Triaging Patients With Artificial Intelligence for Respiratory Symptoms in Primary Care to Improve Patient Outcomes: A Retrospective Diagnostic Accuracy Study

Steindór Ellertsson, *MD*[1]

Hlynur D. Hlynsson, *PhD*[2]

Hrafn Loftsson, *PhD*[2]

Emil L. Sigurðsson, *MD, PhD*[1,3,4]

[1]Primary Health Care of the Capital Area, Iceland

[2]Department of Computer Science, Reykjavik University, Reykjavík, Iceland

[3]Development Center for Primary Health Care in Iceland, Reykjavík, Iceland

[4]Department of Family Medicine, University of Iceland, Reykjavík, Iceland

Conflicts of interest: authors report none.

CORRESPONDING AUTHOR

Emil L. Sigurðsson
Development Center for Primary
Health Care in Iceland
Álfabakka 16
109 Reykjavík, Iceland
emilsig@hi.is

ABSTRACT

PURPOSE Respiratory symptoms are the most common presenting complaint in primary care. Often these symptoms are self resolving, but they can indicate a severe illness. With increasing physician workload and health care costs, triaging patients before in-person consultations would be helpful, possibly offering low-risk patients other means of communication. The objective of this study was to train a machine learning model to triage patients with respiratory symptoms before visiting a primary care clinic and examine patient outcomes in the context of the triage.

METHODS We trained a machine learning model, using clinical features only available before a medical visit. Clinical text notes were extracted from 1,500 records for patients that received 1 of 7 *International Classification of Diseases 10th Revision* codes (J00, J10, JII, J15, J20, J44, J45). All primary care clinics in the Reykjavík area of Iceland were included. The model scored patients in 2 extrinsic data sets and divided them into 10 risk groups (higher values having greater risk). We analyzed selected outcomes in each group.

RESULTS Risk groups 1 through 5 consisted of younger patients with lower C-reactive protein values, re-evaluation rates in primary and emergency care, antibiotic prescription rates, chest x-ray (CXR) referrals, and CXRs with signs of pneumonia, compared with groups 6 through 10. Groups 1 through 5 had no CXRs with signs of pneumonia or diagnosis of pneumonia by a physician.

CONCLUSIONS The model triaged patients in line with expected outcomes. The model can reduce the number of CXR referrals by eliminating them in risk groups 1 through 5, thus decreasing clinically insignificant incidentaloma findings without input from clinicians.

Ann Fam Med 2023;21:240-248. https://doi.org/10.1370/afm.2970

INTRODUCTION

Health care costs have steadily increased in recent decades.[1] General practitioners face a greater number of patients,[2,3] with more comorbidities[4] and demands,[5] and diagnostic test orders have increased substantially.[6] Around 20% of patient visits to general practitioners stem from self-resolving symptoms,[7] and up to 72% of patient visits are due to acute respiratory symptoms.[8] Overuse and misuse of diagnostic tests is a well-known problem in primary care[9,10] that increases incidental findings.[11,12] The same applies to antibiotic prescribing,[13] especially for respiratory tract infections,[14] leading to increased bacterial resistance.[15] The causes for clinical resource misapplication are multifactorial, but patient demands, human biases, and time pressure play substantial roles.[16,17]

Machine learning models (MLMs) are thought to be similar or superior to physicians in multiple clinical tasks.[18-27] Patient triage using MLMs is reportedly comparable to triage by physicians.[28,29] Research in tertiary settings showed MLMs to be superior to physicians at estimating patient risk when ordering diagnostic tests.[30] Clinical guidelines and scoring systems can standardize diagnosis and treatment and improve the quality of care while reducing costs,[31-34] but remain underused.[35-37] Guideline applicability, useability, and time scarcity are cited as reasons why.[38,39]

Structured triage with standardized questionnaires is likely safer than unstructured triage.[40] Assistance from a clinical decision support system increases triage

quality.[41] By design, MLMs use standardized inputs, making them a good fit for integration into a clinical decision support system, and such systems have been shown to reduce health care costs by 14%.[42] Triaging patients at the time of appointment scheduling is even more important since the COVID-19 pandemic. Methods to identify patients well suited to virtual consultations are needed as they now make up 13% to 17% of consultations across all specialties.[43]

Clinical text notes (CTNs) are a written record of a physician's interpretation of the patient's symptoms and signs, reasons for clinical decisions made during the consultation, and actions taken (eg, imaging referrals, prescriptions written). The objective of this study was to train a patient triage MLM on symptoms and signs (clinical features) of patients with respiratory symptoms, using only features the patient could be asked about in order to mimic previsit triage. We extracted the clinical features from CTNs.

This MLM, which we refer to as a respiratory symptom triage model (RSTM), divides patients into 10 risk groups (with increasing risk from groups 1 to 10) based on a score. We validated the RSTM by examining patient outcomes, stratified by risk group, on intrinsic data, and in 2 separate extrinsic (unseen) data sets. Evaluating of MLM performance in a medical context is complex, and knowing which benchmarks to use is often unclear. Many reports benchmark MLMs against physicians' diagnoses which are affected by human biases and errors.[44] Benchmarking the RSTM against multiple patient outcomes likely serves as a better performance metric, and, to our knowledge, no reports have examined MLM triage performance in this way.

METHODS

In this retrospective diagnostic accuracy study, we obtained 44,007 medical records of 23,819 patients from a medical database common to all primary clinics in the Reykjavík area of Iceland. Each record contained a CTN with diagnostic referrals and results, diagnoses, and prescriptions written.

The selection criteria were patients over the age of 18 years who were diagnosed by a physician from January 1, 2016 through December 31, 2018 with 1 of 7 *International Classification of Diseases 10th Revision* (*ICD-10*) codes: J00 (common cold), J10 and J11 (influenza), J15 (bacterial pneumonia), J20 (acute bronchitis), J44 (chronic obstructive pulmonary disease [COPD]), and J45 (asthma), including subgroups. We removed CTNs containing less than 250 characters, resulting in 17,177 CTNs included in this study.

In our previous work, we trained a deep neural network to extract clinical features from CTNs,[45] which we call the clinical feature extraction model. We randomly selected 7,000 CTNs as input to the clinical feature extraction model and discarded CTNs with less than 8 clinical features, increasing the odds of having enough clinical features in each for the RSTM. The clinical feature extraction model also extracted presenting complaints which we used to limit inclusion to

only patients presenting with acute or subacute respiratory symptoms. The complete list of presenting complaints is in Supplemental Table 1. We removed 95 CTNs from follow-up consultations and 223 CTNs with multiple topics to include only CTNs in which patients presented with a new respiratory complaint as a single complaint. Thus, for patients diagnosed with COPD and asthma, only cases of exacerbation were included.

