The Use of Artificial Intelligence for Diagnosis and Outcome Prediction in Primary Care

Steindór Oddur Ellertsson

Thesis for the degree of Philosophiae Doctor

February 2025

School of Health Sciences

FACULTY OF MEDICINE

UNIVERSITY OF ICELAND

The Use of Artificial Intelligence for Diagnosis and Outcome Prediction in Primary Care

Steindór Oddur Ellertsson

Thesis for the degree of Philosophiae Doctor

Supervisors

Emil Lárus Sigurðsson and Hrafn Loftsson

Doctoral committee

Emil Lárus Sigurðsson, Hrafn Loftsson, Jón Steinar Jónsson, Margrét Ólafía Tómasdóttir and Yngvi Björnsson

February 2025

School of Health Sciences

FACULTY OF MEDICINE

UNIVERSITY OF ICELAND

Notkun gervigreindar til greiningar og forspár fyrir horfur sjúklinga í heilsugæslu

Steindór Oddur Ellertsson

Ritgerð til doktorsgráðu

Leiðbeinendur

Emil Lárus Sigurðsson og Hrafn Loftsson

Doktorsnefnd

Emil Lárus Sigurðsson, Hrafn Loftsson, Jón Steinar Jónsson, Margrét Ólafía Tómasdóttir og Yngvi Björnsson

Febrúar 2025

Heilbrigðisvísindasvið

LÆKNADEILD

HÁSKÓLI ÍSLANDS

Thesis for a doctoral degree at the University of Iceland. All rights reserved. No part of this publication may be reproduced in any form without the prior permission of the copyright holder.

© Steindór Oddur Ellertsson 2025

ISBN 978-9935-9781-6-5

ORCID: 0009-0002-4719-6859

Reykjavik, Iceland 2025

Ágrip

Á undanförnum árum hefur getu gervigreindar til þess að greina og spá fyrir um útkomur sjúklinga verulega fleygt fram. Hluta af þessum framförum má þakka aðgengi að stærri rafrænum gagnasöfnum en skortur er á stórum gæðamiklum gagnasöfnum til þess að þjálfa gervigreindarlíkön á til notkunar innan heilbrigðiskerfisins. Rafrænar sjúkraskrár búa yfir gríðarmiklum gögnum og með samtengingu kerfa geta myndast fjölbreytt gagnasöfn sem má samkeyra og hugsanlega nýta til þjálfunar gervigreindarlíkana. Samantektarnótur lækna úr viðtölum við sjúklinga eru stór hluti sjúkraskrár í heilsugæslu og innihalda lýsingu á sögu sjúklings, skoðun, niðurstöðum og ályktun og áætlun læknis. Slíkar nótur ættu því að innihalda allar upplýsingar sem þörf er á til að þjálfa gervigreindarlíkön til að spá fyrir um greiningar og útkomur sjúklinga.

Markmið þessarar doktorsritgerðar var: 1) að rannsaka hvort hægt væri að nýta textagögn úr sjúkraskrám sjúklinga til þess að þjálfa gervigreindaralíkön til þess að spá fyrir um greiningar og horfur sjúklinga í heilsugæslu; 2) að skoða hvaða breytur hafa áhrif á úttak líkananna og bera saman við breytur sem læknar styðjast við; og 3) kanna frammistöðu líkananna við greiningar og forspár fyrir horfur sjúklinga í heilsugæslu.

Ritgerðin byggir á rannsóknum sem hefur verið lýst í þremur greinum—tvær fyrstu rannsóknirnar eru aftursýnar en sú þriðja er framsýn. Í fyrstu greininni var gervigreindarlíkan þjálfað til þess að spá fyrir um frumkomna höfuðverkjagreiningu út frá merktum greiningarsérkennum úr samantektarnótu læknis eftir viðtöl við sjúklinga sem fengu eina af fjórum algengustu greiningum slíkra höfuðverkja í heilsugæslu. Í greininni var frammistaða líkansins borin saman við frammistöðu þriggja sérnámslækna í heimilislækningum og þriggja sérfræðimenntaðra heimilislækna. Einnig var innri virkni líkansins metin með SHapley Additive exPlanations (SHAP) gildum sem gáfu til kynna hvaða þættir inntaksins höfðu mest áhrif á hvert úttak fyrir sig. Í annarri greininni var gervigreindarlíkan þjálfað á greiningarsérkennum úr merktum nótum frá læknum sem hittu sjúklinga í heilsugæslu sem fengu einn af eftirfarandi International Classification of Diseases (ICD) kóðum á ákveðnu tímabili: J00, J15, J20, J44 og J45. Í tveimur fyrstu greinunum var ákveðinni aðferðarfræði beitt til að fjarlægja innihaldslitlar textanótur og hámarka gæði gagnasafnsins sem líkönin læra af. Í þriðju greininni var frammistaða líkansins sem þjálfað var í rannsókn tvö metin, eftir smávægilega aðlögun, með framsýnum hætti á einni heilsugæslustöð á höfuðborgarsvæðinu á Íslandi.

Í fyrstu rannsókninni náði líkanið hærra vegnu meðaltali í næmi, jákvæðu forspárgildi og Matthews Correlation Coefficient (MCC) en vegið meðaltal læknanna. Sértæki fimm lækna af sex var hærra en líkansins. Greining á SHAP gildum sýndi að líkanið styðst við svipuð greiningarsérkenni og læknir gerir fyrir hverja greiningu. Í rannsókn tvö raðaði

líkanið sjúklingum, með einkenni um öndunarfærasýkingu, í áhættuhópa þannig að útkomur sjúklinga sem voru í lægri áhættuhópum endurspegluðu hóp sem líklega var með væg einkenni sem hefðu gengið yfir án inngrips. Engir sjúklingar í lægri áhættuhópum voru með lungnabólgu greinda af lækni eða á röntgenmynd. Í síðustu rannsókninni var einn sjúklingur greindur með lungnabólgu í lægri áhættuhópum en hann reyndist vera með eðlilega röntgemynd af lungum og því líklega rangt greindur. Allir aðrir sjúklingar með lungnabólgu á röntgenmynd voru í há-áhættu hópum. Tveir sjúklingar voru greindir með lungnakrabbamein, báðir í hæsta áhættuhóp.

Niðurstöður fyrstu rannsóknarinnar voru þær að gervigreindarlíkanið stóð sig jafn vel eða ívið betur í greiningum á frumkomnum höfuðverkjum en hóparnir tveir af sérnámslæknum og sérfræðingum í heimilislækningum. Greining á SHAP gildum leiddi í ljós að líkönin studdust við mjög svipuð greiningarsérkenni og læknar við greiningar. Niðurstöður annarrar rannsóknarinnar bentu til þess að gervigreindarlíkaninu tækist að einkennastiga sjúklinga með öndunarfærasýkingar í heilsugæslu með öruggum hætti. Líkanið raðaði sjúklingum sem reyndust raunverulega vera með alvarlega öndunarfærasjúkdóma í há-áhættu hópa en sjúklingum með vægari einkenni í lág-áhættu hópa. Þriðja rannsóknin sýndi fram á að líkaninu virtist takast að einkennastiga sjúklinga í raunverulegum klínískum aðstæðum þannig að sjúklingar með alvarlegar öndunarfærasýkingar voru flokkaðir í háa áhættu en sjúklingar með væg einkenni í lága áhættu. Tveir sjúklingar sem læknar greindu með lungnabólgu en líkanið flokkaði í lág-áhættu hóp reyndust vera með eðlilegar lungnamyndir.

Niðurstöður doktorsritgerðarinnar benda til þess að gervigreindarlíkön, sem þjálfuð eru á greiningarsérkennum úr samantektarnótum lækna, geti haft verulegt notagildi í heilsugæslu.

Lykilorð: Heilsugæsla, Gervigreind, Einkennastigun, Öndunarfærasýkingar, Frumkomnir Höfuðverkir

Abstract

In recent years, the ability of Artificial Intelligence (AI) to analyze and predict various variables within medicine has advanced significantly. Part of this progress can be attributed to the availability of larger electronic datasets; however, there is still a shortage of high-quality data for training AI models for use within the healthcare system. Electronic health records contain vast amounts of data, and with increased interoperability, diverse datasets are emerging that can be cross-referenced and potentially used to train AI models. Doctors' summary notes from patient interviews are a major component of medical records and include descriptions of the patient's medical history, examination, findings, and the doctor's assessment and plan. These notes should, therefore, contain all the necessary information to train AI models to predict patient diagnoses and outcomes.

The objective of this doctoral thesis was: 1) to investigate whether doctors' summary notes from patient medical records could be used to train AI models to predict diagnoses and outcomes for patients in primary care; and 2) to study which factors the models use to reach a conclusion and compare these to the elements doctors rely on.

The thesis is based on studies described in three papers—the first two were retrospective, and the third was prospective. In the first study, an AI model was trained to predict primary headache diagnoses based on labeled diagnostic features from doctors' summary notes following patient interviews in primary care, which resulted in one of four common primary headache diagnoses. The study compared the model's performance to that of three resident physicians and three family medicine specialists. The model's internal functionality was also evaluated using Shapley Additive Explanations (SHAP) values, which indicate which input features have the most significant impact on each output. In the second paper, an AI model was trained on Clinical Features (CFs) from labeled notes of doctors who treated primary care patients diagnosed with one of the following respiratory symptom codes over a specific period: J00, J15, J20, J44, and J45. In both the first and second studies, a methodology was applied to remove low-content text notes and maximize the quality of the dataset from which the models learned. The third study evaluated the model trained in the second study, with minor adjustments, using a prospective approach in a single primary care clinic in Iceland's capital area.

In the first study, the model achieved a higher weighted average sensitivity, positive predictive value, and Matthews Correlation Coefficient (MCC) than the weighted average of the doctors. The specificity of five out of six doctors was higher than that of the model. SHAP value analysis indicated that the model relies on similar diagnostic

features as doctors for each diagnosis. In the second study, the model categorized patients into risk groups, where the outcomes of patients in lower-risk groups reflected those likely to have mild symptoms that resolve without intervention. No cases of pneumonia or patients with lung infiltrates on Chest X-Rays (CXRs) were detected in the lower-risk groups. In the third study, one patient with pneumonia was categorized in a lower-risk group but was later found to have a normal lung CXR, suggesting a likely misdiagnosis. All other patients with pneumonia on CXR were in high-risk groups. Two patients were diagnosed with lung cancer, both in the highest risk group.

The results of the first study indicated that the AI model performed as well or slightly better than the groups of GP trainees and GP specialists in diagnosing primary headaches. SHAP value analysis showed that the models rely on similar clinical features to doctors when making diagnoses. The results of the second study suggested that the AI model could safely triage primary care patients with respiratory infections. The model categorized patients with truly severe respiratory conditions into high-risk groups, while those with milder symptoms were placed in low-risk groups. The third study demonstrated that the model could stratify patients in real clinical settings so that patients with severe respiratory conditions were categorized as high risk, while those with mild symptoms were categorized as low risk. Two patients diagnosed with pneumonia by doctors, whom the model classified as low risk, were subsequently found to have normal lung images.

The doctoral thesis concludes that AI models trained on clinical features extracted from physicians' summary notes can have significant utility in primary healthcare.

Keywords: Primary Health Care, Artificial Intelligence, Triage, Respiratory Tract Infections, Primary Headaches

Acknowledgments

As this educational journey concludes, I am deeply grateful to many individuals whose support and guidance were crucial and inspiring in the development of this doctoral thesis.

