictors among 34 expert-selected variables [26]. *Another study found that Gradient-Boosted Decision Trees and Logistic Regression outperformed pulmonologists, prioritizing different features among 31 variables. Key predictors included age, body mass index, and height in one model, versus blood oxygen saturation, cough, and sputum in another* [27].

Few machine learning studies have focused on pneumonia, with most analyzing X-rays [28] *or ultrasound* [29] *rather than clinical text. Early neural networks used 30–38 expert-selected textual features but lacked feature importance analysis* [30,31].

We aimed to develop an AI-based decision support tool to identify AHF, eCOPD, and pneumonia in dyspneic adults early in the ED visit. The AI would autonomously screen comprehensive, unselected regional healthcare data, enabling the creation of simplified diagnostic models with and without medical expertise.

2. Material and methods

2.1. Setting

This population-based cross-sectional study was conducted in Halland, a southwestern Swedish region with 330,000 inhabitants. Its two EDs receive 46,000 and 42,000 visits annually.

We included all ED visits from patients > 17 years with chief complaint dyspnea from July 1, 2017, to December 31, 2019 (Table 1). The five-level Rapid Emergency Triage and Treatment System (RETTS) [32] defined the complaint.

2.2. Diagnostic labels

Diagnostic labels were defined using World Health Organization (WHO) ICD codes, [33] based on primary diagnoses from ED or in-hospital discharge summaries. More than one label was allowed. A medical expert committee reviewed over 1,000 diagnoses in a prior study, finding diagnostic accuracy acceptable [34]. These expert labels were used in this study. See Appendix for details (Text A.1.).

2.3. Variables

All accessible data were used in an open-ended search for diagnostic variables. *These variables, such as diagnoses, are not isolated but linked to contextual (e.g., care setting) and temporal* (i.e., *number of weeks before the index visit) information, creating a high-dimensional latent space. For example, a heart failure diagnosis in primary care two weeks ago differs from one in emergency care ten weeks ago.* Patient data (ED care, ambulance service, inpatient care, outpatient specialist care, and primary care) within one year before the ED visit were compiled using the region's data analysis platform, [35] covering diagnoses, procedures, medications, blood tests, vitals, and electrocardiograms (ECGs) (Table 2). Socioeconomic factors and redeemed prescription drugs data were added from Statistics Sweden and the National Prescribed Drug Register by The National Board of Health and Welfare, respectively. Images and free text were excluded.

Data were prepared mainly by a rule-based approach. Categorical and ordinal values were treated as categorical. Continuous socioeconomic values were categorized by the 10, 25, 75, and 90 cohort percentiles, vital signs manually based on cohort distribution, and lab tests as normal, high, or low per reference values. Age was grouped in 5-year intervals. ECGs were included as 125 binary features per the manufacturer's algorithm [36]. All values were treated as unordered categorical "words;" those occurring less than 50 times in the complete dataset were excluded. Remaining words were timestamped.

2.4. Model design

We divided the year before the ED visit into ten equal time periods and ten clinical contexts: ambulatory care, ED care, in-hospital care, outpatient specialist care, primary care, human-derived factors, other factors, ECGs, socioeconomic factors, and triage variables at the index visit (Table 2). "Other factors" included variables not fitting specific contexts. Each context appeared in all ten periods, except triage

Table 1
Cohort definition proceeding from all regional ED visits.

Inclusion or Exclusion Criteria	Change (N)	Cohort Size (N)
ED visits registered upon arrival 7/1/ 2017–12/31/ 2019	N/A	221,264
ED visits after exclusion of patients < 18 years	−47,557	173,707
ED visits after exclusion of referrals from triage to other levels of care and exclusion of LWBS*	−16,583	157,124
ED visits after exclusion of visits with other complaints than dyspnea**	−146,255	10,869

*LWBS = patients who left without being seen by a doctor. **i.e., not assigned "dyspnea" according to the RETTS triage system.[32].

Table 2
Included variables in the model's ten contexts. In total, there were 11,656 words, i.e. categorized variables, some appearing in multiple contexts.

Context No.	Context Name	Data	Number of words
1	Primary care (included in all time periods)	Complaints Urgent/planned? Type of encounter (e.g., physical or digital) Care-provider category Procedures Primary and secondary diagnoses Referrals	1,665
2	Outpatient specialist care (included in all time periods)	Complaints Urgent/planned? Type of encounter (e.g., physical or digital) Organization/clinic Care-provider category Procedures Primary and secondary diagnoses Referrals	2,766
3	Emergency department care (included in all time periods)	Which hospital Ambulance/walk-in Complaints Triage priority Care provider category Medications Procedures Primary and secondary diagnoses Hospital admittance or discharge Referrals	1,217
4	Inpatient care (included in all time periods)	Admitted from Urgent/planned? Organization/ward Medications Procedures Primary and secondary diagnoses Discharged to Referrals	2,186
5	Ambulatory care (included in all time periods)	Ambulance priority Medications Oxygen delivery Free airway? Semi-sitting position? Continuous positive airway pressure (CPAP)? Advance notice to ED? Pain Time: acknowledgement of assignment, arrival to and departure from pickup place, completion of assignment	32
6	Other factors (included in all time periods)	Smoking status Vital signs measured in medical history in primary care, outpatient specialist care, ambulatory care, ED care, in-hospital care, but not in index ED visit: – Level of consciousness – Systolic and diastolic blood pressure – Pulse – Temperature – Oxygen saturation – Breathing frequency Ordinary medications, prescribed Ordinary medications, picked-up* Number of picked-up medication packages* Distribution of medication to patient* Blood samples and other laboratory tests Radiology exams, type of	681
7	Human-derived factors (included in all time periods)	Age Sex	18
8	Electrocardiograms (ECGs) (included in all time periods)	Presence or absence of 125 ECG features**	317
9	Socioeconomic factors (included only in the last time period)	Comparison of 216 socioeconomic variables***	2,694
10	Triage variables at index visit (included only in the last time period)	Time at ED registration (hour, day, week) Which hospital Ambulance/walk-in Number of concurrent ambulance assignments ED occupancy Triage priority Vital signs at index visit: – Level of consciousness – Systolic and diastolic blood pressure – Pulse – Temperature – Oxygen saturation – Breathing frequency	80

*Data source: The National Prescribed Drug Register by The Swedish National Board of Health and Welfare; ** Data source: The ECG equipment manufacturer's curve review algorithms [36]; *** Data source: The Longitudinal Integrated Database for Health Insurance and Labor Market Studies database by Statistics Sweden.

variables at the index visit and socioeconomic factors, which were only in the final period. This created ten periods times eight contexts plus two additional contexts in the final period, resulting in 82 categories to which every word of a patient visit was assigned. *This design reflects the hierarchical and contextual nature of clinical data, enabling the model to assess not only which clinical events are important but also when and where they occur.*

We previously developed the multivariate prediction model *Clinical Attention-Based Recurrent Encoder Network (CareNet)* [34], a Hierarchical Attention Network (HAN) inspired by Natural Language Processing (NLP) [37]. CareNet has three hierarchical layers: from bottom to top event, context, and time period. This hierarchical structure mirrors a clinician's consideration of time and context in evaluating clinical events. Each layer includes an encoder and an attention block. The encoder, using bidirectional gated recurrent units (GRUs) [38], integrates neighborhood information and converts inputs into numerical embeddings via pretrained skip-gram initialization [39]. The attention block learns, weighs, and aggregates input embeddings to form higher-level representations.

Thus, the event, context, and time period attention blocks calculate the importance of their respective embeddings. The final patient visit embedding passes through a feed-forward neural network, generating diagnosis label distributions, and evidential loss is calculated using labels. CareNet is trained end-to-end by minimizing evidential loss via backpropagation.

Missing data were handled using "N/A" markers, allowing the model to autonomously deduce missing values as part of its training and to identify existing values for the specific patient visit. Words with equal or less importance than "N/A" within the same time period and context

were considered as noise and excluded in that specific period and clinical context. Thus, all patient visits received their own unique and individual variable set selected from vast possible words based on their medical history.

Further details on CareNet design are in Text A.2 in Appendix.

2.5. Experiments and evaluation

For AUROC analyses, we performed 10-fold cross-validation and 10 bootstrapped evaluations within each fold using 90 % of the evaluation set, yielding a 10x10 matrix of AUROC values reflecting both cross-validation and bootstrapping techniques. Median micro AUROC (2.5–97.5 percentile) was calculated on the evaluation fold, giving each patient visit the same importance. Our multilabel design allowed a probability between 0 and 1 for each diagnosis label, potentially exceeding 100 % in total.