Applying these filters reduced the set of 7,000 CTNs to 2,942. Of those, 2,000 CTNs were randomly selected and manually annotated by a single physician. As annotating CTNs is costly, the final number of 2,000 CTNs was limited by funding. We split the resulting data set randomly into training (75%, 1,500 CTNs) and test (25%, 500 CTNs) sets. We also annotated an additional 664 CTNs with influenza *ICD-10* codes (J10, J11) as a second test set to examine the generalizability of the RSTM further because there were no influenza patients in the training data set.

Subsequently, we trained the RSTM on features that patients can be asked about and measure themselves, imitating a setting where triage takes place before a medical consultation. We chose the input features that a web-based triage system could obtain directly from a patient without other human assistance to ensure the model fits into a clinical workflow. The training objective was to predict the likelihood of a patient having a lower respiratory tract infection. We considered all diagnoses except J00 (common cold) to be a lower respiratory tract infection.

The RSTM had a single output: a score between 0 and 1, where patient scores approaching 1 have an increased probability of a lower respiratory tract infection diagnosis. We performed 25 repeats of a 4-fold stratified nested cross validation for hyperparameter search and intrinsic validation. We then trained the RSTM on the training data set with optimized hyperparameters before splitting patients in the test sets into 10 risk groups based on the score they received. The risk score interval for each group was 0.1, and we refer to groups 1 through 5 as the low-risk groups and 6 through 10 as the high-risk groups.

Annotation

The annotation method was inspired by researchers who applied similar annotations on medical text,[19] which assigned binary and numerical values to clinical features, representing the presence or absence of signs and symptoms in the CTNs, as they were described in text. They constituted the patient's health state as described by a physician during the medical consultation when the CTN was written. A detailed description of the annotation process is in the Supplemental Appendix. We gave missing binary features the value of 0. Missing value features were replaced by randomly sampling the normal distribution for that feature to reduce the odds of the model simply learning where features are missing, which would be more likely for a patient with less severe disease.

Model Architecture, Hyperparameter Optimization, and Training

The classifier we used was a type of logistic regression with Least Absolute Shrinkage and Selection Operator penalty. We used Shapley Additive Explanation[46] values to extract the 50 most impactful clinical features to use as input features into the RSTM to reduce the risk of a spurious correlation between the input and output data. A full list of the model clinical features can be found in Supplemental Table 2. We performed 25 repeats of a 4-fold stratified nested cross validation with grid search on the training set, to optimize the hyperparameters of the RSTM. Only class weight and the penalty (C parameter) were optimized, resulting in use of a balanced parameter for class weight and a C value of 0.1 during training. Then we trained the RSTM on the training data set before running inference on the patients in the test sets.

Outcomes and Statistical Analysis

For each risk group, we examined the following outcomes: mean C-reactive protein (CRP) value, *ICD-10* code distribution, the proportion of patients re-evaluated in primary care and emergency departments within 7 days, the proportion of patients referred for a CXR, CXRs with signs of pneumonia and incidentalomas, and proportion of patients receiving antibiotic prescriptions. C-reactive protein values were only available if the physician deemed it necessary and were extracted from the CTN since rapid-CRP test results are saved only in the CTN not in a structural database in Iceland. Referrals for CXRs and results are linked to a CTN and the textual answer from the radiologist was manually annotated in the same manner as the CTNs. Except for incidentalomas and *ICD-10* codes, a positive or a higher outcome value indicated more severe disease for a given patient. Notes about consolidations, infiltrations, and pneumonia-like signs in the CXR's text description were considered positive signs of pneumonia. All data sources were from the patients' electronic health records. The 95% CIs were calculated by sorting the values for each outcome and calculating the 2.5% and 97.5% percentiles. We used a 2-sided Fisher's exact test to calculate *P* values for binary variables and a 2-sided Mann-Whitney U test for continuous variables. We considered $P < .05$ to be significant. We implemented data analysis in Python (version 3.8) and trained and validated the RSTM with the scikit-learn library (0.22.1).[47]

RESULTS

A total of 2,000 CTNs from 1,915 patients were included in the final data set. There were 26,971 annotations, for an average of 13.5 annotations per CTN. The flowchart of CTN selection of the first test set is shown in Figure 1. In the second test set, 664 CTNs from 652 patients were included. Table 1 shows the demographics for each data set, *ICD-10* code, and mean outcome distribution. The baseline outcome rates are similar to those reported by others.[48-51] Patients with pneumonia on CXRs received antibiotic prescriptions in 46% of cases. All incidentalomas were of nodule subtype and none had clinical significance. Table 2 compares the outcome rates in the low-risk and high-risk groups in the test sets with calculated *P* values. There was a significant difference between the groups in CRP values and antibiotic treatment in test set 1 and only in antibiotic treatment in test set 2. No evaluations in the emergency department resulted in a pneumonia diagnosis, and 83% received the same *ICD-10* code as they received in the initial consultation. No primary care re-evaluations resulted in a pneumonia diagnosis, and 80% received the same *ICD-10* code they initially received.

Outcome distributions stratified by risk group are shown in Figure 2 (training set), Figure 3 (set 1), and Figure 4 (set 2). The low-risk groups in the training set (Figure 2) contain no

Figure 1. Study flowchart for clinical text note selection.
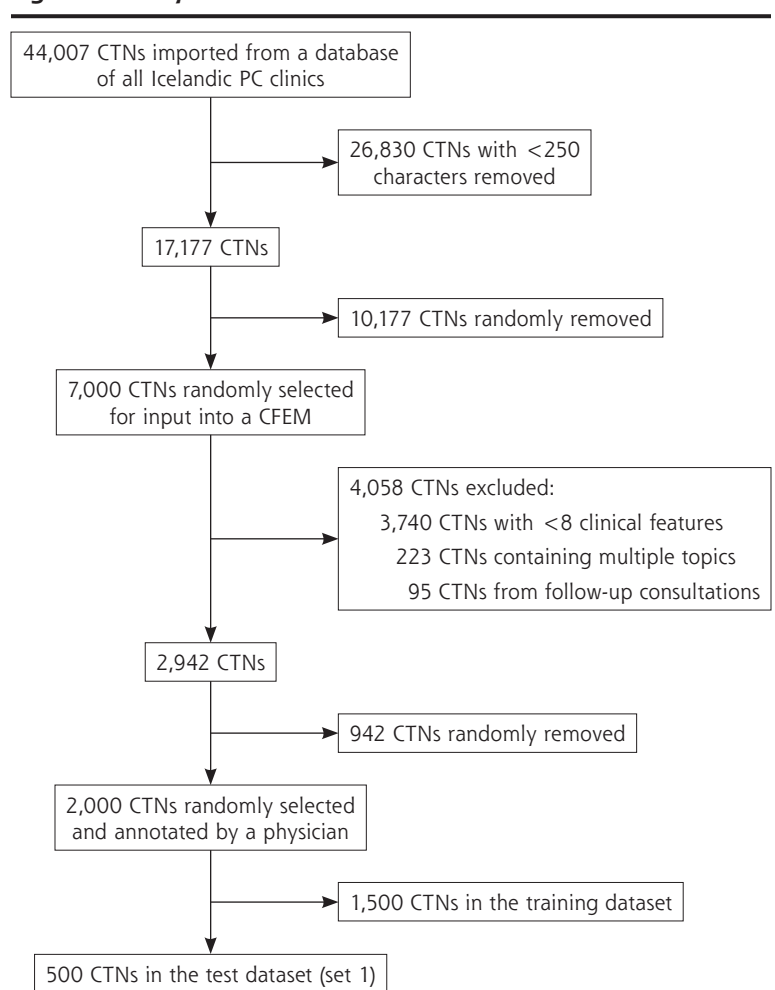
CFEM = clinical feature extraction model; CTN = clinical text note; PC = primary care.