With deep respect and gratitude, I thank both of my supervisors, Emil Lárus Sigurðsson and Hrafn Loftsson, for the opportunity to work on this project under their guidance. Their patience and motivation got me through the challenging periods. I also thank them for their friendship and all the good times we have shared over the years. Under their supervision, I have grown immensely as a scientist, learning the intricacies of research and the resilience needed to navigate its challenges.

I sincerely thank Jón Steinar Jónsson, Margrét Ólafía Tómasdóttir, and Yngvi Björnsson for their valuable input and support as members of my doctoral committee.

I am deeply grateful to both old and new friends who participated in this journey with me. My heartfelt thanks go to Stefán Ólafsson for his unwavering support and invaluable advice throughout this PhD process. I also wish to thank Vilhjálmur Steingrímsson, a long-time friend, for his thoughtful scientific discussions, motivational support, and the many badminton matches we've shared over the years. Special thanks go to Guðmundur Haukur Jörgensen, a colleague and friend, for his meticulous proofreading of all the papers in this thesis and the insightful scientific discussions that have greatly enriched this work.

The studies presented in this thesis were conducted within the Primary Health Care Service of the Capital Area and Reykjavik University. I am deeply grateful for the workspace provided at both locations and my colleagues' educational and emotional support. I extend special thanks to my colleagues at Heilsugæslan Miðbær. Their invaluable assistance in recruiting participants and the camaraderie shared during coffee and lunch breaks provided essential support and much-needed laughter, which got me through the heavier periods of this journey.

I am profoundly grateful for the unwavering love, constant support, and encouragement of my parents and siblings throughout the years. My deepest thanks go to my parents, Ellert Kristján Steindórsson and Rannveig Alma Einarsdóttir,

whose belief in me has been a cornerstone of this journey. I am incredibly fortunate to have such a large and supportive family, without whom this thesis would not have been possible.

Finally, my deepest thanks go to the most important people in my life: my wife, Olga Hrönn Jónsdóttir, and our children, Lísa, Björt, and Styrkár. Your boundless love, patience, and unwavering belief in me have been my greatest motivation. You are my everything, and you make it all worth it. To my children, I want you to know that anything is possible if you are willing to work hard and stay determined.

This work was funded by the Scientific Fund of the Icelandic College of Family Physicians and the Icelandic Technology Fund (RANNÍS).

Contents

List of Abbreviations

AI	Artificial Intelligence
AUC	Area Under the Curve
AUROC	Area Under the Receiving Operating Characteristic curve
AWS	Amazon Web Services
BUN	Blood Urea Nitrogen
CAP	Community Acquired Pneumonia
CDSS	Clinical Decision Support System
CF	Clinical Feature
CFEM	Clinical Feature Extraction Model
CH	Cluster Headache
CI	Confidence Interval
CNN	Convolutional Neural Network
COPD	Chronic Obstructive Pulmonary Disease
CRP	C-Reactive Protein
CT	Computed Tomography
CTM	Clinical Triage Model
CTN	Clinical Text Note
CV	Cross Validation
CXR	Chest X-Ray
DNN	Deep Neural Network
DL	Deep Learning
ED	Emergency Department

EHR	Electronic Health Record
FDA	Food and Drug Administration
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GAS	Group A Streptococcus
GLUE	General Language Understanding Evaluation
GP	General Practitioner
GPS	Global Positioning System
GWh	Gigawatt hours
HAP	Hospital Acquired Pneumonia
ICD	International Classification of Diseases
LASSO	Least Absolute Shrinkage and Selection Operator
LLM	Large Language Model
LR	Logistic Regression
LRTI	Lower Respiratory Tract Infection
MA+	Migraine with Aura
MA-	Migraine without Aura
ML	Machine Learning
MRI	Magnetic Resonance Image
MRSA	Methicillin Resistant Staphylococcus Aureus
NLP	Natural Language Processing
OR	Odds Ratio
PC	Primary Care
PHCCA	Primary Health Care of the Capital Area
PPV	Positive Predictive Value

RF	Random Forests
ROC	Receiver Operating Characteristic
RSTM	Respiratory Symptom Triage Model
SHAP	Shapley Additive Explanation
SOTA	State Of The Art
STRAMA	Swedish Strategic Programme for the Rational use of Antimicrobial Agents
SVM	Support Vector Machines
t-SNE	t-distributed Stochastic Neighbor Embedding
TN	True Negative
TP	True Positive
TPR	True Positive Rate
TTH	Tension-Type Headache
USMLE	United States Medical Licensing Examination

List of Figures

List of Tables

List of original papers

This thesis is based on the following original research papers, which will be referred to in the text by their Roman numerals (I, II, and III):

I. Ellertsson S, Loftsson H, Sigurdsson EL. Artificial intelligence in the GPs office: a retrospective study on diagnostic accuracy. *Scandinavian Journal of Primary Health Care*. 2021;0(0):1-11. doi:10.1080/02813432.2021.1973255

II. Ellertsson S, Hlynsson HD, Loftsson H, Sigurðsson EL. Triaging Patients With Artificial Intelligence for Respiratory Symptoms in Primary Care to Improve Patient Outcomes: A Retrospective Diagnostic Accuracy Study. *The Annals of Family Medicine*. 2023;21(3):240-248. doi:10.1370/afm.2970

III. Ellertsson S, Loftsson H, Sigurðsson EL. Can Artificial Intelligence Save Resources in Primary Care? A Prospective Study. *Submitted and in the review process*.

All papers are reprinted with the permission of the publishers.

Declaration of contribution

The doctoral student, Steindór Oddur Ellertsson, planned the research work for papers I, II, and III in cooperation with his supervisors. The doctoral student is the first author of all papers, applied for the required ethical and research approval, and performed the statistical analyses in cooperation with the co-authors. The doctoral student wrote this thesis under the guidance of his supervisors and the doctoral committee.

1 Introduction

Artificial Intelligence (AI) has rapidly evolved over the past few decades, impacting all areas of society, including medicine. This is especially true in Machine Learning (ML), a subfield of AI where algorithms learn from data without being explicitly programmed. Since the 2010s, Deep Learning (DL), a branch of ML, has advanced rapidly, enabling more sophisticated algorithms with more significant impact on medicine than ever before. However, the adoption of ML in Primary Care (PC) is still limited[1], and research indicates that the number of authors from PC publishing in the field of medical ML is relatively low[2]. Since PC is usually the first point of contact with the healthcare system, the downstream impact of using ML algorithms in PC could have the most significant effect across the entire system. Yet the development and study of ML in the field remain underexplored.

1.1 Artificial Intelligence

The concept of AI is not new. The idea of a machine capable of independent action dates back to ancient times as a part of philosophy and mythology, where it was often depicted as creations of gods or legendary inventors[3]. Historians trace the idea of automatic machines back to the Middle Ages when such constructs were referred to as "automatons"[4] and were designed to follow a sequence of operations or predetermined instructions to carry out a task. In 1950, the English mathematician Alan Turing, in his seminal paper[5], wondered if machines could think and proposed ways to test their intelligence. The term AI was first introduced in 1956 by the computer scientist John McCarthy and his co-authors in a paper created at a summer workshop in Dartmouth[6], marking the inception of a crucial subfield of computer science. No uniform definition of AI exists. However, in this thesis, we define AI as a field that aims to develop systems that mimic human intelligence, encompassing tasks such as learning, reasoning, problem-solving, perception, and language understanding. This includes mimicking cognitive abilities such as pattern recognition, decision-making, language comprehension, adapting to new information, and improving over time. AI comprises several subfields, and Figure 1 outlines the key ones along with their components.

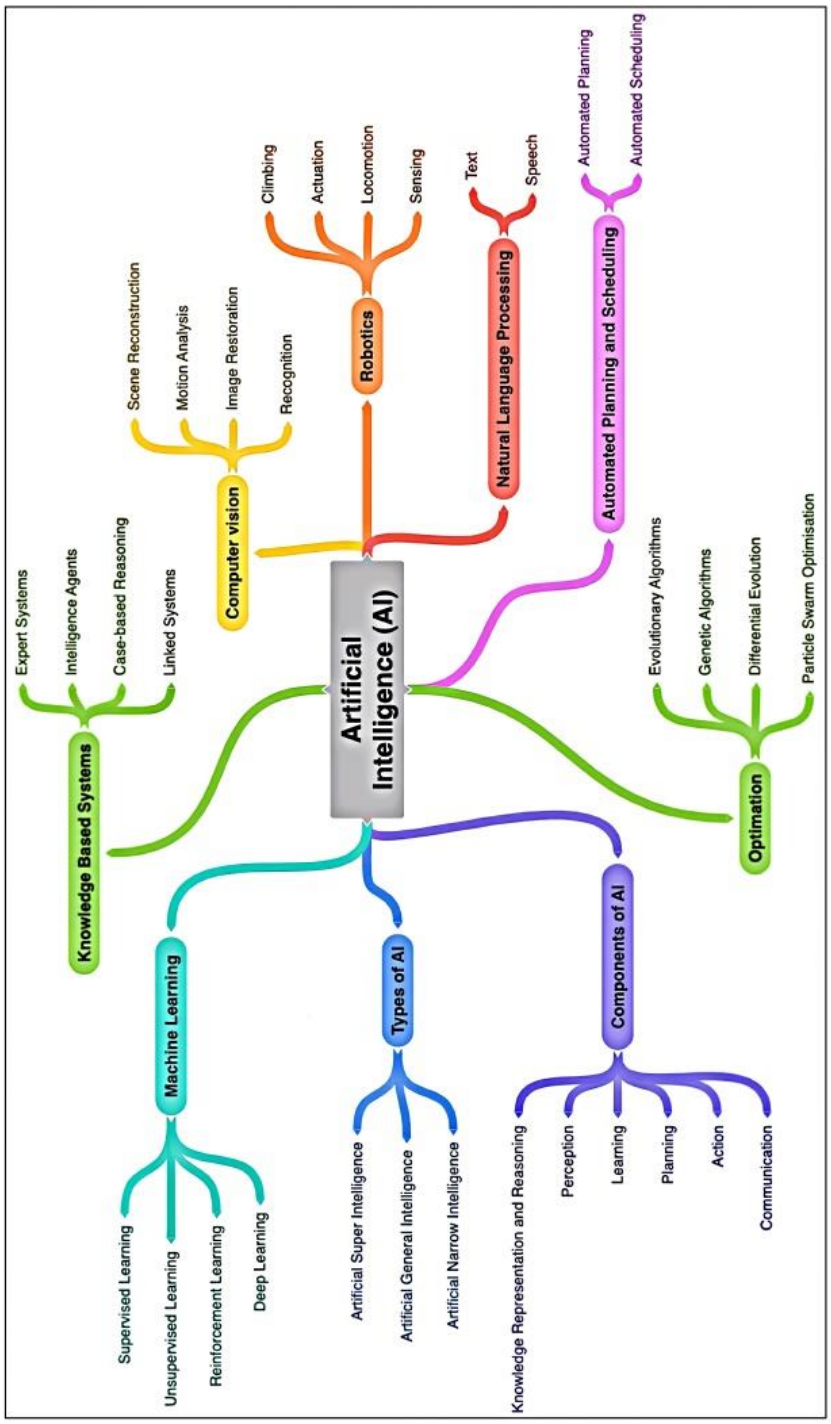


Figure 1. An overview of the various subfields and components of AI. The image is reproduced with the permission of its authors[1].