2.6. Variable list and simpler model

A slight improvement in machine learning performance may not justify reduced transparency and increased real-time data needs in the noisy ED setting. Therefore, we aimed to develop a simpler, more feasible model for ED implementation.

To identify the top variables in the CareNet model, we leveraged its hierarchical structure, which integrates clinical events, context, and timing. For each patient, we calculated a score by multiplying event weights with corresponding context and time weights, reflecting overall importance. This process was repeated in each fold of a 10-fold cross-validation, producing ten variable lists. The final rankings were determined by averaging these lists.

As a post-hoc analysis, four of the authors (ML and UE, consultant physicians in cardiology; ETH and AK, resident and consultant physicians in emergency medicine) selected medically relevant, understandable, and feasible variables from the top 300 variables on the list, choosing those approved by at least three out of the four authors. We analyzed AUROC by including the top 100, 50, and 20 variables, respectively, with and without expert selection using a CatBoost model [40]. *The CatBoost model might be considered simpler as it treats input variables as atomic and independent, disregarding the contextual and temporal dependencies learned by CareNet through its hierarchical attention-based approach.* CatBoost AUROC was assessed in the same manner as CareNet AUROC, utilizing a 10x10 matrix described above, and reported as median micro AUROC (2.5–97.5 percentile). *We report CatBoost results due to its built-in feature importance analysis, proven effectiveness, ease of use, and popularity among clinical researchers* [41]. *An analysis with other machine learning models is presented in the* Appendix *(see* Table A.1*).*

Sensitivity and specificity were defined as the maximum sensitivity with a specificity above 75.0 % and its corresponding specificity. Sensitivity and specificity for each diagnosis label of the CareNet and CatBoost models were reported as medians (2.5–97.5 percentile) calculated using 10x10 matrices of AUROC values after cross-validation and bootstrapping techniques.

The analysis was performed using PyTorch [42] software.

2.7. Descriptive statistics

For descriptive statistics, percentages as well as median and interquartile range (IQR), SPSS [43] and MS Excel [44] software were used.

3. Results

3.1. Descriptive statistics

The study included 10,869 visits by 7,457 unique patients. AHF, eCOPD, and pneumonia were identified in 15.1 %, 13.6 %, and 13.1 % of visits, respectively. A total of 97 patient visits (0.9 % of the cohort) had more than one label. The category labeled “other diagnoses” included pulmonary embolism (3.4 %) and atrial fibrillation/flutter (1.8 %) alongside symptom-related diagnoses and various, less prevalent, conditions (Table A.2 in Appendix). The cohort́s median age was 75 years (IQR 61–83) (Table 3).

3.2. Diagnostic performance

CareNet achieved a performance of 87.8 % (86.4–89.3 %) in the overall cohort (Table 4). Performance was notably higher for younger patients and those with fewer of the three diagnoses in their medical history, while it was similar between male and female patients (Table A.3 in Appendix).

The median number of unique individual variables per ED visit, selected from a total of 11,656 words, was 187 (IQR 111–307). The individual word sets were compiled into an AI-generated list of 2,064 words for the cohort. The highest-ranking variables included prior diagnoses of COPD and heart failure, advanced age, and atrial fibrillation as recorded on ECGs. Age, ECG findings, previous diagnoses, and medication usage were consistently prioritized by the model, whereas socioeconomic factors, vital signs, and sex were considered less influential (Table A.4 in Appendix).

Four medical experts collectively selected 117 variables from the top 300 variables identified by CareNet (Table A.4 in Appendix). Subsequently, two lists were generated: the original CareNet list and a refined list from which non-selected variables were excluded. The top 100, 50, and 20 variables from both lists exhibited overlaps of 43, 34, and 16 variables, respectively.

Evaluation of these variable sets using a simplified CatBoost model indicated minimal variation in median micro AUROC when adjusting the number of variables or incorporating expert selection (Table 4). *The five alternative machine learning models—Random Forest, XGBoost, Logistic Regression, LightGBM, and Extra Trees—showed comparable performance to the CatBoost model, though no formal statistical tests were conducted (*Table A.1 *in the* Appendix*).*

Micro AUROC is shown for the CareNet model in Fig. 1 and for the CatBoost model, including 20 unselected variables, in Fig. 2. Across all models, eCOPD exhibited the highest performance and pneumonia the lowest.

3.3. Sensitivity and specificity

CareNet demonstrated a median sensitivity, with specificity set above 75.0 %, of 78.5 % (69.0–86.2 %) for AHF, 91.9 % (84.5–96.4 %) for eCOPD, 39.6 % (18.0–66.7 %) for pneumonia, and 67.8 % (58.9–73.1 %) for “other diagnoses” (Table 5). The CatBoost model using the top 20 AI-ranked CareNet variables displayed a median sensitivity of 76.4 % (67.6–84.8 %), 87.1 % (79.2–93.0 %), 32.8 % (20.9–44.5 %), and 64.8 % (49.2–70.5 %) for AHF, eCOPD, pneumonia, and “other diagnoses,” respectively (Table 5). Increasing the number of variables to 50 or 100 and incorporating medical expert selection only slightly improved sensitivity while maintaining specificity above 75 % for all diagnoses, particularly noticeable in pneumonia.

4. Discussion

4.1. General discussion

In this cross-sectional population-wide study, we utilized AI to screen vast unselected data from a complete regional healthcare system. We identified the most predictive variables for a simplified model to diagnose AHF, eCOPD, and pneumonia in dyspneic adults at ED triage. Our CareNet model achieved a micro AUROC of 87.8 % (86.4–89.3 %) after ED triage, before physician assessment, blood tests, or x-rays. It identified unique variables for each patient, resulting in a list of 2,064 diverse variables arranged by diagnostic relevance. The top 20 variables from this list were incorporated into a simpler model with an AUROC of

Table 3
Cohort characteristics of visits labeled AHF, eCOPD, pneumonia, and "other diagnoses".