Table 1. Demographics, *ICD-10* Code Distributions, and Outcomes in the Data Sets

Variable	Training Set	Test Set 1	Test Set 2
Demographics			
Patients, No.	1,500	500	664
Age, mean (range), y	54	55	45
	(18-93)	(19-91)	(18-92)
Sex, female, %	66.1	62.6	61.1
***ICD-10* codes, %**			
J00	20.2	19.2	0.0
J15	0.4	0.2	0.0
J20	46.5	48.2	0.0
J44	11.1	9.8	0.0
J45	21.8	22.6	0.0
J10	0.0	0.0	0.8
J11	0.0	0.0	99.2
Patient outcomes			
CRP value, mean (range), mg /dL	20.7	18.8	28.2
	(3-183)	(3-84)	(3-178)
Antibiotics prescribed, %	49.0	50.8	15.5
CXRs ordered, %	6.6	6.8	3.5
CXR signs of pneumonia, %	12.6	9.6	9.1
CXR incidentaloma, %	1.6	3.2	4.3
PC re-evaluation, %	8.2	6.8	10.0
ED re-evaluation, %	0.2	0.6	0.8

CRP = C-reactive protein; CXR = chest x-ray; ED = emergency department; *ICD-10* = *International Classification of Diseases, 10th Revision;* J00 = common cold; J11 = influenza; J12 = viral pneumonia; J15 = bacterial pneumonia; J20 = acute bronchitis; J44 = chronic obstructive pulmonary disease; ; J45 = asthma; PC = primary care.

positive CXRs, 52% of the incidentalomas, and 9% of CXR referrals. In the first test set, the low-risk groups included one-third of the patients who were younger, and had lower CRP values, antibiotic prescription rates, re-evaluation rates, no positive CXRs, and 19% of CXR referrals. In the second test set, 45% of patients and 35% of CXR referrals were in low-risk groups, that had no CXRs with signs of pneumonia and the single incidentaloma found. The outcome trends in Figures 2, 3, and 4 show rising outcome rates with higher outcome groups for all outcomes, except for re-evaluation in primary care and CRP values in the second test set.

DISCUSSION

In this large retrospective study, we show, for the first time, the results of patient triage by MLMs in primary care, using only data available before a medical consultation, in the context of patient outcomes. The RSTM performs the triage such that patients in high-risk groups have more severe outcomes than those in lower-risk groups. Importantly, no patient in the lowest 5 risk groups had a CXR with signs of pneumonia or a pneumonia *ICD-10* code. Despite patients in test set 2 coming from a different population than patients in the training data set, the triage shows an outcome distribution pattern similar to that of test set 1, further validating that the RSTM triages pre-consultation patients similarly. The nested cross validation shows an underlying signal across the whole data set, allowing the RSTM to triage the patients aligned with expected outcomes, regardless of how the data set is split and ordered. The outcome distribution is similar in all data sets, indicating a general good

Table 2. Comparison of Outcome Rates in the Test Sets Between Low-Risk and High-Risk Groups

Outcomes[a]	Low-Risk Group	High-Risk Group	Sensitivity	Specificity	PPV	NPV	*P* Value
Test set 1							
Mean CRP values	7.9	23.0	...	...	...	...	<.05
Antibiotics prescribed	66 (40%)	188 (56.1%)	0.74	0.40	0.56	0.60	<.005
PC re-evaluation	9 (5.4%)	36 (10.7%)	0.80	0.34	0.11	0.95	.07
ED re-evaluation	1 (0.6%)	3 (0.89%)	0.50	0.33	0.01	0.98	.67
CXRs ordered	6 (3.6%)	25 (7.4%)	0.81	0.34	0.075	0.96	.07
Positive CXRs[b]	0 (0.0%)	3 (0.89%)	1.00	0.33	0.01	1.00	.30
Test set 2							
Mean CRP values	29.4	27.4	...	...	...	...	1.00
Antibiotics prescribed	33 (10.9%)	68 (18.7%)	0.67	0.48	0.19	0.89	<.05
PC re-evaluation	30 (9.9%)	35 (9.6%)	0.54	0.45	0.10	0.90	.90
ED re-evaluation	1 (0.3%)	4 (1.2%)	0.80	0.46	0.01	1.00	.38
CXRs ordered	8 (2.7%)	15 (4.1%)	0.65	0.46	0.04	0.97	.40
Positive CXRs[b]	0 (0.0%)	2 (0.6%)	1.00	0.45	0.01	1.00	.50

CRP = C-reactive protein; CXR = chest x-ray; ED = emergency department; NPV = negative predictive value; PC = primary care; PPV = positive predictive value.

Note: Test set 1 had 165 patients in low-risk group and 335 in high risk group; Test set 2 had 301 patients in low-risk group and 363 in high-risk group.

[a] Mean CRP values reported for groups in mg/dL. Other outcomes reported for groups as number of patients (percentage).
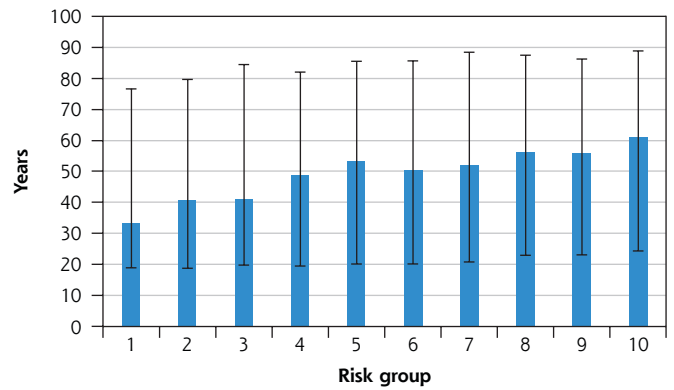[b] Positive CXRs included those with signs of pneumonia only, incidentalomas were excluded.