1. Regona M, Yigitcanlar T, Xia B, Li RYM. Opportunities and Adoption Challenges of AI in the Construction Industry: A PRISMA Review. Journal of Open Innovation Technology Market and Complexity. 2022 Feb 28;8:45.

1.2 Machine Learning

With the digital transformation of medical records and the growth of medical databases in the 1990s, AI began to integrate more deeply with actual patient data. During this period, the emphasis on AI shifted towards data-driven approaches, moving away from reliance on expert systems as they were unable to learn, expensive to maintain, and often gave inappropriate outputs when given unusual inputs[3(p435)]. An expert system is a computer program that uses rules and knowledge databases to solve complex problems in a specific domain by reasoning through the provided information. They are designed to mimic the decision-making ability of a human expert, contrastive to an ML algorithm, which learns directly from data. The development of ML algorithms allowed for the analysis of large datasets, leading to the discovery of previously unknown patterns and correlations in patient information, lab results, and treatment outcomes. The 1990s and 2000s saw significant advancements in ML, with the development of more sophisticated algorithms and increased computational power[4–6].

The DL revolution marked the 2010s. DL and neural networks became powerful tools for handling vast amounts of data with many parameters and surpassed previous benchmarks on many tasks, some by a large margin. This greatly impacted medical research, enabling the development of AI that could match or outperform human experts in specific diagnostic tasks[7–12]. In recent years, significant advances have been made within the field of Natural Language Processing (NLP), where Large Language Models (LLMs) have reached State-Of-The-Art (SOTA) performance each year on most NLP benchmarks[13]. NLP is poised to have a significant impact on how clinical staff interact with Electronic Health Record (EHR) systems, possibly significantly reducing the burden of paper- and administrative work.

ML consists of various components and subfields, each with distinct learning methods and model types. The sections below focus on methods used in our studies and supervised and unsupervised DL and ML algorithms, such as Random Forests (RF) and Logistic Regression (LR). We also used an LLM (in study II), which falls in the text category of NLP, although training that model was not a part of this thesis.

1.2.1 Multivariate Logistic Regression

Multivariate LR is a statistical method used to predict a binary outcome. An example of a binary outcome is a yes or no answer whether a patient has a disease. It extends a simple LR, which uses only one variable, by utilizing multiple predictors that can be continuous or discrete. The model uses a mathematical function to transform the probability of an outcome into a linear equation. It has been the

predominant model for binary regression since the 1970s[14]. LR assumes a linear relationship between the dependent and the independent variables and requires feature scaling for optimal performance. It is less computationally expensive to train compared to many other ML algorithms. LR offers high interpretability, and the model's coefficients show the effects of each predictor on the output. When training an LR model, the training data cannot contain missing values, and they must be imputed by using specific statistical methods, such as calculating an average of the non-missing values.

1.2.2 Random Forests

Random Forests (RF) is an ensemble learning method for classification and regression that combines multiple decision trees to improve predictive accuracy and control overfitting. During training, it builds a large number of decision trees and outputs the mean prediction in regression for each tree. By creating each tree using a random subset of data and features, RF reduces model variance and enhances robustness. It is resistant to noise and outliers and handles high-dimensional data and complex feature interactions, providing built-in feature importance measures. It can be computationally intensive and less interpretable than simpler models, but its ability to handle diverse datasets and provide reliable predictions makes it widely used in ML applications. RF is a robust method that handles missing values and non-linear relationships. There is no need for feature scaling, and it often handles imbalanced datasets well.

1.2.3 Deep Learning

In recent decades, advancements in AI have been driven primarily by ML and DL. Neural networks are computational models inspired by the human brain, designed to recognize patterns, learn from data, and make predictions. Large neural networks are referred to as "deep", hence the name Deep Neural Networks (DNNs). DNNs are simply large neural networks. These multi-layered networks facilitate the intricate, hierarchical processing of data, granting the model the ability to learn from a diverse spectrum of features and patterns at various levels of abstraction. DL has made significant strides in areas requiring high-dimensional data analysis, such as image[15] and text analysis[16] and speech recognition[17]. These capabilities are especially relevant and applicable in the realm of medical applications.

1.2.4 Natural Language Processing

NLP is a branch of AI that enables machines to analyze, interpret, and generate human language[18]. This field has witnessed significant advancements in recent years

thanks to the development of LLMs and the transformer[19] architecture. Such progress is prominently reflected in evaluation benchmarks like the General Language Understanding Evaluation (GLUE)[20]. In these assessments, NLP models have demonstrated capabilities that match or surpass human proficiency in various tasks. LLMs, a specialized subset of generative AI within the broader field of DL, have played a pivotal role in these advancements. Generative AI consists of AI models designed to generate new content or data that closely resembles but is not identical to the data they were trained on. Such models are often used to create images, videos, music, or text that mimic human creativity.

LLMs undergo training in two distinct phases. First, they undergo pre-training, an unsupervised process where the model learns from a vast range of textual data without specific guidance. Then secondly, they are fine-tuned in a supervised manner, where they are specifically trained on targeted downstream tasks with labeled data (correct output). This extensive training enables LLMs to handle a wide range of intricate language-based tasks adeptly. If specifically fine-tuned, they are, for example, capable of understanding and effectively responding to questions posed in natural language, showcasing their advanced comprehension skills. These advancements in NLP signify a leap forward in how machines understand and interact using human language, opening new avenues for application in various sectors and enhancing machine-human interactions. As a support tool, they seem to reduce the time required to solve cognitive writing tasks, such as programming, by 55%[21], reduce frustration, and, at the same time, enhance job satisfaction. Recently, specialized LLMs trained on medical texts have been developed. Notable examples include Med-PaLM[22] and its successor, Med-PaLM 2. According to its developers, Med-PaLM 2 achieved an 85% accuracy score on a question-answering dataset modeled after the U.S. Medical Licensing Examination (USMLE)[23]. Additionally, OpenAI's ChatGPT (GPT-4) represents another significant advancement in this field[24]. ChatGPT has achieved a passing score on the USMLE[25] and has been tested on multiple clinical tasks with various results[26]. It is important to note that these results reflect its performance without fine-tuning for a specific task, which generally improves model performance by a large margin.

LLMs appear to be quite effective in encoding clinical knowledge, but various issues remain that must be addressed before these models can be used safely within clinical settings[22]. One of these issues is their dynamic output capability, which is also the main strength of LLMs. This capability can be a significant drawback in clinical settings since consistency is a key property of medical devices. Other issues include model interpretability, fairness, and biases. LLMs as a group include the biggest AI models in the world today and, as such, become notoriously difficult to interpret. LLMs also tend to hallucinate, where they generate plausible-sounding information that is factually incorrect or nonsensical[27]. Some LLMs, such

as Llama-3[28] and Vicuna[29], are open source, while others, like GPT-4[30], are proprietary. Independent researchers have limited access to evaluate the inner workings of proprietary models. Their performance on clinical questions appears superior to open-source models[31], probably because of more extensive fine-tuning processes.

LLMs represent a unique category in comparison to other ML models, particularly regarding scale, capabilities, and potential influence. These models are notably large, often featuring tens of billions to hundreds of billions of parameters, leading to substantial hardware demands during training and inference. Unlike deep learning models designed for narrow medical applications, LLMs possess a broad range of uses, and their societal influence is likely to be very broad. The complexity of interpreting these models poses a challenge for regulatory bodies like the FDA, which must navigate the difficult task of effectively validating them for use in clinical settings[32].

1.2.5 Learning methods

Within ML, models can be trained in different ways based on how data is processed during the training process. This section discusses how the ML models in this thesis were trained. ML consists of algorithms that learn directly from data and thus utilize data-driven methods.

Supervised learning is an ML learning method where algorithms learn from labeled data. The algorithm learns to map from the input data or features to the corresponding dependent variable or labels. The aim is to create a generalized predictive algorithm that performs well on unseen data from a similar distribution as the labeled data. Examples of algorithms that can be trained with supervised learning methods are linear regression, logistic regression, decision trees, Support Vector Machines (SVM), and DNNs. Supervised learning is widely used for tasks like classification and linear regression.

Unsupervised learning is an ML learning method where algorithms are trained on unlabeled data. In this approach, the model learns to identify patterns and relationships within the data without explicit guidance about the correct output. The training data consists only of input data or features, and the algorithm is tasked with some training objective where it might discover hidden patterns, dimensionality reductions, or groupings. Unsupervised learning algorithms include K-means, hierarchical clustering, t-distributed Stochastic Neighbor Embedding (t-SNE), and principal component analysis. DNNs can also be trained in an unsupervised manner. For example, the pretraining of LLMs consists of unsupervised learning before they are fine-tuned, usually with supervised learning.

1.2.6 Ethical Challenges of Artificial Intelligence

Addressing AI's ethical challenges is critical as AI systems are increasingly integrated into various aspects of society. Bias can be defined roughly as a difference in performance between subpopulations for a predictive task. An example would be an algorithm showing different diagnostic performances for Caucasians compared to other racial or ethnic groups. Bias and fairness are closely related concepts, particularly in ML and AI systems. Bias refers to systematic discrepancies in algorithmic performance across different demographic groups, while fairness focuses on minimizing these discrepancies to ensure that all groups are treated equitably. The interpretability of ML models is crucial in this context to assess whether the predictions are reliable and fair, and they can identify potential biases that may impact patient care. The energy consumption of ever-growing DL models is also of concern.

1.2.6.1 Interpretability of Artificial Intelligence Models

The interpretability of AI models is a significant challenge, particularly for complex models like DNNs, which often function as "black boxes". In the context of DNNs, the term black box describes how they process input data to produce outputs without providing clear insight into the internal logic or decision-making process. The internal workings of the models are not visible, understandable, or accessible. This lack of transparency raises concerns in high-stakes applications, such as healthcare, where understanding the decision-making process is crucial for trust, accountability, and compliance with ethical standards. Interpreting AI models is particularly important in healthcare for several reasons. We must enable doctors and patients to make informed decisions, which is only possible with model transparency. The same can be said for safety and accountability when ML models aid in medical diagnoses or forecast patient outcomes. Legal and ethical views must also be considered. Interpretability ensures compliance with legal standards and ethical guidelines, emphasizing informed consent and patient autonomy.

1.2.6.2 Bias and Fairness

Ensuring that ML models are fair and unbiased is a significant challenge. Multiple examples of biased algorithms have been deployed within the legal system[33], healthcare[34] , and other industries[35]. Removing bias from algorithms is essential to prevent algorithmic output based on prejudiced assumptions that unfairly affect certain groups. Factors such as gender, race, and socioeconomic status may lead to unintended discrimination if used in predictive models without bias mitigation. There are several essential reasons to strive to ensure unbiased and fair ML algorithms. Ensuring equality among decisions made by algorithms is paramount to ensure all people affected by such algorithmic decisions are treated similarly

despite differences in age, gender, and background. Establishing trust in healthcare algorithms is crucial, and a key aspect of this involves ensuring fairness and lack of bias in algorithms utilized within medical environments. Bias in these algorithms can lead to incorrect diagnoses, inappropriate treatment suggestions, or misleading risk evaluations, ultimately harming patient outcomes. Medical decisions should rely on objective, evidence-based data, whereas biased AI systems, which may depend on flawed information, can cloud these objective insights. As the complexity and size of models grow, making them harder to interpret, it becomes increasingly difficult to identify and reduce bias. Therefore, there is a need for a heightened focus on meticulously designed clinical studies that examine algorithmic fairness and bias.