	All Cohort, *N (%) or median (IQR)*	AHF, N (%) or median (IQR)	eCOPD, *N (%) or median (IQR)*	Pneumonia, *N (%) or median (IQR)*	Other Diagnoses, *N (%) or median (IQR)*
*Visits, N (%)**	10,869 (100)	1,640 (15.1)	1,481 (13.6)	1,420 (13.1)	6,425 (59.1)
Unique patients, N	7,457	1,242	800	1,263	5,180
Age, *median (IQR)*	75.0 (61.0–83.0)	83.0 (76.0–89.0)	76.0 (69.0–82.0)	77.0 (66.0–85.3)	71.0 (51.0–81.0)
Sex, N (%)					
Male	5,194 (47.8)	886 (54.0)	650 (43.9)	692 (48.7)	3,011 (46.9)
Female	5,675 (52.2)	754 (46.0)	831 (56.1)	728 (51.3)	3,414 (53.1)
Medical history					
Charlson Comorbidity Index, median (IQR)	1.0 (0.0–2.0)	2.0 (0.0–3.0)	1.0 (1.0–2.0)	0.0 (0.0–2.0)	0.0 (0.0–2.0)
*Heart failure diagnosis previous year**, N (%)*	2,319 (21.3)	881 (53.7)	411 (27.8)	234 (16.5)	824 (12.8)
*COPD diagnosis previous year**, N (%)*	2,366 (21.8)	263 (16.0)	1,194 (80.6)	291 (20.5)	678 (10.6)
*Pneumonia diagnosis previous month**, N (%)*	437 (4.02)	35 (2.13)	48 (3.24)	153 (10.8)	205 (3.19)
*No. of primary care encounters previous year***, median (IQR)*	11 (4.00–21.0)	17 (9.00–29.0)	13 (6.00–24.0)	10 (4.00–19.0)	9 (3.00–19.0)
*No. of outpatient specialist encounters previous year***, median (IQR)*	3.0 (0.0–8.0)	4.0 (1.0–9.0)	3.0 (1.0–8.0)	2.0 (0.0–7.0)	2.0 (0.0–8.0)
No. of emergency department visits previous year, median (IQR)	1.0 (0.0–2.0)	1.0 (0.0–3.0)	2.0 (0.0–4.0)	1.0 (0.0–2.0)	1.0 (0.0–2.0)
No. of in-hospital visits last year, median (IQR)	1.0 (0.0–2.0)	1.0 (0.0–3.0)	1.0 (0.0–3.0)	1.0 (0.0–2.0)	0.0 (0.0–1.0)
Index visit					
*Hospital****, N (%)*					
Hospital 1	5,691 (52.4)	928 (56.6)	818 (55.2)	651 (45.8)	3,345 (52.1)
Hospital 2	5,151 (47.4)	711 (43.4)	661 (44.6)	764 (53.8)	3,060 (47.6)
Arrival time to emergency department, N (%)					
Monday-Friday, 8:00 am-8:59 pm	6,578 (60.5)	1,093 (66.6)	828 (55.9)	833 (58.7)	3,897 (60.7)
Saturday-Sunday, 8:00 am-8:59 pm	2,050 (18.9)	309 (18.8)	282 (19.0)	292 (20.6)	1,183 (18.4)
Nighttime, 9:00 pm-7:59 am	2,241 (20.6)	238 (14.5)	371 (25.1)	295 (20.8)	1,345 (20.9)
Ambulance arrival, N (%)	5,269 (48.5)	905 (55.2)	1,028 (69.4)	841 (59.2)	2,547 (39.6)
Triage priority, N (%)					
Priority 1	614 (5.65)	99 (6.04)	108 (7.29)	140 (9.86)	276 (4.30)
Priority 2	5,221 (48.0)	937 (57.1)	955 (64.5)	816 (57.5)	2,582 (40.2)
Priority 3	4,214 (38.8)	587 (35.8)	398 (26.9)	421 (29.6)	2,827 (44.0)
Priority 4	759 (6.98)	14 (0.854)	17 (1.15)	41 (2.89)	687 (10.7)
Priority 5	57 (0.524)	3 (0.183)	3 (0.203)	2 (0.141)	49 (0.763)
Socioeconomic factors					
*Civil status*****, N (%)*					
Unmarried	1,917 (17.6)	146 (8.90)	143 (9.66)	212 (14.9)	1,425 (22.2)
Married/registered partnership	4,633 (42.6)	686 (41.8)	581 (39.2)	622 (43.8)	2,781 (43.3)
Divorced	1,982 (18.2)	258 (15.7)	414 (28.0)	242 (17.0)	1,093 (17.0)
Widow/widower	2,251 (20.7)	546 (33.3)	338 (22.8)	337 (23.7)	1,055 (16.4)
*Household disposable income per consumption unit (100 Swedish kronor/year)*****, median (IQR)*	1,867 (1,519–2,625)	1,743 (1,506–2,625)	1,740 (1,507–2,625)	1,796 (1,503–2,625)	1,981 (1,534–2,791)
*Education******, N (%)*					
Primary school	4,084 (37.6)	800 (48.8)	643 (43.4)	557 (39.2)	2,125 (33.1)
Upper secondary school	4,521 (41.6)	576 (35.1)	630 (42.5)	571 (40.2)	2,786 (43.4)
Post-secondary education	2,019 (18.6)	245 (14.9)	185 (12.5)	262 (18.5)	1,339 (20.8)

Heart failure: ICD-10 code I11.0, I13.0, I13.2 or I50; Chronic obstructive pulmonary disease (COPD): ICD-10 code J44; Pneumonia: ICD-10 code J10.0, J11.0 or J12-J18; Other diagnoses: all other ICD-10 codes. * N = 97 (0.9 %) of the cohort has two labels; ** Registered anywhere in the regional health care system; *** Encounters included visits, digital meetings and phone calls; **** Missing value in all cohort: N = 27 (0.25 %); ***** Missing value in all cohort: N = 86 (0.79 %); ****** Missing value in all cohort: N = 245 (2.25 %).

Table 4
Comparison of CatBoost diagnostic performance with 100, 50, and 20 variables from the top of the CareNet List with and without selection of plausible variables by medical experts.

Model	Median Micro AUROC (%, 2.5–97.5 percentile)
CareNet, all data	87.8 (86.4–89.3)
CatBoost, 20 variables, unselected	86.6 (85.1–88.1)
CatBoost, 20 variables, selected by medical experts	86.7 (85.1–88.2)
CatBoost, 50 variables, unselected	87.0 (85.2–88.1)
CatBoost, 50 variables, selected by medical experts	87.2 (85.7–88.7)
CatBoost, 100 variables, unselected	87.4 (85.8–88.6)
CatBoost, 100 variables, selected by medical experts	88.7 (87.5–89.8)

86.6 % (85.1–88.1 %). Adding medical expertise had no significant impact.

CareNet selected unique variables for each patient, with a median of 187 (IQR 111–307) per visit from 11,656 possible words. This selection process allowed the model to autonomously recognize missing values during training and exclude low-weighted variables as noise, with variable noise thresholds adapted to different contexts and time periods. Aggregating these patient-specific "fingerprints" yielded 2,064 key variables, reflecting significant variables across the cohort. Variables may carry higher weight due to critical importance to specific individuals or broader relevance across patient groups. *While the attention mechanism in our CareNet model offers a useful preliminary measure of feature importance for the model's internal evaluation of its latent variables, it is important to recognize that it may not fully capture semantic significance. As noted by others,* [45,46] *attention mechanisms do not always serve as definitive indicators of feature importance, necessitating cautious interpretation. However, they can provide valuable initial assessments, though they do not enable causal inference on these inputs* [37,47]. CareNet effectively analyzes complex clinical relationships in real clinical scenarios, integrating data within the same context and timeframe, surpassing standard feature selection methods.

CareNet's highest-ranked variables (Table A.4 in Appendix) seem medically quite reasonable: prior underlying, chronic conditions of COPD and heart failure, advanced age, and atrial fibrillation recorded on

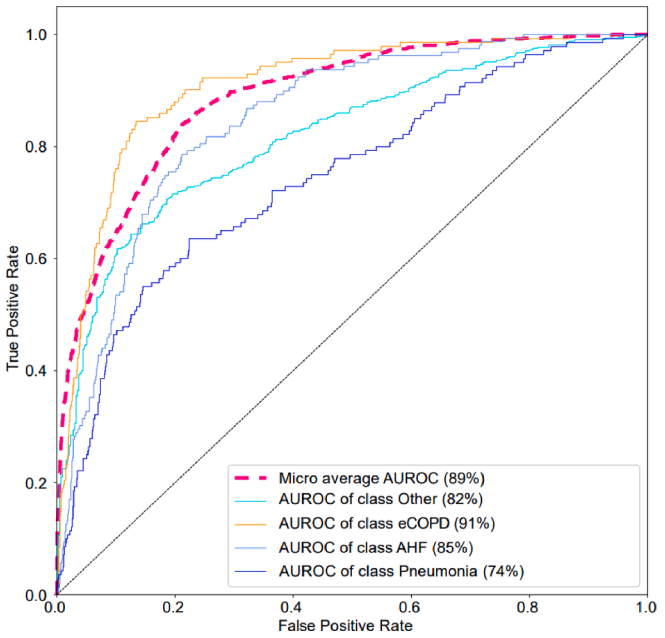


Fig. 1. CareNet Performance. CareNet micro AUROC using one year of data prior to index visit. An illustrative example from one of the validation folds with the highest micro AUROC. AHF = acute heart failure, eCOPD = exacerbation of chronic obstructive pulmonary disease.

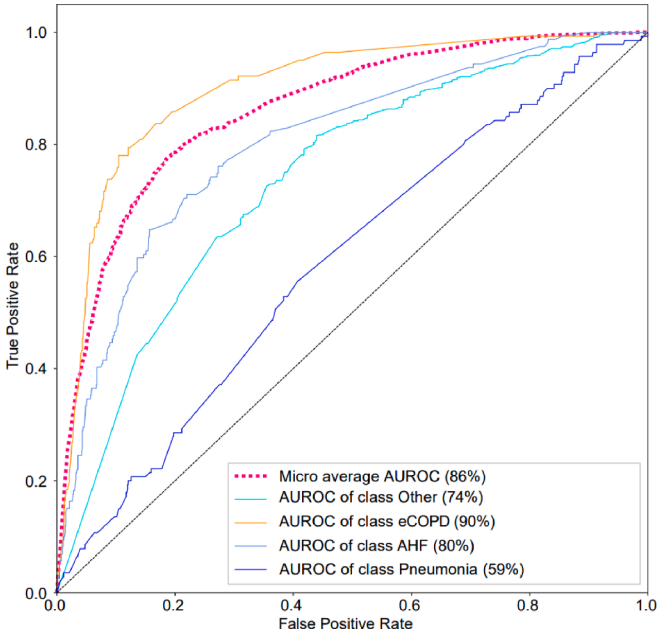


Fig. 2. CatBoost Performance. CatBoost micro AUROC models using the top 20 variables, not selected by medical experts. An illustrative example from one of the validation folds with the highest micro AUROC. AHF = acute heart failure, eCOPD = exacerbation of chronic obstructive pulmonary disease.

an ECG. In fact, 16 of the top 20 variables were clearly medically plausible according to the medical experts (Table A.4 in Appendix), supporting its validity.