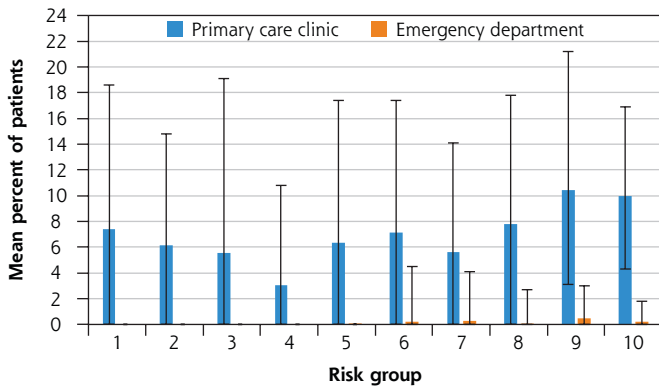## Figure 2. The outcome distribution in the cross-validated data set.
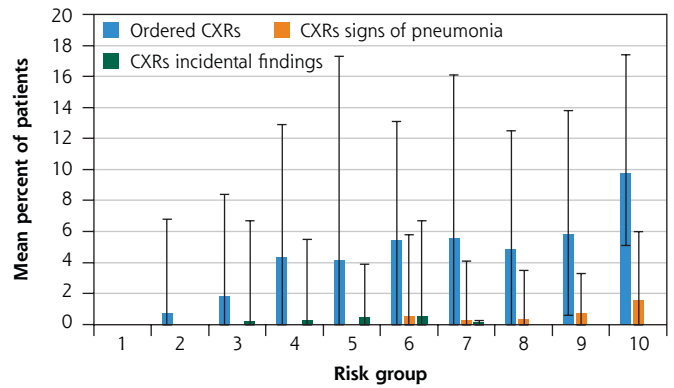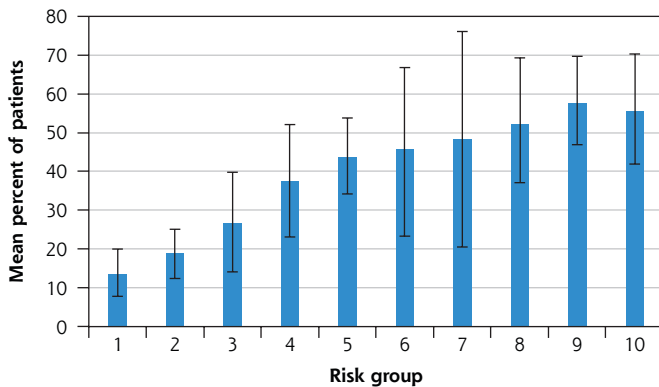

A. Mean patient count
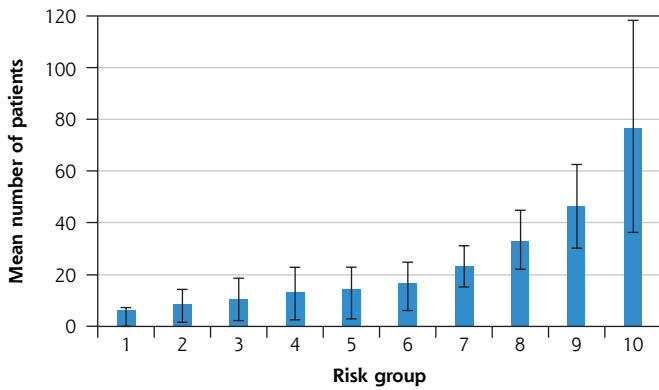
B. Mean patient age

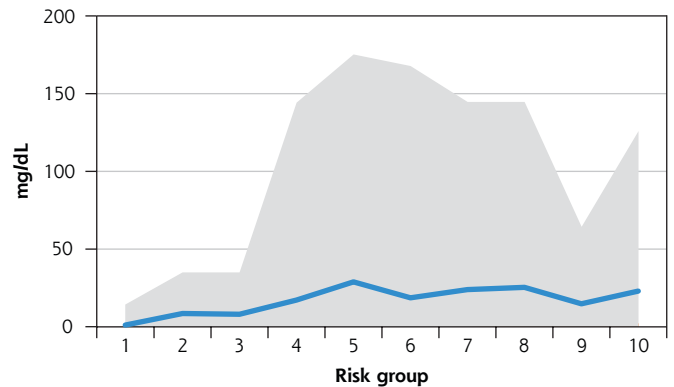**C. Re-evaluation within 7 days**
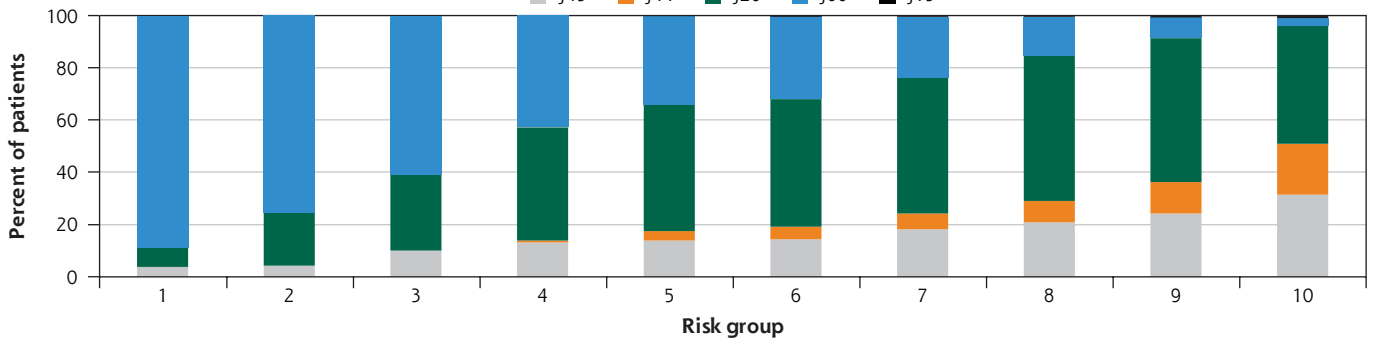
D. Chest x-rays