1.2.6.3 Energy consumption

DNNs, compared to traditional ML methods, require more extensive hardware for training and inference. After the takeoff of ChatGPT and similar services, the race to create larger models trained on more data has intensified, and larger models need more computational power. For context, in 2024, training a single 200 billion parameter LLM on Amazon Web Services (AWS) hardware consumed around 11.9 Gigawatt hours (GWh)[36]. That is a similar amount needed to power more than 3,000 homes in Iceland for a year[37]. No reliable methods exist to estimate the energy usage of an AI model during training or inference. The companies that own and run the best-performing LLMs today, which are also probably the largest ones, do not disclose the size of their models or energy consumption. In November 2024, ChatGPT had 180 million users, and it was the fastest-growing service in history, taking only 5 days to surpass 1 million users[38]. AI researchers are optimizing hardware, models, and training processes to lower power consumption. Yet, with the ongoing trend of achieving better performance through larger models and increased data and the uncertainty of when performance improvements will reach their limits, it is evident that the AI industry will continue to consume more resources in the coming years.

1.3 Artificial Intelligence in Healthcare

AI within healthcare needs dedicated subsections due to its unique applications and challenges in the field. Its potential to transform diagnostics, patient care, and administrative efficiency is immense, yet healthcare demands strict adherence to privacy, security, and high ethical standards. Moreover, integrating AI into healthcare workflows requires careful consideration of clinical accuracy, patient safety, and regulatory compliance. These factors set healthcare apart from other industries, making AI implementation more complex but, at the same time, more impactful.

1.3.1 A short historical overview of Artificial Intelligence in Healthcare

Initially, AI's application in medicine predominantly revolved around rule-based systems. A notable early example is the Dendral project in the 1960s[39], which was instrumental in analyzing mass spectrometry data in organic chemistry and set the foundation for AI's role in diagnostic systems. MYCIN, a pioneering system developed at Stanford University, followed in the 1970s[40]. MYCIN could diagnose bacterial infections and suggest appropriate antibiotics. The subsequent decades, particularly the 1970s and 1980s, witnessed the emergence of expert systems, such as CADUCEUS[41] and ONCOCIN[42]. CADUCEUS was developed at the University of Pittsburgh in the late 1970s and focused on internal medicine. ONCOCIN was developed at Stanford in the early 1980s and was designed to guide the management of cancer patients, particularly in the use of chemotherapy protocols. These systems, designed to replicate the decision-making process of human experts, aimed to encapsulate specialized knowledge within specific medical domains to assist in diagnosing and recommending treatments. They demonstrated the potential of AI in assisting with diagnosis and treatment decisions. However, their adoption in clinical practice was limited due to technological limitations, lack of integration with other systems, and concerns about accuracy and reliability.

With a general shift of the field from rule-based systems towards ML in the 1990s, AI in medicine began transitioning to data-driven methods as well. Within medicine, the advent of ML-enabled AI to analyze large and complex patient datasets to forecast patient outcomes. Predictive models started to play a role in identifying patients at risk for sepsis[43,44], diabetes[45,46], and heart disease[47,48]. The increasing adoption of EHRs in the 2000s and 2010s gave AI models access to larger datasets. With EHR systems with more interconnected databases, AI models could suddenly leverage information from various resources, such as lab results, medical history, imaging, and patient demographics. Digitation of image data had broad effects, as medical imaging was one of the early applications of ML in medicine. Statistical models and neural networks were developed to assist in tasks like image recognition and diagnosis, where they could be trained on large volumes of radiology images. For example, neural networks trained on X-rays and Magnetic Resonance Image (MRI) scans could help radiologists identify abnormalities, such as tumors, fractures, or lesions, with increasing accuracy[7,11,12]. This ability to learn from images and improve performance over time marked a significant advancement over rule-based systems limited by the rules they were programmed to follow.

However, integrating AI with EHR systems has faced significant challenges. One major obstacle is the completeness and quality of data across systems. Data is inconsistent and often missing, and differences in coding and variables within systems make it challenging to create reliable AI models. Additionally, few EHR systems were designed with AI in mind, making integrating AI models into the workflow difficult.

1.3.2 The current state of Artificial Intelligence in Healthcare

The development of algorithms within medicine is accelerating, as evidenced by the sharp increase in scientific publications on AI models featured in peer-reviewed medical journals in recent years[49]. AI has been making headlines in the medical field by showing similar or better performance than a human specialist on a range of diagnostic tasks within multiple specialties[7–9,50–52] , and it is a potential support tool for radiologists[10–12,53]. Many of these studies are retrospective, and the models are tested on pre-annotated datasets in controlled environments. There is still quite a gap in the literature regarding prospective clinical studies, as demonstrated by findings that only 4% of AI models approved by the Food and Drug Administration (FDA) between 2015 and 2020 were backed up by prospective data[54]. Furthermore, criticism has been raised on the validation process for AI models[55,56], highlighting the challenges regulators face when approving AI models as medical devices.

As of October 2023, the FDA had approved 692 medical devices, with 76% of these approvals within radiology[57]. This pronounced inclination towards radiology is primarily attributed to the advancements in Convolutional Neural Networks (CNNs). CNNs, a type of DL network, were among the first to gain significant traction during the early 2010s, marking the beginning of the DL era. As a result, image analysis algorithms are furthest along the technology maturity axis compared to other types of DL algorithms within medicine. This has led to such algorithms being among the first to witness a significant integration and application within healthcare. This trend is reflected in the noticeable bias towards radiology in the number of FDA-approved medical devices and the volume of clinical studies initiated and undertaken in recent years[57,58].

1.3.2.1 LLMs

LLMs are DL models designed to predict the next word in a sequence based on the preceding context. These models leverage vast datasets and complex architectures to generate coherent and contextually appropriate text, enabling a wide range of NLP applications. LLMs present unique validation challenges in healthcare due to their dynamic text-generation capabilities. While this feature is one of their main strengths, it can also be a drawback in medical contexts. This means that when LLMs

receive the same input repeated N times, they produce N different outputs. Although some outputs might be conceptually similar, current training methods do not guarantee consistency. Historically, clinical validation of algorithms has focused on models with static outputs. Therefore, governing bodies are now faced with determining appropriate regulatory frameworks for LLMs within a healthcare context, ensuring both efficacy and reliability. As of October 19th, 2023, no AI device powered by LLMs had been approved by the FDA[57].

Another possible use case for LLMs is their possibility of reducing administration- and paperwork within healthcare. As such tools, LLMs have the potential to positively impact the clinical work environment, as paperwork is a leading contributor to physician burnout[59]. By leveraging their advanced text analysis and generation capabilities, LLMs can reduce the workload on clinical staff, who can then focus on reviewing, editing, and confirming the generated content rather than creating it from scratch. Research shows that ChatGPT can create answers to questions from actual patients, deemed more preferable by healthcare professionals in 79% of cases than actual answers[60]. The model also produced a much higher volume of highly empathetic answers than humans. Patients appear to prefer AI-generated responses as well[61]. These studies do not take hallucination into account, which seems to be a significant problem, at least when using ChatGPT, which is not finetuned[62]. Non-peer-reviewed data has shown that LLMs reduce the time spent on coding tasks by 55%, where they pre-generate code for developers while reducing cognitive load and improving job satisfaction[21].

1.3.2.2 ML models

Data analysis methods can broadly be categorized as descriptive, exploratory, inferential, predictive, and causal[63]. In ML, statistical methods are used to train models using observed data to either infer or predict future data and would fall into the inferential or predictive groups. ML models have been used for decades in medicine and research for various purposes, such as simple diagnostic rules for disease diagnosis or risk prediction, predicting patient outcomes, and analyzing medical data. These models, which include linear regression and decision trees, are valuable for their interpretability and ease of implementation. They provide essential insights into medical data, supporting clinical decision-making and foundational research. A good example of such models is all the multivariate LR models[47,64] developed from data gathered from the Framingham Heart Study[65]. These models serve as essential tools in clinical practice, enabling physicians to calculate objective risk scores that inform treatment decisions. A collection of similar risk prediction tools with explanations and guidelines is conveniently available on a single webpage for easy access[66].

1.3.2.3 Computer Vision

Computer vision models gained the most attention as the DL era began. DL models, like CNNs, achieved SOTA on numerous medical diagnostic tasks, often surpassing human specialists, though mainly in controlled retrospective research settings[7–12]. Much of this research has historically been conducted retrospectively in isolated, controlled environments without assessing model performance in real clinical settings. This is now changing as a growing number of clinical trials on CNNs are being launched. CNNs are currently the most advanced DL models regarding clinical applicability, as they demonstrated potential earlier than models like LLMs. Computer vision algorithms have been studied more extensively and have progressed further in the clinical validation process than many other ML algorithms, as evidenced by the significant proportion of FDA-approved AI medical devices based on computer vision models[67].

1.3.2.4 Clinical Triage Models (CTMs)

CTMs are ML models designed to optimize patient assessment and decision-making processes in healthcare settings. CTMs are often integrated into Clinical Decision Support Systems (CDSSs) in a clinical environment. Such systems have been researched in various settings, such as risk prediction of sepsis[68,69], emergency admissions[70], and mortality[71]. Similarly, AI-driven predictive modeling using EHRs has demonstrated superior performance over traditional predictive models in forecasting in-hospital mortality, 30-day unplanned readmission, prolonged length of stay, and all of a patient's final discharge diagnoses[72]. Most of this research focuses on CTMs, which utilize available data from EHRs directly without the patient in the loop. CDSSs in use today are mainly rule-based and created from expert knowledge and clinical guidelines, and results indicate that CTMs can improve upon current predictive methods used in medicine[73].

CTMs that include information retrieved directly from the patient remain less extensively investigated. Such models could enhance diagnosis, treatment, and patient flow in outpatient settings, where web integration enables patients to input their data before seeking medical assistance.

1.3.3 The cooperation of AI and humans

Things become more complicated as humans interact more with AI models within the healthcare system. While AI has shown great performance in isolated clinical tasks, more systematic research is needed on how AI tools are used alongside healthcare professionals in real-world settings and their effect on actual patient outcomes. In healthcare, we must consider both the patients and the clinical workers.

More research has been done on the clinical worker side. Increasing evidence shows that human-AI hybrid workflows outperform human or AI working independently[74]. Such cooperation might ensure an expert review of the outputs of AI models as we gather more data on their safety, possibly making this transition of incorporating AI into medicine safer. Humans also appear to over-rely on AI systems[75], where users trust the output of an AI model without assessing the reliability of the output. Trusting model output blindly might increase the probability of humans losing skills that are delegated to AI, much like how the widespread use of Global Positioning System (GPS) systems appears to diminish our spatial cognition and navigation abilities[76]. Algorithmic aversion or AI under-reliance is also a problem where clinical workers avoid algorithmic counsel after seeing them make a mistake[77]. Healthcare professionals struggle with knowing when their own conclusions might be better than those of an AI system. Making the models explain themselves has shown mixed results[78]. There is increasing research on methods to mitigate these problems, but the cooperation between humans and AI in clinical settings must be better documented.

2 Aims

This thesis aimed to assess whether ML models can be effectively trained using data extracted from Clinical Text Notes (CTNs) written by human physicians after a patient visit. Since there is a shortage of high-quality data for training ML models in healthcare, the abundance of data within Electronic Health Records (EHRs) offers a potential solution. However, the viability of this approach still needs to be determined. Other research has shown potential in using CTN data for ML training[79], but more in-depth studies are required. Such models could have significant applications in PC as it is the first patient contact point in the healthcare system.