Among 2,064 variables, prior diagnoses, medication, age, and ECGs ranked highest, while sex, vital signs, and socioeconomic factors ranked lower. Top ECG variables—atrial fibrillation, ventricular-paced complexes, left ventricular hypertrophy, and bundle-branch block—are expected, as atrial fibrillation is the most common condition in the "other diagnoses" label (Table A.2, Appendix), and because patients with AHF, eCOPD, and pneumonia have altered likelihoods of most of these findings on their ECG [48–50]. The low ranking of socioeconomic factors, despite prior findings, [51–53] may reflect the cohort's advanced age and multimorbidity, making factors like profession less relevant. Income differences were also negligible among diagnostic groups (Table 3). Diagnostic precision was lower in elderly patients and those with all three diagnoses in their history (Table A.3 in Appendix), consistent with routine care. [20].

To our knowledge, no clinical score or published study identifies the most effective variables for classifying adult patients with dyspnea into four categories simultaneously: AHF, eCOPD, pneumonia, and all other diagnoses. Few studies use machine learning to classify a single diagnosis against all others, which is a different approach. These studies also rely on a limited number of expert-selected variables, [24–27,30,31] *free-text extractions,* [24] *and data that are not typically found in structured form in routine healthcare records, but have to be manually curated* [25,26]. *Meanwhile, other studies do not present any assessment of variable importance* [30,31]. *This hinders direct comparisons, as our study incorporates a comprehensive set of structured data but excludes free text. Additionally, different models analyzing the same dataset may assign vastly different importance to variables* [27].

We tested the top 100, 50, and 20 CareNet variables in a simpler CatBoost model. A simpler model, less variables, and medical expertise had little impact on micro AUROC (Table 4). This suggests a few strong predictors in this research question, ranked highest by CareNet, and validated by experts, drive performance. However, sensitivity, with a specificity set above 75 %, altered in pneumonia (Table 5), likely because pneumonia does not necessarily presume an underlying chronic disease, and is therefore more elusive.

4.2. Strengths and limitations

This study's strength is its basis in a complete regional population and an integrated healthcare system, including all emergency care. Data were analyzed without variable selection and with minimal manual modification to reduce bias. Though, it does not adjust for biases from care consumption and clinical work procedures.

Another strength is the inclusion of both admitted and discharged patients, despite less reliable ED discharge diagnoses [54]. To address this, expert labels were assigned to over 1,000 reviewed visits [34]. Prior research showed no performance difference in diagnosing AHF, eCOPD, and pneumonia with or without expert labels [34].

Only structured data, not free text or images, were included, and continuous values were categorized, while ordinal values were treated as categorical, losing some information. *Furthermore, our hierarchical data structure uses contextual embeddings, meaning the model interprets latent variables rather than isolated features. These variables incorporate context (e.g., location) and time, limiting the direct use of standard feature importance methods designed for independent variables.*

4.3. Future implications

CareNet's generic design allows application to any diagnosis. Adding unstructured data, like physician notes, should be feasible given its NLP properties. Future research needs to explore advanced interpretability methods for hierarchical and latent variables. *Future studies should investigate more robust techniques to analyze the model's black-box behavior, requiring interpretability methods specialized for hierarchical and latent variables.*

Code availability statement

The code for this study is available on GitHub and can be accessed via: https://github.com/aaq109/Carenet_v2.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used OpenAI's ChatGPT4 for English language review. After using this service, the

Table 5
Sensitivity and specificity for the CareNet model and six CatBoost models. Values reported are maximum sensitivity with a specificity above 75.0% and its corresponding specificity. The variables may be selected or unselected by the authors as medically feasible.

		AHF	eCOPD	Pneumonia	Other
CareNet	Median sensitivity (%, 2.5–97.5 percentile)	78.5 (69.0–86.2)	91.9 (84.5–96.4)	39.6 (18.0–66.7)	67.8 (58.9–73.1)
	Median specificity (%, 2.5–97.5 percentile)	75.3 (75.0–76.0)	75.4 (75.0–76.2)	76.3 (75.1–81.9)	77.6 (75.1–82.9)
CatBoost					
20 variables, unselected	Median sensitivity (%, 2.5–97.5 percentile)	76.4 (67.6–84.8)	87.1 (79.2–93.0)	32.8 (20.9–44.5)	64.8 (49.2–70.5)
	Median specificity (%, 2.5–97.5 percentile)	75.3 (75.0–78.3)	80.9 (75.1–84.6)	76.7 (75.1–81.4)	75.7 (75.0–80.8)
20 variables, selected	Median sensitivity (%, 2.5–97.5 percentile)	75.9 (65.7–83.3)	90.1 (82.6–94.1)	34.3 (24.3–44.6)	65.0 (56.9–70.9)
	Median specificity (%, 2.5–97.5 percentile)	75.3 (75.0–78.9)	77.5 (75.1–80.9)	75.5 (75.0–79.0)	75.5 (75.0–78.8)
50 variables, unselected	Median sensitivity (%, 2.5–97.5 percentile)	76.1 (66.0–83.7)	90.6 (83.5–96.5)	35.0 (24.2–49.4)	64.3 (56.8–72.9)
	Median specificity (%, 2.5–97.5 percentile)	75.2 (75.0–76.8)	75.3 (75.0–77.4)	75.4 (75.0–78.5)	75.4 (75.0–77.8)
50 variables, selected	Median sensitivity (%, 2.5–97.5 percentile)	75.9 (68.5–84.2)	92.7 (85.4–97.8)	36.9 (25.4–47.3)	65.1 (58.3–75.3)
	Median specificity (%, 2.5–97.5 percentile)	75.2 (75.0–76.1)	75.3 (75.0–78.4)	75.4 (75.0–77.8)	75.3 (75.0–76.5)
100 variables, unselected	Median sensitivity (%, 2.5–97.5 percentile)	77.6 (68.0–87.0)	91.4 (84.0–96.7)	40.5 (27.2–52.5)	66.2 (59.5–72.5)
	Median specificity (%, 2.5–97.5 percentile)	75.2 (75.0–75.6)	75.2 (75.0–77.0)	75.3 (75.0–78.1)	75.2 (75.0–76.2)
100 variables, selected	Median sensitivity (%, 2.5–97.5 percentile)	81.4 (71.7–89.2)	92.2 (84.8–97.3)	55.9 (44.6–67.2)	68.6 (61.4–78.1)
	Median specificity (%, 2.5–97.5 percentile)	75.1 (75.0–75.5)	75.2 (75.0–76.0)	75.2 (75.0–75.9)	75.2 (75.0–75.8)

authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Summary table

What was already known on the topic

- Dyspnea, or breathing difficulty, is a high-risk emergency department complaint, often misdiagnosed.
- AI has been used to predict individual diagnoses but usually relies on fewer expert-selected variables.

What this study added to our knowledge

- We developed an AI tool for early emergency department diagnosis of acute heart failure, COPD exacerbation, and pneumonia.
- We screened 12,000 unselected, categorized variables from a complete healthcare system.
- A hierarchical attention model autonomously selected top predictive variables for each patient visit.
- Aggregating individual top variables enabled a simpler, high-accuracy model; medical expertise added little.

Funding

This study was part of the AIR Lund (Artificially Intelligent use of Registers at Lund University) research network in Lund, Sweden. The work was funded by the Swedish Research Council [Grant no. 2019–00198]; the Scientific Council of Region Halland, Sweden [Grant no. 979314]; Sparbanksstiftelsen Varberg, Sweden [Grant no. 980763]; and the foundation Stiftelsen Landshövding Per Westlings Minnesfond, Sweden [Grant no. RMh2020-0007]. The funders had no role in the study design, data collection, analysis, interpretation, writing of the report, or the decision to submit the article for publication.

Ethical and legal statement

This study was performed in compliance with relevant laws and institutional guidelines. The privacy rights of human subjects were observed. The study was approved by the Swedish Ethical Review Authority (no. 2021–02520 on 07–12-2021 and 2021–05989-02 on 12–06-2021). Informed consent was waived, and the participants were instead given an opt-out possibility in accordance with the ethical approval.