E. Antibiotics prescribed

F. Mean CRP value

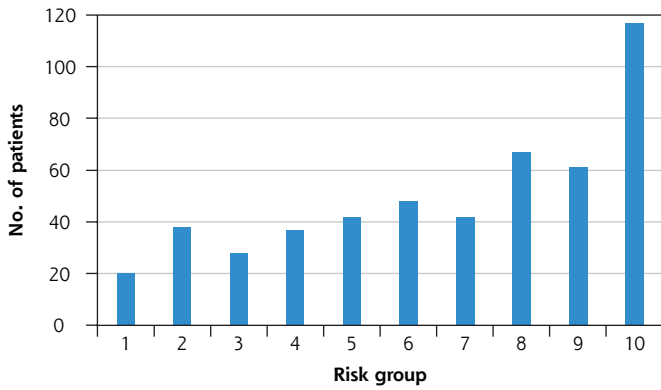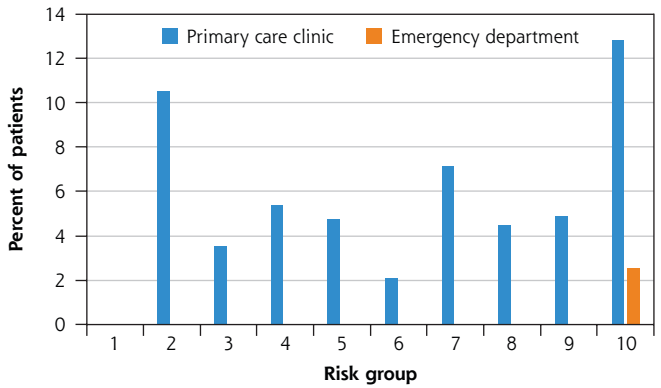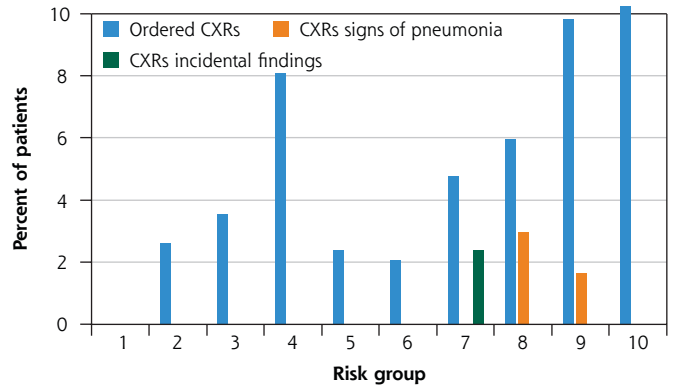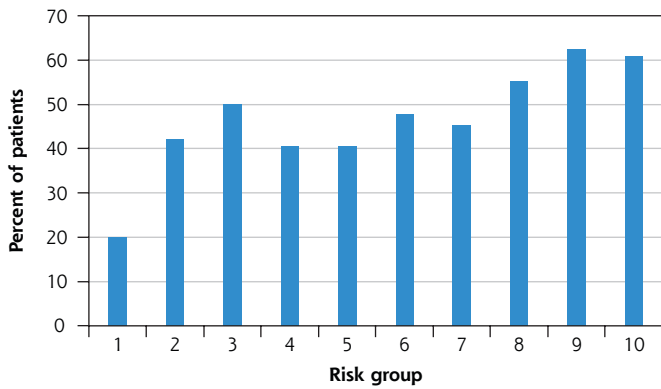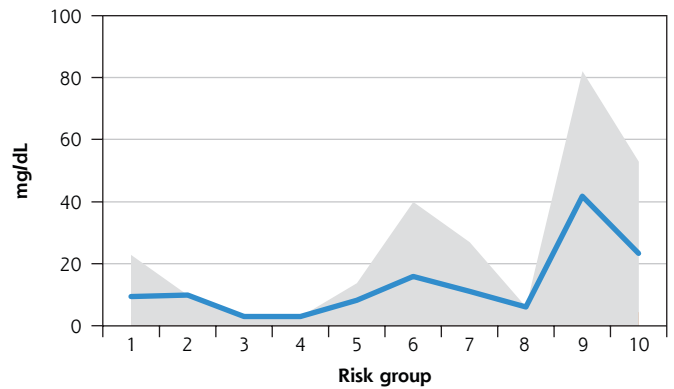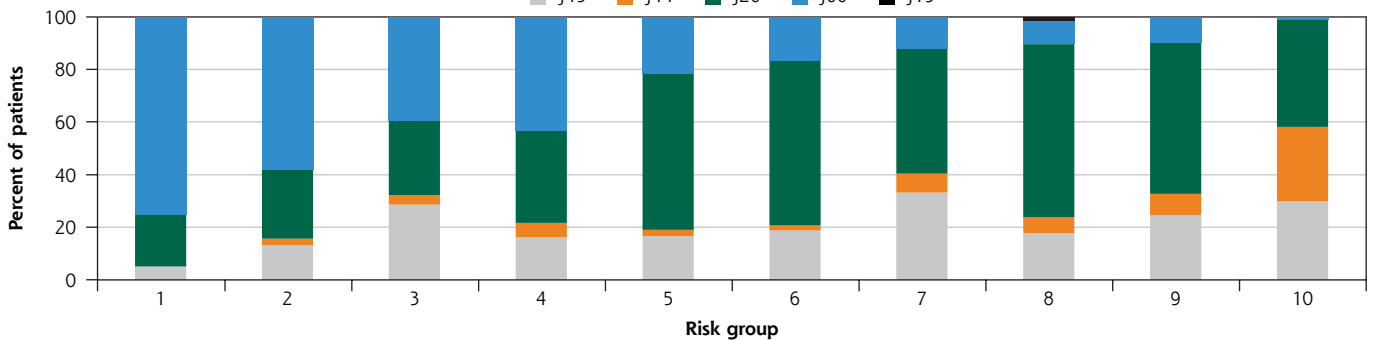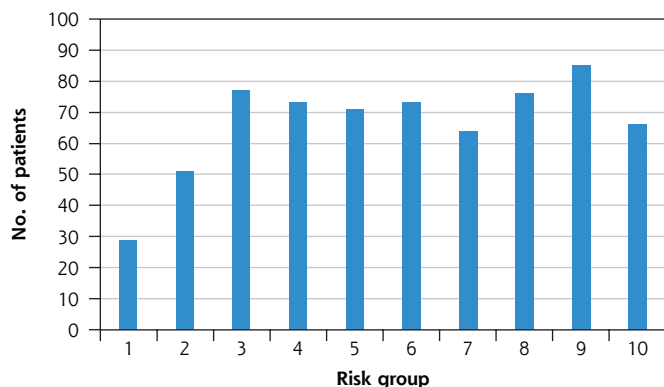**G. *ICD-10* code distribution**

CXR = chest x-ray; CRP = C-reactive protein; *ICD-10 = International Classification of Diseases, 10th Revision*; J00 = common cold; J15 = bacterial pneumonia; J20 = acute bronchitis; J44 = chronic obstructive pulmonary disease; J45 = asthma.

Notes: (A-E) bars represent 95% CIs. (F) shaded area represents 95% CIs.

Figure 3. The outcome distribution in test set 1.