2.1 Study I

The primary objective of Study I was to evaluate the practicality of training an ML classifier on annotated Clinical Features (CFs) extracted from CTNs authored by physicians to forecast primary headache diagnoses in General Practice (GP). A secondary aim was to investigate the inner workings of such model using Shapley Additive Explanation (SHAP) values.

2.2 Study II

Study II aimed to expand upon the methodology established in Study I by applying it to a different group of diagnoses and larger patient cohorts while also employing alternative metrics to assess the performance of the models. Patient outcomes were used to evaluate model performance instead of benchmarking against physicians. The study explored the feasibility of training an ML model using only CFs from patients' symptomatic history and to assess if the model could triage individuals with certain acute or subacute respiratory symptoms effectively and identify those with self-resolving symptoms before they sought assistance in primary care settings. We refer to the ML model trained in this study as the Respiratory Symptom Triage Model (RSTM).

2.3 Study III

Study III was designed to conduct a prospective analysis of the RSTM developed in Study II in a single PC clinic. It focused on assessing its safety and efficacy in clinical environments and comparing its predictions with patient outcomes. It allowed us to examine a much larger group of outcomes than we could in study II and evaluate how the RSTM fared in real clinical settings.

3 Materials and methods

The first two studies were retrospective, while the last was a prospective single-center study conducted in clinical settings at one of the primary care clinics of the Primary Health Care of the Capital Area (PHCCA). All three studies were approved by the Icelandic Data Protection Authority, the Icelandic National Bioethics Committee, and the ethics committee of the PHCCA. The reference numbers for the studies are VSN-19-153, VSN-20-198, and VSN-21-204.

3.1 Study populations

In the first two studies, International Classification of Diseases (ICD) codes were used to select the original patient cohort. For the third study, patients presenting with respiratory symptoms were recruited, and filtering methods were applied to ensure the final population was suitable for the ML model. In all three studies, only adult patients were included. In the first two studies the populations were selected from a database from 19 primary care clinics in the capital area of Iceland. On 1st of January 2024 there were 239,733 people living in this area[80]. In the third study, participants were recruited from a single PC clinic in the capital area of Iceland. In June of 2024 there were 15,393 patients registered at the clinic. The following subsections describe the study populations in more detail.

3.1.1 Study I

Study I was a retrospective analysis encompassing patients, older than 18, diagnosed with one of four specific primary headache disorders, defined by their ICD codes, from January 1, 2006, to April 30, 2020. The selected ICD codes were G44.2 (Tension-type headache; TTH), G43.9 (Migraine without aura; MA-), G43.1 (Migraine with aura; MA-) and G44.0 (Cluster headaches; CH).

3.1.2 Study II

The second study, retrospective as well, investigated the possibility of training an ML model to predict the likelihood of severe lower respiratory tract infections based on symptomatic history only. This study involved patients older than 18, diagnosed between January 1, 2016, and December 31, 2018. The selected diagnostic codes for inclusion were J00 (common cold), J10 and J11 (influenza), J15 (pneumonia), J20 (acute bronchitis), J44 (Chronic Obstructive Pulmonary Disease, COPD), and J45

(asthma). J15 was the only code signifying pneumonia in the second study, but the diagnostic codes for pneumonia ranged from J12 to J18. We sampled ICD codes before we got the final dataset for the second study, and in that sample, the overwhelming proportion of pneumonia cases were coded using J15.

3.1.3 Study III

The third study was a prospective evaluation of the safety and performance of the ML model developed in the second study. To ensure comprehensive coverage of patients exhibiting respiratory symptoms, healthcare staff were asked to report every patient, 18 years or older, who reported such symptoms to a study recruiter at the Centrum primary care clinic. The Centrum clinic ("Heilsugæslan Miðbæ") is one of the 15 clinics under the PHCCA. This decision was made due to the impossibility of determining the diagnoses patients will receive before they consult with a physician. Travelers were excluded from the study to avoid biasing the results, as their transient nature precluded the necessary follow-up required for the study. Since elderly individuals are less likely to possess a mobile phone or an electronic ID, the option to sign the research consent with a pen was also provided. Subsequently, participants completed the survey on a desktop computer at the primary care clinic. Table 1 provides an overview of the three studies presented in this thesis, their design, populations, number of participants, time period, and reported outcomes.

Table 1. Summary of material and methods in the three studies. ICD, International classification of diseases; SHAP, Shapley additive explanations; CRP, C-reactive protein; PC, primary care; ED, emergency department; CXR, chest x-ray; CT, computed tomography.

Study number	Study I	Study II	Study III
Study cohort	Patients diagnosed with G44.2, G43.0, G43.1, and G44.0	Patients diagnosed with J00, J10, J11, J15, J20, J44, and J45 including subgroups	All patients with respiratory symptoms as presenting complaints
Study type	Retrospective	Retrospective	Prospective
Number of participants	800	2,567	512
Period	2006 - 2021	2016 - 2018	2023 - 2024
Outcome measures	ICD code, SHAP values	CRP value, antibiotic treatment, re-evaluation in PC within seven days, re-evaluation in ED within seven days, ICD code, CXR referral, signs of pneumonia on CXR, incidentalomas	CRP value, antibiotic treatment, re-evaluation in PC within seven days, re-evaluation in ED within seven days, ED referral, ICD code, CXR referral, signs of pneumonia on CXR, CT referrals, lung cancer, incidentalomas

3.2 Data collection

The data for the first two studies were obtained from the database of the common EHR system of the PHCCA and the four private primary care clinics in the capital area of Iceland. The first study's original dataset comprised 15,024 medical records from 13,682 patients. Each record included a CTN and an ICD diagnostic code. In the second study, the original dataset contained 44,007 medical records from 23,918 patients. Each record consisted of a CTN, with diagnostic referrals, results, diagnoses, and prescriptions. In the third study, a custom web-based program was developed specifically for data collection. This platform enabled participants to respond to the questions necessary for the model to generate predictions. The process unfolded in the following manner: Upon agreeing to participate in the study, individuals were directed to a designated URL in a web browser hosting the program. Participants had the option to set the language to either Icelandic or English. When participants started the program, they were presented with a research consent form, which included an option for electronic signature at the bottom. The study recruiter then reviewed the research consent with each participant and addressed any questions during the explanation. Once consent was given, an interface appeared, enabling participants to input their main complaints by typing into an input field with dropdown suggestions or selecting from a list of common complaints. Next, participants were prompted to rank their primary complaints in order of severity. Following this, participants responded to a series of questions required by the model to generate its output. Upon completing the required questions, participants received a thank-you notification, marking the end of their input.

3.2.1 The database of all primary care clinics in the capital area

The data for the first two studies were gathered from the shared EHR database of primary care clinics in the capital region of Iceland. Only the CTN and a diagnostic code were obtained in the first study. For the second study, all clinical encounters involving the selected diagnoses were retrieved, including imaging referrals, results, prescriptions, and all associated ICD codes, to ensure a comprehensive outcome review.

3.3 Population selection

Different criteria and methods were used to select the final population in all three studies. This section describes the process in each study.

3.3.1 Study I

The population selection in the first study is described in section 3.1.1. The data was subsequently preprocessed before model training. The data processing pipeline is described in detail in chapter 4.1.2.

3.3.2 Study II

A similar method was used in the second study as the first one, with slight adjustments. CTNs with less than 250 characters were removed, and then 7,000 CTNs were chosen randomly for analysis. These 7,000 CTNs were input into an LLM that had been previously trained, as detailed in earlier work and publications[81]. The LLM outputs annotations for each CTN, which we used to pre-select for CTNs rich in CFs. The LLM also outputs the main complaints for each patient (if present), and we discarded CTNs with less than eight CFs in total and then selected CTNs based on main complaints (which were also annotated in each CTN) to include only patients presenting with acute/subacute respiratory symptoms.

When training the RSTM, we analyzed the performance of multiple ML models on the evaluation dataset. The multivariate LR with the Least Absolute Shrinkage and Selection Operator penalty (LASSO) showed the best performance of these models. Subsequently, the 50 most impactful input features were selected based on SHAP values during training, and the model was retrained using only these features to train the RSTM.

3.3.3 Study III

In the third study, we used primary complaints to exclude participants who would likely be diagnosed with respiratory infections on which the RSTM was originally not trained. As the RSTM was trained initially on patients with respiratory infections, excluding ear, throat, and sinus infections, we excluded patients with pharyngitis, ear pain, or pain/pressure in maxillary or frontal sinuses as one of their primary complaints. The reason for this approach is discussed in detail in the chapter 5.

3.4 Data processing

In all studies, the data was preprocessed before statistical analysis was conducted. Different types of ML models were trained in the first two studies. This section describes data processing and ML model training.

3.4.1 Data processing and model training

The dataset underwent preprocessing before training the ML models in the first two studies. The following subsections provide a detailed description of the preprocessing steps.

3.4.1.1 Study I

A medical keyword dictionary was created to address the issue of empty or minimally informative CTNs. This dictionary includes a range of symptoms that physicians commonly inquire about or use to describe patient conditions during diagnosis, such as nausea, vomiting, phonophobia, etc. Its purpose is to help identify and eliminate CTNs lacking diagnostically informative content. The dictionary was created by manually reviewing a subsample of the dataset, approximately 1% from each category, and identifying the relevant keywords for all diagnoses. Subsequently, this dictionary was applied to the entire dataset, filtering out CTNs that did not contain the selected keywords, mirroring a methodological approach used in a similar research study[79].

After filtering the CTNs with the keyword search and removing 93 duplicate CTNs, a dataset of 2,563 information-dense CTNs remained. This accounts for 17% of the original dataset. Subsequently, 800 CTNs were selected randomly to be annotated by a single physician. The data annotation process is described in detail in section 3.4.2. The annotation resulted in a dataset of 8,595 question-answer pairs; a single pair is called a CF. Thus, a single CF contains a question and a value that can be binary or an integer. The dataset included 254 different types of CFs. This dataset was randomly split into training, validation, and test sets containing 600, 100, and 100 CTNs, respectively. Multiple ML classifiers were trained and evaluated on the evaluation dataset where RF performed best. The other algorithms trained and evaluated were SVM, multivariate LR, decision trees, k-Nearest Neighbors, Naïve Bayes and XGBoost. RF was chosen since it scored highest according to the Matthews Correlation Coefficient (MCC) metric. After the training, the test set was input into the model, and the output was used for statistical analysis.

3.4.1.2 Study II

The original dataset in the second study contained 44,007 CTNs. All CTNs with less than 250 characters were removed, resulting in 17,177 CTNs. Before the second study, a DNN had been trained, called the Clinical Feature Extraction Model (CFEM)[81]. The CFEM is an LLM that takes CTNs as an input and gives CFs as output. Seven thousand randomly selected CTNs were input into the CFEM, and using the CFs from the CFEM, all CTNs with less than 8 CFs were removed. This

process increases the odds of the resulting ML model having enough CFs in each CTN to learn from. The CFEM also extracted presenting complaints that were used to increase the number of patients presenting with a first occurrence of acute or subacute respiratory infection in the final dataset. Using presenting complaints, 95 CTNs from follow-up consultations and 223 CTNs with multiple topics were removed. This means that for COPD and asthma patients, only cases of exacerbation were included, and all CTNs from regular follow-up were removed. After applying these filters, 2,942 CTNs remained, whereas 2,000 CTNs were randomly selected for annotation by a physician. The training dataset consisted of 1,500 CTNs, and 500 CTNs were held out for test set 1. Additionally, we annotated 652 CTNs from patients diagnosed with influenza (J10 or J11) to serve as a second test set.