CRediT authorship contribution statement

Ellen T. Heyman: Writing – original draft, Investigation, Project administration. **Awais Ashfaq:** Writing – review & editing, Methodology, Formal analysis, Data curation. **Ulf Ekelund:** Writing – review & editing, Supervision, Methodology. **Mattias Ohlsson:** Writing – review & editing, Supervision, Methodology. **Jonas Björk:** Writing – review & editing, Supervision, Methodology. **Alexander Marcel Schubert:** Writing – review & editing, Supervision, Methodology. **Markus Lingman:** Writing – review & editing, Supervision, Methodology. **Ardavan M. Khoshnood:** Writing – review & editing, Supervision, Methodology.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ellen Tolestam Heyman reports financial support was provided by Swedish Research Council. Ellen Tolestam Heyman reports financial support was provided by Region Halland. Ellen Tolestam Heyman reports financial support was provided by Sparbanksstiftelsen Varberg. Ellen Tolestam Heyman reports financial support was provided by Stiftelsen Landshövding Per Westlings minnesfond. Markus Lingman reports a relationship with Tandem Health Inc. that includes: consulting or advisory. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

Text A.1. Diagnostic Labels.

The study's four diagnostic labels were defined in accordance with WHO diagnostic ICD-10 codes: [33] AHF was represented by I11.0, I13.0, I13.2, and I50; eCOPD by J44; and pneumonia by J10.0, J11.0, or J12-J18. "Other diagnoses" was defined by the remaining ICD-10 codes. The primary diagnosis listed in the hospital discharge summary labeled patients admitted to the hospital, whereas the main diagnosis at ED discharge labeled patients who were not admitted. More than one label was allowed, which applied to a few visits with two main diagnoses documented in their electronic health record (EHR).

Our study cohort was manually reviewed for diagnostic accuracy in a previously published study, [34] including 95 % of the present study's cohort. The review identified ED visits discharged to home from the ED with the non-specific diagnosis "R06.0 dyspnea" as high risk of misdiagnosis. An expert committee of emergency medicine physicians manually reviewed the diagnoses in these 1,070 visits with access to data up to 30 days after the ED visit. Diagnostic inaccuracy was 4.5 % in AHF, 6.6 % in eCOPD, and 1.9 % in pneumonia cases, which we considered acceptable. We saw no

difference in model performance with and without expert labels in the previous study. [34] These expert labels are used throughout the present study.

Text A.2. Mathematical Background of the Clinical Attention-Based Recurrent Encoder Network (CareNet) Design.

The mathematical background was originally published in a previous paper by the same authors. [34].

1. Problem Setup

In an electronic health record, the journey of a patient's care during a single visit involves gathering various details specific to the patient, reported at different times and from different clinical contexts across the healthcare system. For example, a diagnosis might be made during primary care, outpatient specialist care, ED visits, or during inpatient care. The aim was to compile all this clinical information along with its respective timeframes and contexts and then represent it numerically at what we call the index time. This index time is crucial for categorizing patients in the ED into different diagnosis labels. Table 2 provides a comprehensive list of the types of data derived from various clinical contexts.

To capture the patient's health status at index time, we employed attention modules inspired by Natural Language Processing (NLP) [37]. Our approach involved breaking down the patient's visit trajectory over the past year into ten equal time periods, akin to paragraphs. Within each period, we organized data from ten clinical contexts resembling sentences originating from five different healthcare settings: primary care, outpatient specialist care, emergency care, ambulance care, inpatient care, as well as additional contexts such as ECG data, human-derived factors, socioeconomic factors, triage data, and "other factors" not easily connected to any specific clinical context due to the healthcare organization or IT-system structure. Notably, socioeconomic factors and triage data were only included in the most recent time segment.

Each of these contexts consisted of clinical events akin to words documented within its specific clinical context and timeframe. Ultimately, the ordered sequence of these time periods comprised the patient's health state or document for that visit.

2. Generating patient visit representation

In the context of NLP, we developed a three-layer hierarchical document (patient visit) embedding module consisting of word-level (clinical events), sentence-level (clinical contexts), and paragraph-level (time periods) embeddings and attention modules.

Let us consider a single patient visit p with M periods, where each period consists of K contexts (ten in our case) and each context consists of N clinical events. Vectors and matrices are represented in lowercase and uppercase bold respectively. $\boldsymbol{p}$ is the vector representation of patient visit p.$\boldsymbol{t}_m$ is the vector representation of period m of patient visit p.$\boldsymbol{s}_{mk}$ is the vector representation of context k within period m of patient visit p.$\boldsymbol{c}_{mkn}$ is the n^{th} code vector from context k in period m of patient visit p. For simplicity, we ignore the suffixes hereafter.

For word (or event) level embedding and attention, we first embed all the N clinical events from a care context s and period m of patient visit p into a vector space via a pretrained skip-gram initialization [38]. These event vectors $\boldsymbol{c}$ are then passed through bidirectional Gated Recurrent Units (GRUs) to obtain an intermediate representation $\boldsymbol{h}^{\boldsymbol{c}}$ for each event, such that it also incorporates the contextual information of other events in s. Since not every event contributes equally to the overall context representation, we apply Bahdanau attention using a single-layer NN with weights $\boldsymbol{W}^{\boldsymbol{c}}$ and bias $\boldsymbol{b}^{\boldsymbol{c}}$ as follows to obtain event-level attention $\boldsymbol{a}_{\boldsymbol{c}}$ and context representation $\boldsymbol{s}$:

$$\boldsymbol{u}_{\boldsymbol{c}} = \tanh(\boldsymbol{W}^{\boldsymbol{c}}\boldsymbol{h}^{\boldsymbol{c}} + \boldsymbol{b}^{\boldsymbol{c}}); \boldsymbol{a}_{\boldsymbol{c}} = \frac{\exp(\boldsymbol{u}_{\boldsymbol{c}}^T \boldsymbol{u}^{c'})}{\sum_{n=1}^{N} \exp(\boldsymbol{u}_{\boldsymbol{c}}^T \boldsymbol{u}^{c'})}; \boldsymbol{s} = \sum_{n=1}^{N} \boldsymbol{a}_c \boldsymbol{h}^c$$

where $\boldsymbol{u}^{c'}$ is the event-level self-attention vector which is randomly initialized and jointly learned during training.

For context-level attention, we follow similar steps. We stack another bidirectional GRU and a single-layer attention NN to generate care context-level attention $\boldsymbol{a}_{\boldsymbol{s}}$ and time period representation $\boldsymbol{t}$ as follows:

$$\boldsymbol{u}_{\boldsymbol{s}} = \tanh(\boldsymbol{W}^{\boldsymbol{s}}\boldsymbol{h}^{\boldsymbol{s}} + \boldsymbol{b}^{\boldsymbol{s}}); \boldsymbol{a}_{\boldsymbol{s}} = \frac{\exp(\boldsymbol{u}_{\boldsymbol{s}}^T \boldsymbol{u}^{s'})}{\sum_{k=1}^{K} \exp(\boldsymbol{u}_{\boldsymbol{s}}^T \boldsymbol{u}^{s'})}; \boldsymbol{t} = \sum_{k=1}^{K} \boldsymbol{a}_s \boldsymbol{h}^s$$

where $\boldsymbol{h}^{\boldsymbol{s}}$ is the intermediate context representation (after passing $\boldsymbol{s}$ through a bidirectional GRU), $\boldsymbol{u}^{s'}$ is the context-level self-attention vector, and $\boldsymbol{W}^{\boldsymbol{s}}$ and $\boldsymbol{b}^{\boldsymbol{s}}$ are context-level NN weights and bias respectively.

Next, for time period level attention, we repeat the process. We stack another bidirectional GRU and a single-layer attention NN to generate time-level attention $\boldsymbol{a}_{\boldsymbol{t}}$ and patient visit representation $\boldsymbol{p}$ as follows:

$$\boldsymbol{u}_{\boldsymbol{t}} = \tanh(\boldsymbol{W}^{\boldsymbol{t}}\boldsymbol{h}^{\boldsymbol{t}} + \boldsymbol{b}^{\boldsymbol{t}}); \boldsymbol{a}_{\boldsymbol{t}} = \frac{\exp(\boldsymbol{u}_{\boldsymbol{t}}^T \boldsymbol{u}^{t'})}{\sum_{m=1}^{M} \exp(\boldsymbol{u}_{\boldsymbol{t}}^T \boldsymbol{u}^{t'})}; \boldsymbol{p} = \sum_{m=1}^{M} \boldsymbol{a}_t \boldsymbol{h}^t$$

where $\boldsymbol{h}^{\boldsymbol{t}}$ is the intermediate time period and context representation (after passing $\boldsymbol{t}$ through a bidirectional GRU), $\boldsymbol{u}^{t'}$ is the time period level self-attention vector, and $\boldsymbol{W}^{\boldsymbol{t}}$ and $\boldsymbol{b}^{\boldsymbol{t}}$ are time-level NN weights and bias, respectively.