CXR = chest x-ray; CRP = C-reactive protein; dL= deciliter; ED = emergency department; *ICD-10 = International Classification of Diseases, 10th revision;* J00 = common cold; J15 = bacterial pneumonia; J20 = acute bronchitis; J44 = chronic obstructive pulmonary disease; J45 = asthma; mg = milligram; PC = primary care.
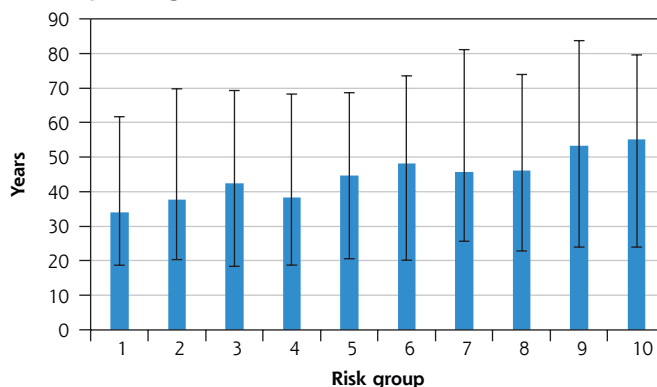
Notes: (B) bars represent 95% CIs. (F) shaded area represents 95% CIs.
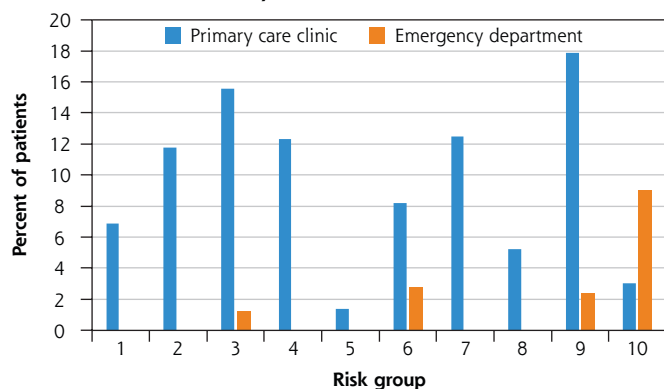
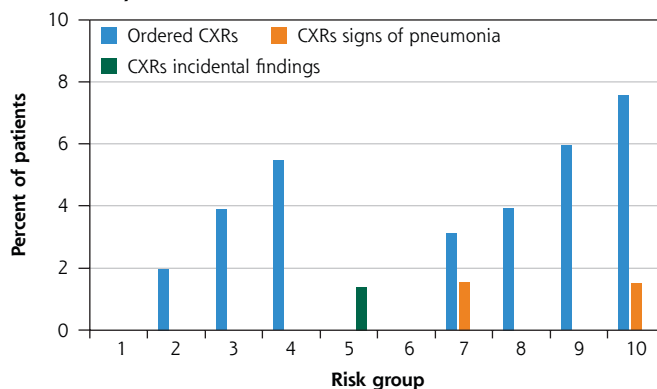**Figure 4. The outcome distribution in test set 2.**
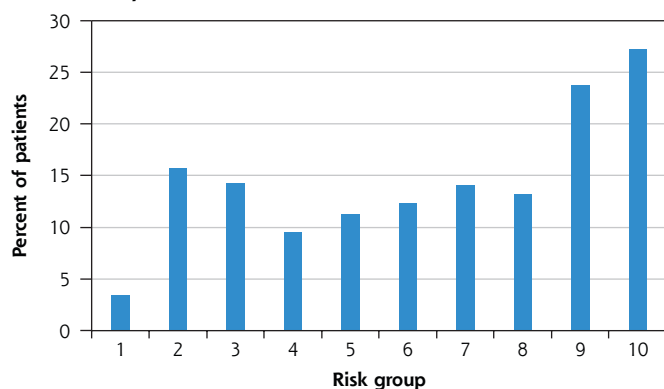
A. Patient count

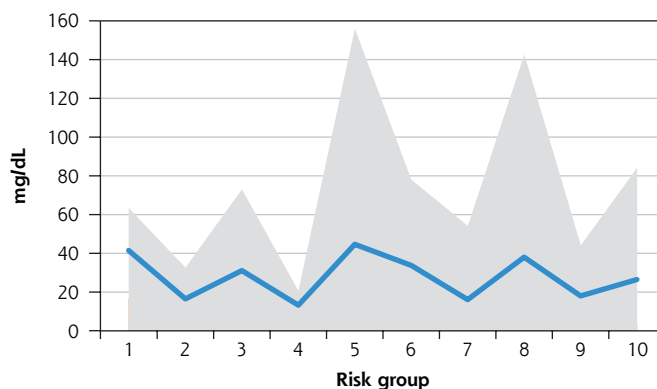B. Mean patient age

C. Re-evaluation within 7 days

D. Chest x-rays

E. Antibiotics prescribed

F. Mean CRP value

CXR = chest x-ray; CRP = C-reactive protein; dL= deciliter; *ICD-10 = International Classification of Diseases, 10th revision;* mg = milligram; PC = primary care.

Notes: (B) bars represent 95% CIs. (F) shaded area represents 95% CIs. *ICD-10* code distribution in test set 2 was not examined for these symptomatically similar patients.

model fit to the data. Interestingly, the RSTM is ignorant of *ICD-10* code subtypes but scores J15 (bacterial pneumonia) patients at an increasing rate in groups 4 through 10, while J00 (common cold) and J20 (acute bronchitis) decrease proportionally. J44 (COPD) was only found in groups 2 though 8, indicating that the model considers patients with pneumonia (J15) and COPD (J44) most likely for worse outcomes, matching reality.

Findings Compared With Other Studies

We were unable to find similar studies, but multiple studies have attempted to derive a diagnostic rule for pneumonia from the signs and symptoms of patients. All but 1 include features in their rules which make them incomparable to the RSTM. When we compare the scores of the RSTM and the diagnostic rule from the 1 comparable study,[52] we see a linear correlation (**Supplemental Figure 1**). Those authors

concluded that using the diagnostic rule in clinical settings would substantially reduce antibiotic use and CXR imaging,[52] which coincides with our findings. We also compared the score of the RSTM to the Anthonisen score,[53] which recommends antibiotic treatment for exacerbation of COPD if 2 of 3 cardinal symptoms are positive (increased sputum expectoration, increased dyspnea, purulent sputum production). Their score coincides well with the risk prediction of the RSTM (Supplemental Figure 2) and recommends that COPD patients in the low-risk groups should not be treated with antibiotics.