3.4.1.3 Study III

Using a model with 50 input features would be impractical in clinical practice, as it would require collecting a large amount of patient data, making it cumbersome in real-world settings. Therefore, before prospectively collecting data in Study III, we refined the model by analyzing its co-variates and selecting only co-variates with p-values less than 0.05. When reviewing the co-variates removed during this process, the question regarding blood in sputum/hemoptysis was not removed from the input features of the RSTM despite the co-variate for this feature not having a p-value less than 0.05. During the evaluation of the RSTM, we deemed it better to err on the side of safety. Hemoptysis, being a common symptom of lung cancer, requires prompt medical evaluation in clinical practice. Therefore, patients who reported hemoptysis were assigned an additional 0.5 to their risk score, leading to a high-risk classification. We then retrained the model on the whole training dataset in study II using only these selected co-variates before conducting study III with originally used hyperparameters. The retrospective performance of the RSTM remained similar or slightly better after this adaptation.

3.4.2 Data annotation

In the first two studies, the data was annotated using a specific method before the ML models were trained on the annotated data. The annotation method was inspired by researchers who applied similar annotations on medical text[79], which assigned binary and numerical values to CFs, representing the presence or absence of signs and symptoms in the CTNs as described in the text. The text in the CTNs constituted the patient's health state as described by a physician during the medical consultation when the CTN was written. Each annotation is formed as a question-answer pair where the question is stated to ask about a specific clinical variable,

and the answer is annotated depending on the variable type. For example, consider a frequent text description in CTNs: 'The patient has a cough and a sore throat but no fever.'. Two question-answer pairs regarding the patient's cough and sore throat are annotated as positive, and the third one, regarding the fever, is annotated as negative. Numerical values variables were assigned a number value, such as a variable for temperature (e.g., with the value of 38.2°C). Missing binary variables were given the value of 0. Missing value variables were replaced by randomly sampling the normal distribution for that variable to reduce the odds of the model simply learning where value variables are missing, which would be more likely for a patient with a less severe disease. The annotation in both studies was done blindly, where the annotator only saw a single CTN when annotating, nothing else. In study I, all textual descriptions of symptoms, past medical history, and examination were annotated, but only symptoms and past medical history were annotated in study II.

3.4.3 Data gathering process in study III

Patients were asked to respond to the questions below according to their present condition unless instructed otherwise (disregarding the progression of their illness). For each of the following questions, they could answer with "yes," "no," or "I don't know":

- Do you have symptoms of the common cold?
- Have you been diagnosed with asthma by a physician?
- Is there mucus / expectorate coming up from your lungs?
- Have you been coughing up blood, or has there been blood in the mucus from your lungs?
- Have you experienced a fever in the course of your symptoms?
- Have you experienced shortness of breath?
- Have you been diagnosed with COPD by a physician?
- Do you have any allergies?
- Have you vomited?
- Do you have a fever?

If patients answered "Yes" to the question "Do you have a fever?" they were prompted to enter their temperature in degrees Celsius. The input allowed for floating-point numbers from 35 to 42. If they hadn't taken their temperature, they

could select "I have not measured my temperature." A default value of 37°C was used for the model input in these instances. The following items were presented as a list of checkboxes associated with the question, "Do you have any of the following symptoms?":

- Cough
- Sore throat
- Runny Nose (Rhinorrhea)
- Nasal congestion / Stuffy nose

Additionally, the patients answered the question "Do you have a smoking history?" by selecting one of the four choices:

- active smoker
- former smoker
- social smoker
- no

Age and sex are also part of the predictors of the RSTM and were collected automatically when each participant signed the consent electronically.

3.5 Statistical analysis

Data processing was done through custom scripts within the Jupyter Notebook environment, utilizing the Python programming language[82]. All statistical analyses and model training procedures were conducted using the scikit-learn library, version 0.22.1. Descriptive statistics were reported as numbers (%) and mean $\pm$ with a 95% confidence interval where applicable. The 95% confidence intervals were calculated by sorting the values for each outcome and calculating the 2.5% and 97.5% percentiles.

In the first study, the following metrics were used for the performance assessment: sensitivity, specificity, Positive Predictive Value (PPV), MCC, Receiver Operating Characteristics (ROC) curve, and Area under the ROC curve (AUROC score). A confusion matrix is a statistical presentation of the performance of a classifier and consists of four quadrants. Each quadrant contains a value count for one of four performance variables: true-positive count (TP), false-positive count (FP), true-negative count (TN), and false-negative count (FN). The MCC is a valuable metric for imbalanced datasets since it only produces a high score if the prediction obtains

good results in all four quadrants of the confusion matrix. It ranges from 1 to 1, where 1 represents a perfect prediction. The MCC is defined as follows:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(FN + FP)(TN + FN)}}$$

The MCC was calculated for each diagnosis as a dichotomous outcome. The AUROC score was determined by plotting the true positive rate (TPR) against the false positive rate (FPR) at various thresholds for each dichotomous diagnosis and calculating the area under the curve. The ROC curve for each diagnosis, as a dichotomous outcome, was also plotted. The calculated means were weighted to account for a substantial imbalance in diagnostic codes. This adjustment provided a more accurate performance comparison between the model and the physicians. The inner workings of the ML model were examined using SHAP values.

The second study evaluated the model's performance using patient outcomes from two separate test sets. The first test set comprised patients with the same diagnostic distribution as the training dataset. In contrast, the second test set included patients diagnosed with influenza (J10 and J11 including subgroups), a group theoretically outside the training set but including patients that often present with symptoms similar to those in the training dataset. Based on the model's risk score, patients in each test set were divided into ten risk groups, numbered from 1 to 10.

3.5.1 Shapley additive explanations (SHAP) values

SHAP values are a tool used in ML to explain the output of predictive models. They are based on Shapley values from cooperative game theory, which allocate a fair distribution of payoffs to players depending on their contribution to the total payout. In the context of ML, SHAP values quantify the contribution of each feature in a model to a specific prediction. This approach provides insights into how different features impact the model's output, making it a potent method for interpreting complex ML models. In study I, SHAP values were used to examine the inner workings of the ML classifier, where the top-20 most impactful input features for each diagnosis were plotted in a SHAP value summary plot. SHAP values were also used to choose the top 50 input features in study II.

There are multiple ways of plotting SHAP values. We created a summary plot that combines all features of SHAP values across a dataset in one chart. Each dot represents a SHAP value for a single sample, plotted along the x-axis according to the magnitude of its effect on the prediction. The x-axis shows the impact on the

model output. A SHAP value of 0 means no impact, while positive and negative values indicate positive or negative influences on the prediction. The color of each feature represents its value from low (blue) to high (red). For instance, if age is high for red dots and low for blue, you can see how higher or lower age values contribute to model predictions. The features are also ranked from top to bottom by their importance across the dataset. To interpret the model in the first study, we used SHAP values to analyze the top 20 features influencing the model's predictions.

3.5.2 Receiver-operating characteristic curve

In study I, we used a ROC curve to assess diagnostic accuracy for each condition, comparing the performance of individual physicians and the group mean to the performance of the ML model. We created a single ROC curve for each diagnosis as a dichotomous outcome. The ROC curve shows the trade-off between sensitivity and specificity at various thresholds, with the area under the curve (AUC) summarizing overall performance. An AUC of 1 indicates perfect accuracy, while 0.5 represents random guessing. By plotting each physician's results, we identified variations in diagnostic ability, providing insights into individual and group performance.

3.5.3 Risk group stratification and outcomes

In studies II and III, we used the RSTM risk score to stratify patients into one to ten risk groups. We then analyzed the selected outcomes in each group to evaluate the safety of this risk stratification. Furthermore, we used the terms Low-Risk (LR) and High-Risk (HR) groups for all patients in risk groups 1-5 and 6-10, respectively, to simplify the discussion and comparison of results.

3.5.4 Cross-Validation

Cross-Validation (CV) is a method used in ML where a dataset is divided into a specified number of subsets or folds. One fold is held out as a validation set for testing the model, and the other folds are used as training data. This process helps assess the model's generalization on unseen data. There are three main types of CV: K-fold CV, Leave-One-Out CV, and stratified K-fold. Then, there is a standard CV and a nested CV. Hyperparameter optimization and model evaluation are performed in the same loop in a standard CV grid search. This approach is more straightforward but more prone to overfitting than nested CV. A nested CV addresses this problem by creating an inner and outer loop. The inner loop is used for hyperparameter tuning with its own CV, and the outer loop is used to evaluate the model with the tuned parameters from the inner loop. A nested CV ensures that

model evaluation is unbiased by separating the data used for hyperparameter search and assessment. In study 2, when training the RSTM, we used a 25-repeated stratified nested CV and grid search to optimize the hyperparameters, optimizing only class weight and the penalty (C parameter). This resulted in the balanced hyperparameter being used with a C value of 0,1.

3.5.5 Outcomes

In the first study, the only outcome measured was ICD code diagnosis. In the second study, we reported on the mean C-Reactive Protein (CRP) value, ICD-10 code distribution, the proportion of patients re-evaluated in primary care and emergency departments within seven days, the proportion of patients referred to a Chest X-Ray (CXR), CXRs with signs of pneumonia and incidentalomas, and the proportion of patients receiving antibiotic prescriptions. In study III we analyzed mean CRP value, ICD code distribution, the proportion of patients re-evaluated in PC within seven days, the proportion of patients seeking evaluation in the Emergency Department (ED) within seven days, the proportion of patients referred to ED immediately by the receiving doctor, the proportion of patients referred to a CXR, CXRs with signs of pneumonia, tumors, and incidentalomas, the proportion of patients receiving antibiotic prescriptions, the proportion of patients referred for CT of the thorax, and the proportion of patients referred for CT thorax with pneumonia, cancer, and incidentalomas.

4 Results

The main findings of this research indicate that ML models trained to predict ICD codes based on clinical features extracted from CTNs perform similarly or slightly better than GPs and rely on the same clinical features used by humans. Additionally, such models show positive possibilities in triaging patients with respiratory symptoms based on their symptomatic history, as demonstrated in Studies II and III. The RSTM performed well when stratifying patients by severity of symptoms and by means of using the probability of having a Lower Respiratory Tract Infection (LRTI) as a marker. This was further supported by prospective analysis, where the RSTM accurately placed two patients in lower-risk categories despite physicians diagnosing them with pneumonia when subsequent CXRs confirmed the absence of pneumonia in these cases.

4.1 Study I

The main result of study I was that the ML classifier achieved a similar or slightly better diagnostic performance in diagnosing primary headaches in primary care than the mean performance of GPs and GP trainees. Figure 2 shows the flowchart of CTNs in the first study.