3. Loss computation and training

Given our patient visit representation $\boldsymbol{p}$, we finally map it to a 4-node output layer. We then calculate the multilabel evidential loss proposed in [55] using ground truth labels for training all the NN weights.

Hyperparameter tuning for Carenet was systematically conducted to evaluate the impact of various parameters on model performance. The embedding sizes were varied from 100 to 300 in increments of 50, drop out from 0.1 to 0.5 with increments of 0.1 and batch sizes of 32, 64, and 128 were tested. A learning rate decay strategy was employed (using pytorch optimizer class), where the learning rate was progressively reduced with each training step. Additionally, slight modifications were made to code-level RNN, context-level RNN, and time period-level RNN architectures. The results, assessed through cross-validation, indicated that the default architecture (reasonable accepted ranges) is robust, as variations in neural network sizes and layers did not significantly affect the performance metrics.

4. Experiments and evaluation

In each experiment, we employed 10-fold cross-validation with minimal hyperparameter tuning. Of note, we are aware of the potential minor impact stemming from the absence of a separate hold-out set, which could influence the reported results. Further validation through external datasets could enhance the robustness of our findings.

Table A1
Analyses Using Alternative Models. Analysis using Random Forest, XGBoost, Logistic Regression, LightGBM, and Extra Trees on the top 20, top 50, and top 100 variables, with and without expert selection.

Model	Top 20 SelectedMean AUROC (SD) (%)	Top 20 UnselectedMean AUROC (SD) (%)	Top 50 SelectedMean AUROC (SD) (%)	Top 50 UnselectedMean AUROC (SD) (%)	Top 100 SelectedMean AUROC (SD) (%)	Top 100 UnselectedMean AUROC (SD) (%)
Random Forest	85 (1.0)	85 (1.1)	86 (0.9)	86 (1.0)	88 (0.8)	86 (1.0)
XGBoost	86 (1.0)	86 (1.1)	86 (0.9)	86 (1.0)	88 (0.9)	87 (0.8)
Logistic Regression	86 (0.9)	86 (0.9)	87 (0.8)	86 (0.9)	88 (0.7)	87 (0.8)
LightGBM	87 (0.9)	86 (1.0)	87 (0.9)	87 (1.0)	88 (0.8)	87 (0.8)
Extra Trees	83 (1.1)	84 (1.2)	83 (1.1)	83 (1.3)	87 (1.0)	83 (1.1)

Table A2
Prevalence, based on the total study cohort of the five most common diagnosis codes within the "other diagnosis" label group.

	ICD-10 Code and Name	Prevalence (%)
1.	R06 Abnormalities of breathing	10.9
2.	R07 Pain in throat and chest	4.5
3.	I26 Pulmonary embolism	3.4
4.	J45 Asthma	2.5
5.	I48 Atrial fibrillation and flutter	1.8

Table A3
Comparison of CareNet's diagnostic performance in different cohort subgroups.

CareNet Model and Cohort Subgroup	Median Micro AUROC (%, 2.5–97.5 Percentile)
All cohort	87.8 (86.4–89.3)
Females	88.3 (86.4–90.3)
Males	87.7 (85.0–90.0)
0–2 diagnoses in medical history*	88.2 (86.8–89.7)
3 diagnoses in medical history*	68.2 (54.1–86.1)
Age $\leq$ 75 years	92.1 (90.3–94.1)
Age $>$ 75 years	81.7 (78.5–84.6)

* Visits having none, one, two, or all three diagnoses of heart failure, chronic obstructive pulmonary disease (COPD), and pneumonia registered anywhere in the regional electronic healthcare system in the preceding 12 months.

Table A4
Top 100 variables assembled by the CareNet Model Using Data Up to One Year Prior to Index Visit. In the right column, "selected" assigns that the variable was selected by the medical experts.

1	*Chronic obstructive pulmonary disease (COPD), primary diagnosis*	Selected
2	*Heart failure, primary diagnosis*	Selected
3	*ECG: Atrial fibrillation (unknown atrial activity)*	Selected
4	*Chronic obstructive pulmonary disease (COPD), secondary diagnosis*	Selected
5	*Age $\geq$ 95 years*	Selected
6	*Age 90–94 years*	Selected
7	*ECG: Ventricular-paced complexes, other complexes also detected*	
8	*ECG: Left Ventricular Hypertrophy with secondary repolarization abnormality (multi-LVH criteria, repolarisation abnormality)*	Selected
9	*Chronic obstructive pulmonary disease (COPD), primary care complaint*	Selected
10	*ECG: Atrial fibrillation (ventricular rate **-**, irregular ventricular activity)*	Selected
11	*Atrioventricular and left bundle-branch block, secondary diagnosis*	Selected
12	*Age 55–59 years*	Selected
13	*Heart failure, secondary diagnosis*	Selected
14	*Veterinary medication for cardiovascular system, collected medication*	
15	*Cutaneous abscess, furuncle and carbuncle, secondary diagnosis*	
16	*Age 85–89 years*	Selected
17	*Diseases of vocal cords and larynx, not elsewhere classified, secondary diagnosis*	
18	*Medication against obstructive airways, collected medication*	Selected
19	*Age 70–74 years*	Selected
20	*ECG: Anterior Q waves, possibly due to LVH (Q $>$ 30mS, V1 V2 & LVH)*	Selected
21	*Age 75–79 years*	Selected
22	*Age 80–84 years*	Selected
23	*"Heart failure," primary care and outpatient specialist care complaint*	Selected
24	*"Amputation," primary care and outpatient specialist care complaint*	Selected
25	*Age 45–49 years*	Selected

(*continued on next page*)

Table A4 (*continued*)