Clinical Implications

If the RSTM shows similar performance in clinical settings, it could be implemented as a web-based tool, potentially triaging patients online before they make an appointment. The triage could potentially identify patients with low risk of lower respiratory tract infection, that could be attended to without the need for face-to-face consultations. The RSTM could eliminate CXR referrals for patients in groups where the probability of them being positive is low or nonexistent, which would remove up to one-third of CXRs and possibly one-half of the incidentalomas without missing a positive CXR. Despite all patients in the low-risk groups receiving diagnoses where the benefit of antibiotics is debatable, antiobiotics were substantially prescribed. Reducing antibiotic prescriptions in the low-risk groups would increase prescription quality. The RSTM score needs no input from clinicians and can be ready when a patient enters the examination room, resulting in an easy-to-use, unambiguous, applicable score with a meaningful effect. Thus, the RTSM can possibly reduce costs for patients, the health care system, and society.

Strengths and Limitations

The strengths of this study include a large data set of patients with 2 distinct test sets. Using multiple patient outcomes, stratified by risk groups, gives more insight into the performance and safety of the triage instead of using only physicians' diagnoses as benchmarks. The study is subject to limitations and biases of a retrospective methodology, and the findings must be validated prospectively. The CTNs are a written record of the physician's interpretation of patients' symptoms and signs and contain human errors and biases, making the RSTM erroneous and biased. Removing CTNs with less than 8 clinical features creates selection bias, likely toward patients with more severe symptoms. Direct data collection from patients would provide more quality training data. There is availability bias in the CRP values and CXR outcomes. Performing annotation with multiple physicians would likely result in more quality annotations.

🐦 **Read or post commentaries in response to this article.**

Key words: artificial intelligence; clinical decision support systems; primary care; triage; respiratory symptoms

🐦 **Supplemental materials**

REFERENCES

1. OECD. Health spending. OECD Data. Accessed Mar 21, 2022. https://data.oecd.org/healthres/health-spending.htm

2. Svedahl ER, Pape K, Toch-Marquardt M, et al. Increasing workload in Norwegian general practice - a qualitative study. *BMC Fam Pract.* 2019;20(1):68. 10.1186/s12875-019-0952-5

3. Hobbs FDR, Bankhead C, Mukhtar T, et al; National Institute for Health Research School for Primary Care Research. Clinical workload in UK primary care: a retrospective analysis of 100 million consultations in England, 2007-14. *Lancet.* 2016;387(10035):2323-2330. 10.1016/S0140-6736(16)00620-6

4. NHS. Long Term Conditions Compendium of Information: third edition. Published May 30, 2012. Accessed Mar 21, 2022. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/216528/dh_134486.pdf

5. Deloitte. Under pressure: the funding of patient care in general practice. Published Apr 2, 2014. Accessed Mar 21, 2022. https://www.queensroad partnership.co.uk/mf.ashx?ID=406a083a-144f-457d-b14b-aad537f67fc9

6. Smith-Bindman R, Miglioretti DL, Larson EB. Rising use of diagnostic medical imaging in a large integrated health system. *Health Aff (Millwood).* 2008; 27(6):1491-1502. 10.1377/hlthaff.27.6.1491

7. Sihvonen M, Kekki P. Unnecessary visits to health centres as perceived by the staff. *Scand J Prim Health Care.* 1990;8(4):233-239. 10.3109/0281343900 8994964

8. Renati S, Linder JA. Necessity of office visits for acute respiratory infections in primary care. *Fam Pract.* 2016;33(3):312-317. 10.1093/fampra/cmw019

9. Simpson GC, Forbes K, Teasdale E, Tyagi A, Santosh C. Impact of GP direct-access computerised tomography for the investigation of chronic daily headache. *Br J Gen Pract.* 2010;60(581):897-901. 10.3399/bjgp10X544069

10. O'Sullivan JW, Albasri A, Nicholson BD, et al. Overtesting and undertesting in primary care: a systematic review and meta-analysis. *BMJ Open.* 2018; 8(2):e018557. 10.1136/bmjopen-2017-018557

11. Anjum O, Bleeker H, Ohle R. Computed tomography for suspected pulmonary embolism results in a large number of non-significant incidental findings and follow-up investigations. *Emerg Radiol.* 2019;26(1):29-35. 10.1007/s10140-018-1641-8

12. Waterbrook AL, Manning MA, Dalen JE. The Significance of Incidental Findings on Computed Tomography of the Chest. *J Emerg Med.* 2018;55(4):503-506. 10.1016/j.jemermed.2018.06.001

13. Hawker JI, Smith S, Smith GE, et al. Trends in antibiotic prescribing in primary care for clinical syndromes subject to national recommendations to reduce antibiotic resistance, UK 1995-2011: analysis of a large database of primary care consultations. *J Antimicrob Chemother.* 2014;69(12):3423-3430. 10.1093/jac/dku291

14. Gulliford MC, Dregan A, Moore MV, et al. Continued high rates of antibiotic prescribing to adults with respiratory tract infection: survey of 568 UK general practices. *BMJ Open.* 2014;4(10):e006245. 10.1136/bmjopen-2014-006245

15. Costelloe C, Metcalfe C, Lovering A, Mant D, Hay AD. Effect of antibiotic prescribing in primary care on antimicrobial resistance in individual patients: systematic review and meta-analysis. *BMJ.* 2010;340:c2096. 10.1136/bmj.c2096

16. The ABIM foundation. Choosing wisely. DataBrief: findings from a national survey of physicians. Published 2017. Accessed Mar 21, 2022. https://www.choosingwisely.org/wp-content/uploads/2017/10/Summary-Research-Report-Survey-2017.pdf

17. Fletcher-Lartey S, Yee M, Gaarslev C, Khan R. Why do general practitioners prescribe antibiotics for upper respiratory tract infections to meet patient expectations: a mixed methods study. *BMJ Open*. 2016;6(10):e012244. 10.1136/bmjopen-2016-012244

18. Ellertsson S, Loftsson H, Sigurdsson EL. Artificial intelligence in the GPs office: a retrospective study on diagnostic accuracy. *Scand J Prim Health Care*. 2021;39(4):448-458. 10.1080/02813432.2021.1973255

19. Liang H, Tsui BY, Ni H, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med*. 2019;25(3):433-438. 10.1038/s41591-018-0335-9

20. Ribeiro AH, Ribeiro MH, Paixão GMM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun*. 2020;11(1):1760. 10.1038/s41467-020-15432-4

21. Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med*. 2019;25(1):65-69. 10.1038/s41591-018-0268-3

22. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402-2410. 10.1001/jama.2016.17216

23. Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*. 2018;172(5):1122-1131.e9. 10.1016/j.cell.2018.02.010

24. Tomita N, Cheung YY, Hassanpour S. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. *Comput Biol Med*. 2018;98:8-15. 10.1016/j.compbiomed.2018.05.011

25. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv:171105225 [cs, stat]*. Published online Dec 25, 2017. Accessed Febr 15, 2022. https://arxiv.org/abs/1711.05225