4.1.1 Descriptive statistics

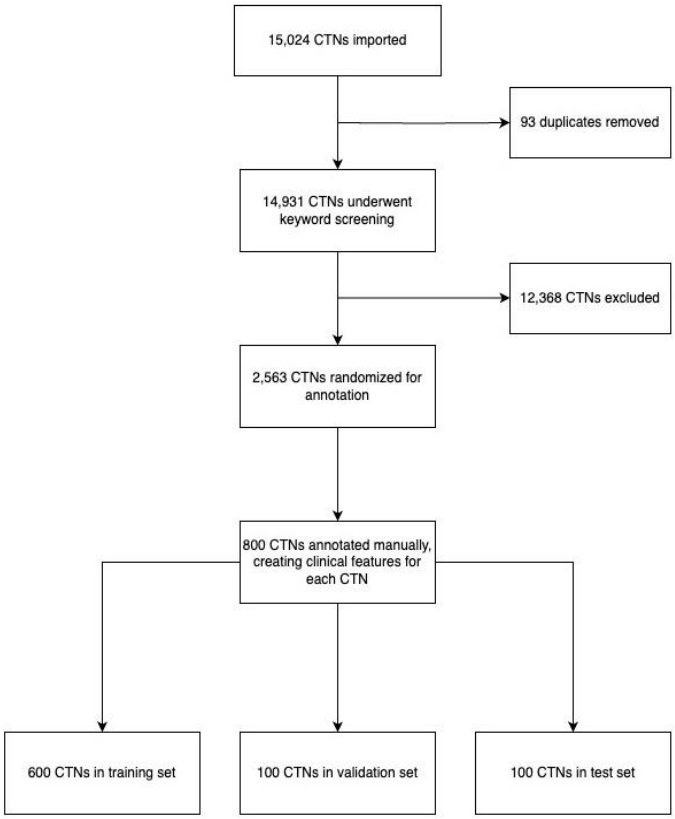


Figure 2. The flowchart in the first study.

A total of 15,024 CTNs were imported from the database. Then, 93 duplicate CTNs were removed, and 12,638 CTNs were excluded after applying the keyword dictionary, leaving 2,653 CTNs. These CTNs were randomized, and then 800 CTNs were annotated by a physician. A total of 8,595 CFs were annotated, where 4,913 CFs were positive and 3,682 CFs were negative, resulting in 1,33 positive CFs for each negative. The mean number of CFs per CTN was 10,74. Table 2 shows the demographics of the train, validation, and test sets in study I. Females are more likely to experience primary headaches and seek help in primary care, which explains the imbalance in the dataset. The demographics were similar in all datasets.

Table 2. The demographical distribution in study I in the training, validation, and test set.

Variable	Training set	Validation set	Test set
Total size	600	100	100
Female (%)	434 (72.3)	65 (65.0)	72 (72.0)
Mean age (min-max)	34.56 (6-90)	33.53 (8-78)	31.29 (8-77)

4.1.2 Model and Physician Performance

Table 3 shows the ML classifiers and physicians' performance with weighted averages. The ML classifier outperforms or is equal to the physicians on the weighted average on all metrics except specificity. The ML classifier performs best on CH and MA+ but has a more challenging time discerning between TTH and MA- in some cases, achieving the lowest scores for MA-.

Table 3. Performance metrics for ML model and GP specialists and trainees. CH, cluster headache; MA+, *migraine with aura;* MA-, migraine without aura; TTH, tension-type headache; PPV, positive predictive value; MCC, Matthew's correlation coefficient.

Diagnosis	Sensitivity	Specificity	PPV	MCC
		ML classifier		
CH	0.83	1.00	1.00	0.91
MA+	0.92	0.99	0.92	0.9
MA-	0.67	0.99	0.86	0.73
TTH	0.99	0.85	0.95	0.87
Weighted average	0.95	0.88	0.94	0.86
		GP specialist 1		
CH	0.83	1.00	1.00	0.91

MA+	0.92	0.95	0.73	0.79
MA-	0.8	0.92	0.53	0.61
TTH	0.89	0.96	0.98	0.79
Weighted average	0.88	0.95	0.88	0.77
		GP specialist 2		
CH	0.67	1.00	1.00	0.81
MA+	0.58	0.96	0.70	0.59
MA-	0.90	0.87	0.45	0.58
TTH	0.90	0.95	0.98	0.79
Weighted average	0.86	0.94	0.85	0.73
		GP specialist 3		
CH	0.83	1.00	1.00	0.91
MA+	0.67	0.99	0.89	0.74
MA-	0.5	0.95	0.56	0.48
TTH	0.97	0.72	0.91	0.75
Weighted average	0.90	0.78	0.88	0.73
		GP trainee 1		
CH	0.83	1.00	1.00	0.91
MA+	0.92	0.99	0.92	0.90

MA-	0.70	0.93	0.54	0.56
TTH	0.93	0.88	0.96	0.80
Weighted average	0.89	0.91	0.90	0.78
		GP trainee 2		
CH	0.83	0.95	0.56	0.65
MA+	0.42	0.99	0.83	0.55
MA-	0.80	0.79	0.31	0.41
TTH	0.81	0.95	0.98	0.64
Weighted average	0.78	0.91	0.76	0.58
		GP trainee 3		
CH	0.83	0.98	0.71	0.75
MA+	0.50	0.99	0.86	0.62
MA-	0.70	0.90	0.44	0.49
TTH	0.94	0.90	0.97	0.82
Weighted average	0.87	0.91	0.86	0.75

4.1.3 ROC curves

A ROC curve was plotted for each diagnosis as a dichotomous outcome for the ML classifier outputs, and the performance of each physician and the mean of both groups were also plotted. Figure 3 displays the ROC curves. When comparing classifiers, the one further up to the left on the ROC curve achieves a higher ratio of TPR and FPR. The ROC curves show that the classifier achieves the highest performance score when diagnosing MA+ and CH but struggles more with TTH and MA- diagnosis. No mean physician scores outperformed the ML classifier on any diagnosis on the ROC curve. Still, two GP specialists outperformed in MA- diagnosis, and one GP trainee outperformed the classifier on TTH.

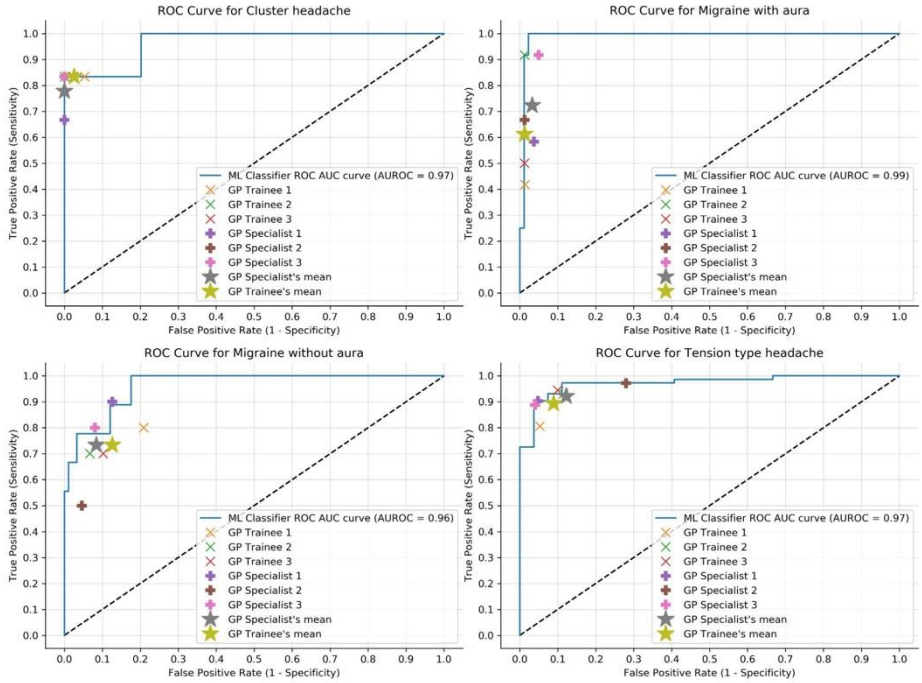


Figure 3. The ROC curve for the classifier for each diagnosis can be seen in blue, along with the performance of each physician and the mean score of both physician groups. The + and X icons indicate the score for an individual GP trainee and specialist, the green star shows the mean score for the GP trainees, and the gray star shows the score of the GP specialists. The area under the receiver operating characteristic (AUROC) score is calculated as the area under the ROC curve, and a higher AUROC score indicates a better classifier performance. The performance of a classifier is better as the score approaches the upper left corner of the graph.

4.1.4 SHAP values

Figures 4-7 show the SHAP values for each diagnosis's top 20 input features. In Figure 4, the SHAP values for CH are displayed. The features include CHs autonomic

symptoms such as ptosis, conjunctivitis, lacrimation, runny or stuffy nose accompanying a unilateral headache around a single eye. Myalgia symptoms reduce CHs diagnostic probability as does visual disturbance which often accompanies MA+. In Figure 5, SHAP values for MA+ are visualized. The two most impactful features of MA+ are prodromal symptoms and visual disturbance, the former is a hallmark of an aura and the latter for migraines. Unilateral symptoms and photophobia positively impact MA+ diagnostic probability but less since MA- also typically has these features. Aura can present itself as limb numbness, which also positively affects the prediction of MA+. In the case of MA-, we see a different distribution of SHAP values, displayed in Figure 6. The model uses prodromal symptoms and visual disturbance to distinguish between MA+ and MA-. The SHAP values describe a unilateral headache with accompanying nausea, vomiting, photophobia, phonophobia and without symptoms of myalgia. This picture fits well with MA-. Figure 7 displays the SHAP values for TTH. This distribution differs from the other three, where it shows an absence of symptoms, and only myalgia symptoms affect the prediction probability positively.

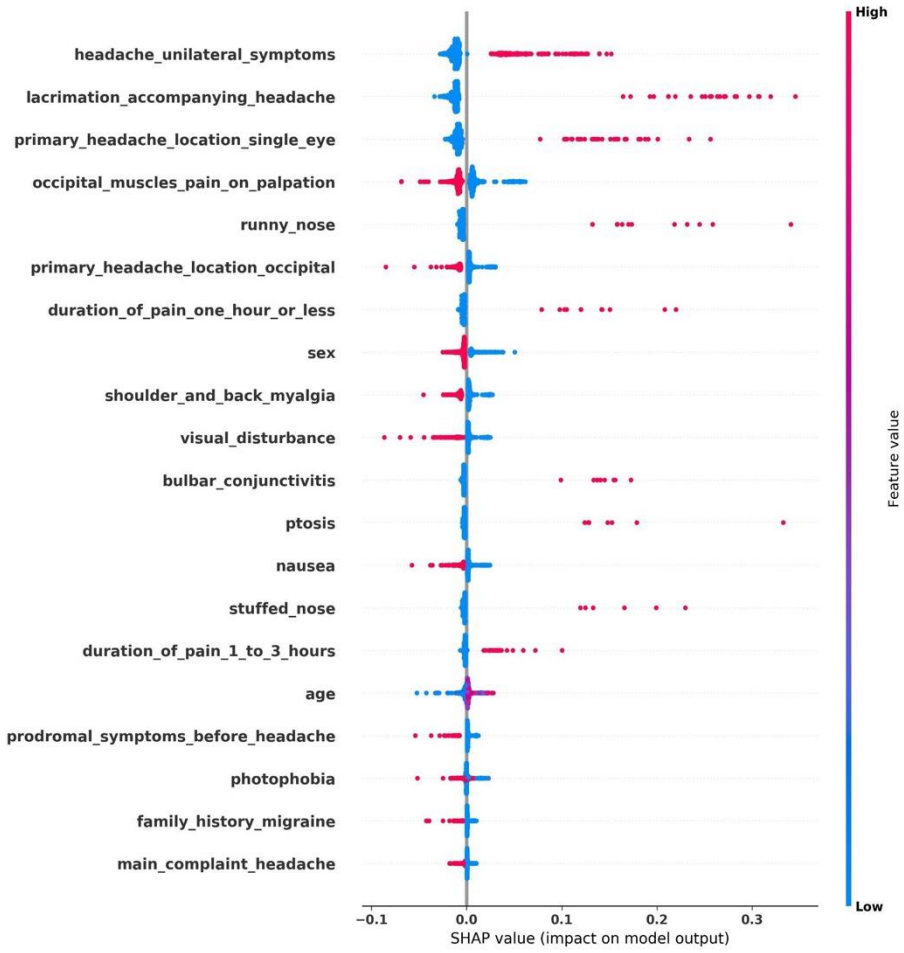


Figure 4. SHAP values for the most impactful features for a CH prediction. Each dot represents a SHAP value for a single sample, plotted along the x-axis according to the magnitude of its effect on the prediction. The x-axis shows the impact on the model output. The features are also ranked from top to bottom by their importance across the dataset.