26	*ECG: Abnormal T, consider ischemia, diffuse leads (T < -0.20 mV, ant/lat/inf)*	Selected
27	*Hypertensive chronic kidney disease, secondary diagnosis*	
28	*Atrial fibrillation and flutter, secondary diagnosis*	Selected
29	*Other gynecological, collected medication*	
30	*Non-pressure chronic ulcer of lower limb, not elsewhere classified, primary diagnosis*	
31	*Age 65–69 years*	Selected
32	*ECG: Repolarisation abnormality, severe global ischemia ((LM/3VD) STe aVR, STd & Tneg, ant/lat/in)*	
33	*Age 50–54 years*	Selected
34	*ECG: Atrial-sensed ventricular-paced complexes (other complexes also detected)*	
35	*"Makula", primary care and outpatient specialist care complaint*	
36	*Antihypertensives, prescribed medication*	Selected
37	*Retention of urine, secondary diagnosis*	
38	*"Aid for personal needs", primary care and outpatient specialist care complaint*	
39	*Influenza due to identified seasonal influenza virus, secondary diagnosis*	Selected
40	*Fracture of lower leg, including ankle, primary diagnosis*	
41	*ECG: Repolarisation abnormality suggests ischemia, diffuse leads (ST-T neg, ant/lat/inf)*	Selected
42	*Type 2 diabetes mellitus, primary diagnosis*	Selected
43	*"Assessment", primary care and outpatient specialist care complaint*	
44	*ECG: Abnormal T, consider ischemia, lateral leads (T < -0.20 mV, I aVL V5 V6)*	Selected
45	*Other peripheral vascular diseases, secondary diagnosis*	Selected
46	*ECG: Intraventricular conduction delay, consider atypical LBBB (QRSd> **, notch/slur R I aVL V5-6)*	
47	*ECG: Sinus or ectopic atrial rhythm (P axis (−45,135))*	
48	*ECG: Left bundle-branch block (QRSd> **, broad/notched R)*	Selected
49	*Lab: Digoxin within normal range*	Selected
50	*Age 60–64 years*	Selected
51	*Antigout preparations, collected medication*	
52	*Mental and behavioral disorders due to use of opioids, primary diagnosis*	
53	*ECG: Atrial-paced complexes (other complexes also detected)*	
54	*ECG: Repolarisation abnormality suggests ischemia, lateral leads (ST dep, T neg, I aVL V5 V6)*	Selected
55	*Fracture of forearm, primary diagnosis*	
56	*Diabetes, chiropody, primary care and outpatient specialist care complaint**	Selected
57	*Neoplasm of uncertain behavior of urinary organs, primary diagnosis*	
58	*Gastric ulcer, primary diagnosis*	
59	*Other anemia, secondary diagnosis*	
60	*"Patient with ICD" (implantable cardioverter-defibrillator), primary care and outpatient specialist care complaint*	Selected
61	*Antigout preparations, prescribed medication*	
62	*ECG: Paired ventricular premature complexes (sequence of 2 V complexes)*	Selected
63	*"Prothrombin time" (PT), primary care and outpatient specialist care complaint*	
64	*"Chiropody", primary care and outpatient specialist care complaint*	
65	*ECG: Anterior ST elevation, probably due to LVH (ST > 0.20 mV in V1-V4 & LVH)*	Selected
66	*Rheumatic tricuspid valve diseases, secondary diagnosis*	
67	*Other gynecological, prescribed medication*	
68	*Complications of internal orthopedic prosthetic devices, implants and grafts, secondary diagnosis*	
69	*"Care planning", primary care and outpatient specialist care complaint*	
70	*Other and unspecified dorsopathies, not elsewhere classified, secondary diagnosis*	
71	*Osteoarthritis of hip, primary diagnosis*	
72	*Nonrheumatic mitral valve disorders, secondary diagnosis*	Selected
73	*"Nose bleeding", ED complaint*	
74	*ECG: Nonspecific T abnormalities, lateral leads (T < -0.10 mV, I aVL V5 V6)*	
75	*Erysipelas, primary diagnosis*	
76	*Cytology, primary care and outpatient specialist care complaint**	
77	*ECG: Ventricular bigeminy (bigeminy string > 4 w/ V complexes)*	Selected
78	*ECG: Nonspecific repolarisation abnormality, diffuse leads (ST dep, T flat/neg, ant/lat/inf)*	
79	*Problems related to lifestyle, secondary diagnosis*	Selected
80	*Lab: Creatinine above normal range*	Selected
81	*Measurement of blood pressure in toe or finger*	
82	*Other diseases of the digestive system, primary diagnosis*	
83	*Dislocation and sprain of joints and ligaments at ankle, foot and toe level, primary diagnosis*	
84	*"Check-up", primary care and outpatient specialist care complaint*	
85	*ECG: No further rhythm analysis attempted due to paced rhythm*	
86	*Symptoms and signs concerning food and fluid intake, secondary diagnosis*	
87	*Atopic dermatitis, primary diagnosis*	
88	*Right side*	
89	*ECG: Minimal ST depression, lateral leads (ST < -0.04 mV, I aVL V5 V6)*	Selected
90	*Radiology: X-ray of elbow*	
91	*Emphysema, secondary diagnosis*	
92	*Paralytic ileus and intestinal obstruction without hernia, primary diagnosis*	
93	*ECG: Repolarization abnormality, probably rate related (ST depression, T-negativity, tachycardia)*	Selected
94	*ECG: Anterior infarct, old (Q > 30mS, abnormal ST-T, V2-V5)*	Selected
95	*"Wound", primary care and outpatient specialist care complaint*	
96	*Lab: Prothrombin time (PT) above normal range*	
97	*Bacterial pneumonia, not elsewhere classified, secondary diagnosis*	Selected
98	*ECG: Ventricular premature complex (V complex w/ short R-R interval)*	Selected
99	*Triage: Oxygen saturation 80–85 % (regardless of oxygen gas treatment)*	Selected

References

[1] S. Ibsen, T.A. Lindskou, C.H. Nickel, T. Kløjgård, E.F. Christensen, M.B. Søvsø, "Which symptoms pose the highest risk in patients calling for an ambulance? A population-based cohort study from Denmark," (in eng), Scand J Trauma Resusc Emerg Med, vol. 29, no. 1, p. 59, Apr 20 2021, doi: 10.1186/s13049-021-00874-6.

[2] E. Jemt, M. Ekström, U. Ekelund, Outcomes in emergency department patients with dyspnea versus chest pain: a retrospective consecutive cohort study," (in eng), Emerg. Med. Int. 2022 (2022) 4031684, https://doi.org/10.1155/2022/4031684.

[3] M.D. Arvig, C.B. Mogensen, H. Skjøt-Arkil, I.S. Johansen, F.S. Rosenvinge, A. T. Lassen, "Chief complaints, underlying diagnoses, and mortality in adult, non-trauma emergency department visits: a population-based, multicenter cohort study," (in eng), West J. Emerg. Med. 23 (6) (2022) 855–863, https://doi.org/10.5811/westjem.2022.9.56332.

[4] G. Phipps, et al., "Contemporary management of acute heart failure in the emergency department and the potential impact of early diuretic therapy on outcomes," (in eng), Emerg. Med. Australas. 36 (1) (Feb 2024) 71–77, https://doi.org/10.1111/1742-6723.14301.

[5] D. Dzikowicz, A. Zemanek, M. Carey, Delay in door-to-diuretic time is associated with greater odds of 30-day readmission and mortality, J. Card. Fail. vol. 28, no. 5, Supplement 2022/04/01/ (2022) S5, https://doi.org/10.1016/j.cardfail.2022.03.016.

[6] P.M. Houck, D.W. Bratzler, W. Nsa, A. Ma, J.G. Bartlett, "Timing of antibiotic administration and outcomes for Medicare patients hospitalized with community-acquired pneumonia," (in eng), Arch. Intern. Med. 164 (6) (2004) 637–644, https://doi.org/10.1001/archinte.164.6.637.

[7] Ò. Miró, et al., "Early intravenous nitroglycerin use in prehospital setting and in the emergency department to treat patients with acute heart failure: Insights from the EAHFE Spanish registry," (in eng), Int. J. Cardiol. 344 (2021) 127–134, https://doi.org/10.1016/j.ijcard.2021.09.031.

[8] Y. Matsue, et al., Time-to-furosemide treatment and mortality in patients hospitalized with acute heart failure, J. Am. Coll. Cardiol. 69 (25) (2017) 3042–3051, https://doi.org/10.1016/j.jacc.2017.04.042, 2017/06/27/.

[9] N.T. Vozoris, et al., "Incident diuretic drug use and adverse respiratory events among older adults with chronic obstructive pulmonary disease," (in eng), Br. J. Clin. Pharmacol. 84 (3) (Mar 2018) 579–589, https://doi.org/10.1111/bcp.13465.

[10] K. Takagi, et al., "Safety of diuretic administration during the early management of dyspnea patients who are not finally diagnosed with acute heart failure," (in eng), Eur. J. Emerg. Med. 27 (6) (Dec 2020) 422–428, https://doi.org/10.1097/mej.0000000000000695.

[11] A.J. Singer, et al., "Bronchodilator therapy in acute decompensated heart failure patients without a history of chronic obstructive pulmonary disease," (in eng), Ann. Emerg. Med. 51 (1) (Jan 2008) 25–34, https://doi.org/10.1016/j.annemergmed.2007.04.005.

[12] P. Ray, et al., Acute respiratory failure in the elderly: etiology, emergency diagnosis and prognosis, Crit. Care 10 (3) (2006) R82, https://doi.org/10.1186/cc4926.

[13] S. Laribi, et al., Epidemiology of patients presenting with dyspnea to emergency departments in Europe and the Asia-Pacific region, Eur. J. Emerg. Med. 26 (5) (2019) 345–349, https://doi.org/10.1097/MEJ.0000000000000571.

[14] W. Kauppi, J. Herlitz, C. Magnusson, L. Palmér, and C. Axelsson, "Characteristics and outcomes of patients with dyspnoea as the main symptom, assessed by prehospital emergency nurses- a retrospective observational study," (in eng), BMC Emerg Med, vol. 20, no. 1, p. 67, Aug 28 2020, doi: 10.1186/s12873-020-00363-6.

[15] B.R. Celli, et al., "Differential diagnosis of suspected chronic obstructive pulmonary disease exacerbations in the acute care setting: best practice," (in eng), Am. J. Respir. Crit. Care Med. 207 (9) (2023) 1134–1144, https://doi.org/10.1164/rccm.202209-1795CI.

[16] S. Caravita, J.L. Vachiéry, "Obstructive ventilatory disorder in heart failure-caused by the heart or the lung?," (in eng), Curr. Heart Fail. Rep. 13 (6) (Dec 2016) 310–318, https://doi.org/10.1007/s11897-016-0309-5.