26. Teramoto A, Fujita H, Yamamuro O, Tamaki T. Automated detection of pulmonary nodules in PET/CT images: ensemble false-positive reduction using a convolutional neural network technique. *Med Phys*. 2016;43(6):2821-2827. 10.1118/1.4948498

27. Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med*. 2019;25(6):954-961. 10.1038/s41591-019-0447-x

28. Kim CK, Choi JW, Jiao Z, et al. An automated COVID-19 triage pipeline using artificial intelligence based on chest radiographs and clinical data. *NPJ Digit Med*. 2022;5(1):5. 10.1038/s41746-021-00546-w

29. Baker A, Perov Y, Middleton K, et al. A comparison of artificial intelligence and human doctors for the purpose of triage and diagnosis. *Frontiers in Artificial Intelligence*. Published 2020. Accessed Mar 21, 2022. https://www.frontiersin.org/article/10.3389/frai.2020.543405

30. Mullainathan S, Obermeyer Z. Diagnosing physician error: a machine learning approach to low-value health care. Published Aug 2019. Updated Nov 2021. Accessed Mar 21, 2022. https://click.endnote.com/viewer?doi=10.1093%2Fqje%2Fqjab046&token=WzcyMDUwOSwiMTAuMTA5My9xamUvcWphYjA0NiJd.LTNw5P_WqY_RWsjNaONiygyvKb0

31. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med*. 2015;7(299):299ra122. 10.1126/scitranslmed.aab3719

32. Hall KK, Shoemaker-Hunt S, Hoffman L, et al. *Making Healthcare Safer III: A Critical Analysis of Existing and Emerging Patient Safety Practices*. Agency for Healthcare Research and Quality; 2020. Accessed Aug 29, 2022. https://www.ncbi.nlm.nih.gov/books/NBK555525/

33. Pestotnik SL, Classen DC, Evans RS, Burke JP. Implementing antibiotic practice guidelines through computer-assisted decision support: clinical and financial outcomes. *Ann Intern Med*. 1996;124(10):884-890. 10.7326/0003-4819-124-10-199605150-00004

34. Podda M, Pisanu A, Sartelli M, et al. Diagnosis of acute appendicitis based on clinical scores: is it a myth or reality? *Acta Biomed*. 2021;92(4):e2021231. 10.23750/abm.v92i4.11666

35. Cabana MD, Rand CS, Powe NR, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. *JAMA*. 1999;282(15):1458-1465. 10.1001/jama.282.15.1458

36. Logan GS, Dawe RE, Aubrey-Bassler K, et al. Are general practitioners referring patients with low back pain for CTs appropriately according to the guidelines: a retrospective review of 3609 medical records in Newfoundland using routinely collected data. *BMC Fam Pract*. 2020;21(1):236. 10.1186/s12875-020-01308-5

37. Morgan B, Mullick S, Harper WM, Finlay DB. An audit of knee radiographs performed for general practitioners. *Br J Radiol*. 1997;70(831):256-260. 10.1259/bjr.70.831.9166050

38. Carlsen B, Glenton C, Pope C. Thou shalt versus thou shalt not: a meta-synthesis of GPs' attitudes to clinical practice guidelines. *Br J Gen Pract*. 2007; 57(545):971-978. 10.3399/096016407782604820

39. Dambha-Miller H, Everitt H, Little P. Clinical scores in primary care. *Br J Gen Pract*. 2020;70(693):163-163. 10.3399/bjgp20X708941

40. Khan MNB. Telephone consultations in primary care, how to improve their safety, effectiveness and quality. *BMJ Open Quality*. 2013;2(1):u202013.w1227. 10.1136/bmjquality.u202013.w1227

41. Graversen DS, Christensen MB, Pedersen AF, et al. Safety, efficiency and health-related quality of telephone triage conducted by general practitioners, nurses, or physicians in out-of-hours primary care: a quasi-experimental study using the Assessment of Quality in Telephone Triage (AQTT) to assess audio-recorded telephone calls. *BMC Fam Pract*. 2020;21(1):84. 10.1186/s12875-020-01122-z

42. Tenhunen H, Hirvonen P, Linna M, Halminen O, Hörhammer I. Intelligent patient flow management system at a primary healthcare center - the effect on service use and costs. *Stud Health Technol Inform*. 2018;255:142-146.

43. McKinsey and Company. Telehealth: a quarter-trillion-dollar post-COVID-19 reality? Published Jul 9, 2021. Accessed Aug 22, 2022. https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/telehealth-a-quarter-trillion-dollar-post-covid-19-reality

44. Balogh EP, Miller BT, Ball JR, et al. *Improving Diagnosis in Health Care*. National Academies Press; 2015. Accessed Mar 28, 2022. https://www.ncbi.nlm.nih.gov/books/NBK338594/

45. Hlynsson HD, Ellertsson S, Daðason JF, Sigurdsson EL, Loftsson H. Semi-self-supervised automated ICD coding. Published May 20, 2022. Accessed May 23, 2022. https://arxiv.org/abs/2205.10088

46. Lundberg S, Lee SI. A unified approach to interpreting model predictions. Published May 22, 2017. 10.48550/arXiv.1705.07874

47. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.

48. Speets AM, Hoes AW, van der Graaf Y, Kalmijn S, Sachs APE, Mali WPTM. Chest radiography and pneumonia in primary care: diagnostic yield and consequences for patient management. *Eur Respir J*. 2006;28(5):933-938. 10.1183/09031936.06.00008306

49. van Vugt S, Broekhuizen L, Zuithoff N, et al; GRACE Project Group. Incidental chest radiographic findings in adult patients with acute cough. *Ann Fam Med*. 2012;10(6):510-515. 10.1370/afm.1384

50. Havers FP, Hicks LA, Chung JR, et al. Outpatient antibiotic prescribing for acute respiratory infections during influenza seasons. *JAMA Netw Open*. 2018;1(2):e180243. 10.1001/jamanetworkopen.2018.0243

51. Wood J, Butler CC, Hood K, et al. Antibiotic prescribing for adults with acute cough/lower respiratory tract infection: congruence with guidelines. *Eur Respir J*. 2011;38(1):112-118. 10.1183/09031936.00145810

52. Diehr P, Wood RW, Bushyhead J, Krueger L, Wolcott B, Tompkins RK. Prediction of pneumonia in outpatients with acute cough—a statistical approach. *J Chronic Dis*. 1984;37(3):215-225. 10.1016/0021-9681(84)90149-8

53. Anthonisen NR, Manfreda J, Warren CPW, Hershfield ES, Harding GKM, Nelson NA. Antibiotic therapy in exacerbations of chronic obstructive pulmonary disease. *Ann Intern Med*. 1987;106(2):196-204. 10.7326/0003-4819-106-2-196