[17] N.M. Hawkins, S. Virani, C. Ceconi, "Heart failure and chronic obstructive pulmonary disease: the challenges facing physicians and health services," (in eng), Eur. Heart J. 34 (36) (Sep 2013) 2795–2803, https://doi.org/10.1093/eurheartj/eht192.

[18] P. Spörl, S.K. Beckers, R. Rossaint, M. Felzen, H. Schröder, "Shedding light into the black box of out-of-hospital respiratory distress-A retrospective cohort analysis of discharge diagnoses, prehospital diagnostic accuracy, and predictors of mortality," (in eng), PLoS One 17 (8) (2022) e0271982, https://doi.org/10.1371/journal.pone.0271982.

[19] S. H. Ovesen, S. F. Sørensen, M. Lisby, M. H. Mandau, I. K. Thomsen, and H. Kirkegaard, "Change in diagnosis from the emergency department to hospital discharge in dyspnoeic patients," (in eng), Dan Med J, vol. 69, no. 2, Jan 27 2022.

[20] K.M. Hunold, J.M. Caterino, High diagnostic uncertainty and inaccuracy in adult emergency department patients with dyspnea: a national database analysis, Acad. Emerg. Med. 26 (2) (2019) 267–271, https://doi.org/10.1111/acem.13553.

[21] R. Sikka, L.H. Tommaso, C. Kaucky, E.B. Kulstad, "Diagnosis of pneumonia in the ED has poor accuracy despite diagnostic uncertainty," (in eng), Am. J. Emerg. Med. 30 (6) (Jul 2012) 881–885, https://doi.org/10.1016/j.ajem.2011.06.006.

[22] A. Chandra, B. Nicks, E. Maniago, A. Nouh, A. Limkakeng, "A multicenter analysis of the ED diagnosis of pneumonia," (in eng), Am. J. Emerg. Med. 28 (8) (Oct 2010) 862–865, https://doi.org/10.1016/j.ajem.2009.04.014.

[23] A. Atamna, S. Shiber, M. Yassin, M.J. Drescher, J. Bishara, "The accuracy of a diagnosis of pneumonia in the emergency department," (in eng), Int. J. Infect. Dis. 89 (Dec 2019) 62–65, https://doi.org/10.1016/j.ijid.2019.08.027.

[24] S. Blecker, et al., Early identification of patients with acute decompensated heart failure, J. Card. Fail. 24 (6) (2018) 357–362, https://doi.org/10.1016/j.cardfail.2017.08.458.

[25] D.-J. Choi, J.J. Park, T. Ali, S. Lee, Artificial intelligence for the diagnosis of heart failure, npj Digital Med. 3 (1) (2020) 54.

[26] C.-T. Kor, Y.-R. Li, P.-R. Lin, S.-H. Lin, B.-Y. Wang, C.-H. Lin, Explainable machine learning model for predicting first-time acute exacerbation in patients with chronic obstructive pulmonary disease, J. Personalized Med. 12 (2) (2022) 228.

[27] S. Swaminathan, et al., A machine learning approach to triaging patients with chronic obstructive pulmonary disease, PLoS One 12 (11) (2017) e0188532, https://doi.org/10.1371/journal.pone.0188532.

[28] M. Chumbita, et al., Can artificial intelligence improve the management of pneumonia, J. Clin. Med. 9 (1) (2020) 248, https://doi.org/10.3390/jcm9010248.

[29] D. Orso, N. Guglielmo, R. Copetti, Lung ultrasound in diagnosing pneumonia in the emergency department: a systematic review and meta-analysis, Eur. J. Emerg. Med. 25 (5) (2018) 312–321.

[30] P.S. Heckerling, B.S. Gerber, T.G. Tape, R.S. Wigton, Prediction of community-acquired pneumonia using artificial neural networks, Med. Decis. Making 23 (2) (2003) 112–121, https://doi.org/10.1177/0272989X03251247.

[31] O. Er, C. Sertkaya, F. Temurtas, A.C. Tanrikulu, A comparative study on chronic obstructive pulmonary and pneumonia diseases diagnosis using neural networks and artificial immune system, J. Med. Syst. 33 (6) (2009) 485–492, https://doi.org/10.1007/s10916-008-9209-x.

[32] Predicare, "Rapid Emergency Triage Treatment Scale (RETTS©) online version 2019," ed. Göteborg, Sweden: Predicare AB, 2020.

[33] O. World Health, "International Classification of Diseases (ICD) ICD-10 2019," ed. Geneva, 2020.

[34] E.T. Heyman, et al., A novel interpretable deep learning model for diagnosis in emergency department dyspnoea patients based on complete data from an entire health care system, PLoS One 19 (12) (2024) e0311081.

[35] A. Ashfaq, et al., Data resource profile: Regional healthcare information platform in Halland, Sweden, a dedicated environment for healthcare research, Int. J. Epidemiol. (2020).

[36] M. R. Philips Medical Systems, Andover, MA 01810 USA, "Philips DXL ECG algorithm, physician's guide," April 2009. [Online]. Available: https://www.documents.philips.com/doclib/enc/fetch/2000/4504/577242/577243/577246/581601/711562/DXL_ECG_Algorithm_Physician_s_Guide_%28ENG%29_Ed.2.pdf.

[37] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1480–1489.

[38] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv: 1412.3555, 2014.

[39] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Adv. Neural Inf. Proces. Syst. 26 (2013).

[40] A.V. Dorogush, V. Ershov, A. Gulin, CatBoost: gradient boosting with categorical features support. arXiv Preprint arXiv:1810.11363, 2018.

[41] J.T. Hancock, T.M. Khoshgoftaar, CatBoost for big data: an interdisciplinary review, Journal of Big Data 7 (1) (2020) 94, https://doi.org/10.1186/s40537-020-00369-8, 2020/11/04.

[42] A. Paszke, et al., Pytorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Proces. Syst. 32 (2019).

[43] N. Ibm Corp, IBM SPSS Statistics for Windows vol. 29.0 (2022).

[44] M. W. Microsoft Corporation, Redmond, WA, USA, "Microsoft Excel för Microsoft 365 MSO " vol. Version 2402, ed, 2024.

[45] S. Jain and B. C. Wallace, "Attention is not explanation," arXiv preprint arXiv: 1902.10186, 2019.

[46] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," arXiv preprint arXiv: 1908.04626, 2019.

[47] D. Bahdanau, Neural machine translation by jointly learning to align and translate. arXiv Preprint arXiv:1409.0473, 2014.

[48] S. Pokhrel Bhattarai, R.C. Block, Y. Xue, D.H. Rodriguez, R.G. Tucker, M.G. Carey, Integrative review of electrocardiographic characteristics in patients with reduced, mildly reduced, and preserved heart failure, Heart Lung 63 (2024) 142–158, https://doi.org/10.1016/j.hrtlng.2023.10.012, 2024/01/01/.

[49] T. Rusinowicz, T.M. Zielonka, K. Zycinska, Cardiac Arrhythmias in Patients with Exacerbation of COPD (Clinical Management of Pulmonary Disorders and Diseases), Springer International Publishing, Cham, 2017, pp. 53–62.

[50] P.D. Stein, et al., Electrocardiogram in pneumonia, Am. J. Cardiol. 110 (12) (2012) 1836–1840, https://doi.org/10.1016/j.amjcard.2012.08.019, 2012/12/15/.

[51] E.L. Potter, I. Hopper, J. Sen, A. Salim, T.H. Marwick, "Impact of socioeconomic status on incident heart failure and left ventricular dysfunction: systematic review and meta-analysis," (in eng), Eur Heart J Qual Care Clin Outcomes 5 (2) (2019) 169–179, https://doi.org/10.1093/ehjqcco/qcy047.

[52] A.S. Gershon, T.E. Dolmage, A. Stephenson, B. Jackson, Chronic obstructive pulmonary disease and socioeconomic status: a systematic review, COPD: J. Chron. Obstruct. Pulmon. Dis. 9 (3) (2012) 216–226, https://doi.org/10.3109/15412555.2011.648030, 2012/05/23.

[53] P.D. Blanc, et al., "The occupational burden of nonmalignant respiratory diseases. an official american thoracic society and european respiratory society statement,"

(in eng), Am. J. Respir. Crit. Care Med. 199 (11) (2019) 1312–1334, https://doi.org/10.1164/rccm.201904-0717ST.
[54] D. E. Newman-Toker et al., "Diagnostic errors in the emergency department: a systematic review," 2022.
[55] A. Ashfaq, Deep Evidential Doctor, Halmstad University Press, 2022.