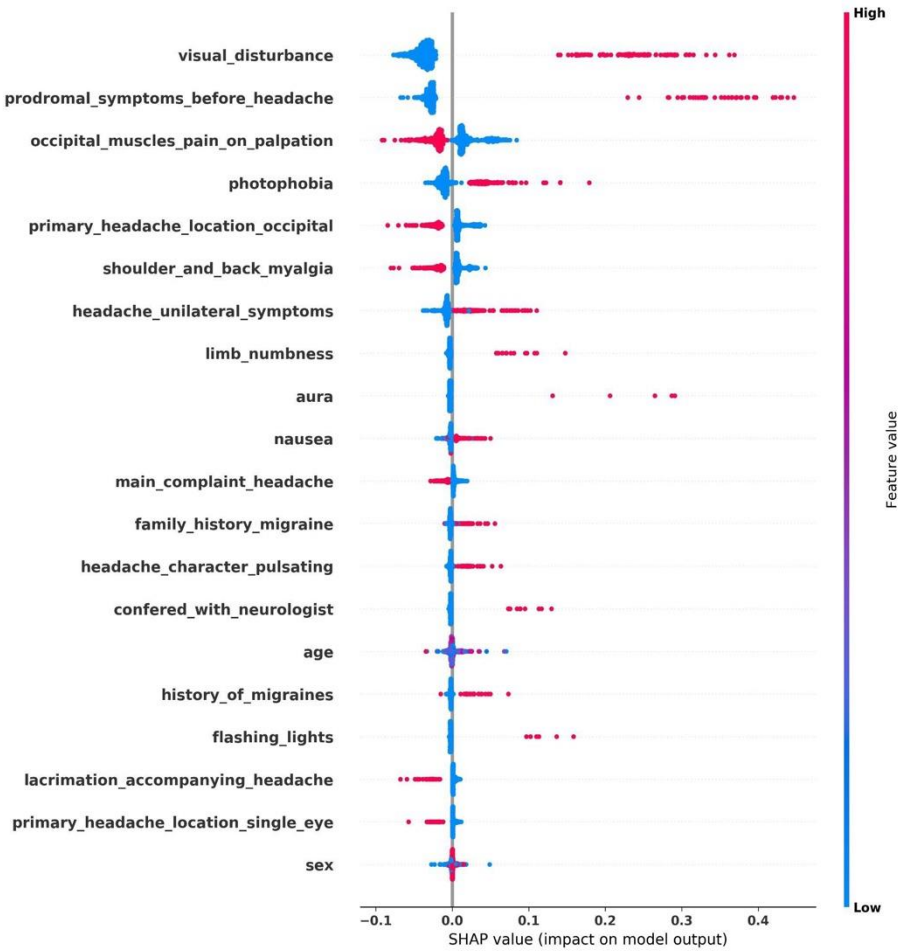


Figure 5. SHAP values for the most impactful features for a MA+ prediction. Each dot represents a SHAP value for a single sample, plotted along the x-axis according to the magnitude of its effect on the prediction. The x-axis shows the impact on the model output. The features are also ranked from top to bottom by their importance across the dataset.

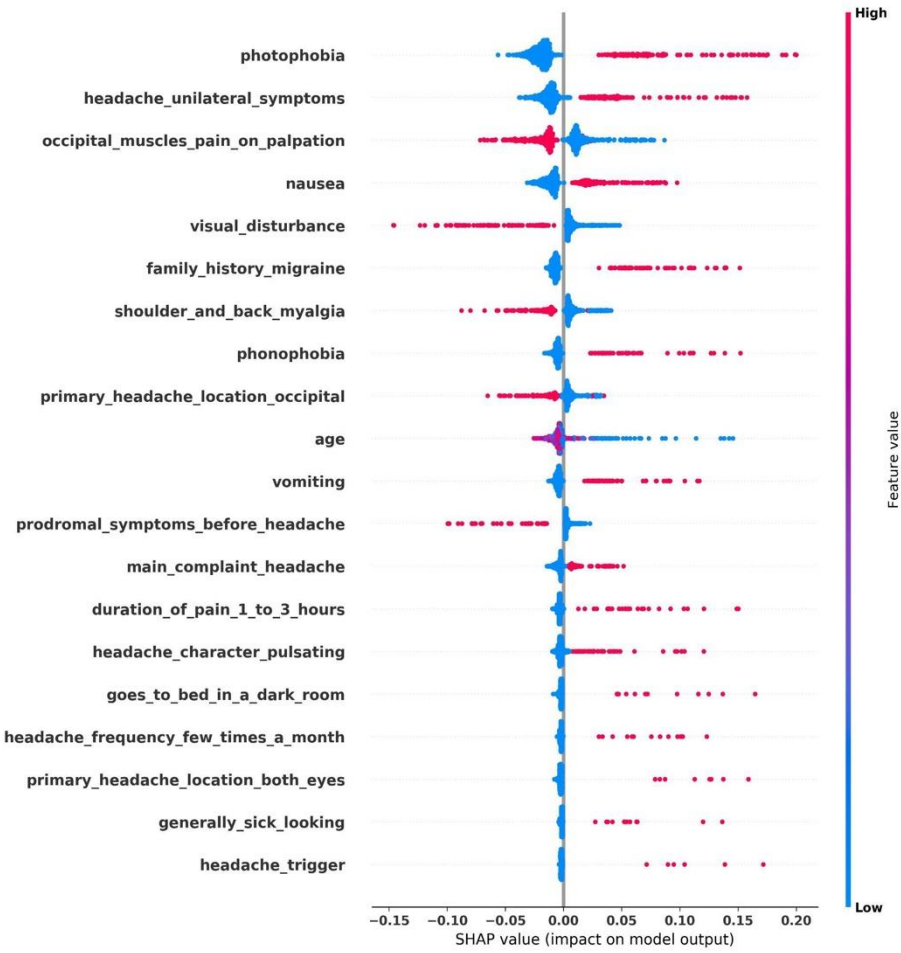


Figure 6. SHAP values for the most impactful features for a MA- prediction. Each dot represents a SHAP value for a single sample, plotted along the x-axis according to the magnitude of its effect on the prediction. The x-axis shows the impact on the model output. The features are also ranked from top to bottom by their importance across the dataset.

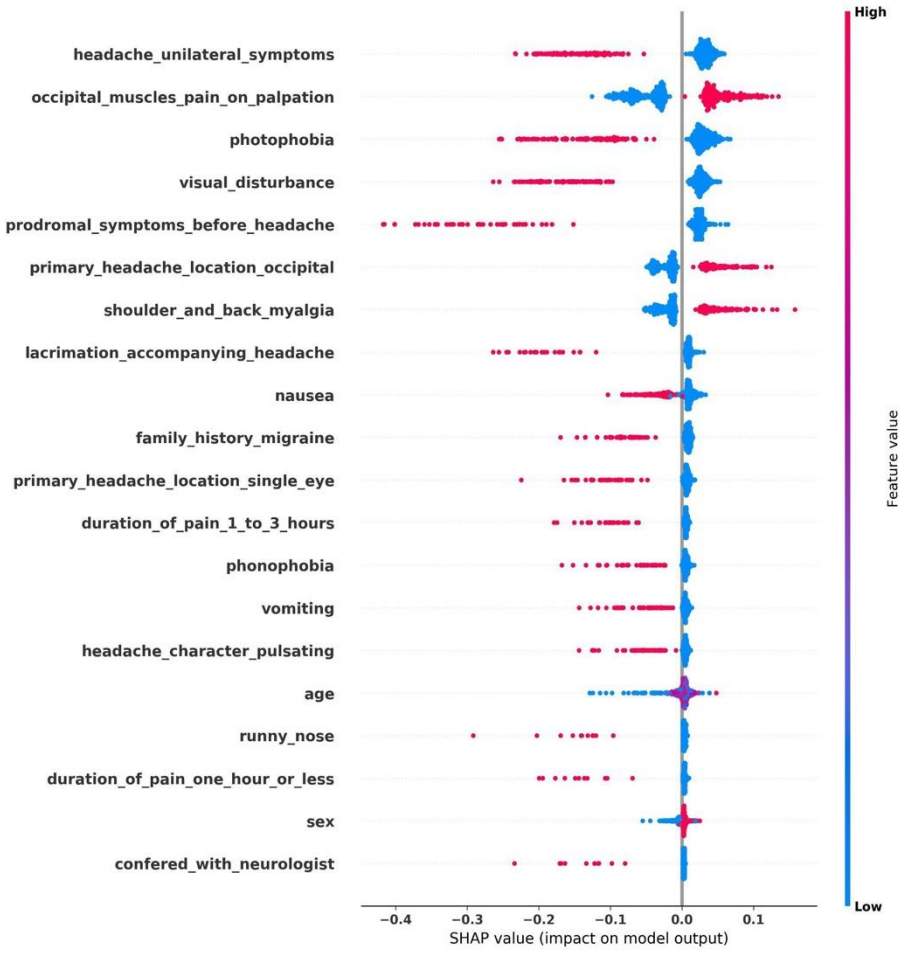


Figure 7. SHAP values for the most impactful features for a TTH prediction. Each dot represents a SHAP value for a single sample, plotted along the x-axis according to the magnitude of its effect on the prediction. The x-axis shows the impact on the model output. The features are also ranked from top to bottom by their importance across the dataset

4.2 Study II

The main result of study II was that the RSTM triaged patients with the selected ICD codes so that no severe outcomes, such as pneumonia diagnoses or CXRs positive for pneumonia, were observed in the lower-risk groups. All outcomes increased proportionally with higher risk groups.

4.2.1 Descriptive statistics

We received 44,007 medical records from 23,819 patients from the database. To remove CTNs with little or no information, we removed all CTNs containing less than 250 characters, which removed 26,830 CTNs from the dataset. We then selected 7,000 CTNs randomly to be used as input into the CFEM, which we had created in previous work and has been described and cited in section 3.4.1.2. We then removed all CTNs containing less than 8 CFs per CTN, which resulted in the removal of 4,058 CTNs. The CFEM also marked CTNs with multiple topics and where patients were coming for follow-up. Due to this, 223 and 95 CTNs were removed, leaving 2,942 CTNs. Then, 2,000 CTNs were selected randomly for annotation by a physician. After annotation, these 2,000 CTNs were split randomly into training and test sets. Figure 8 displays the flow of the CTNs in study II.

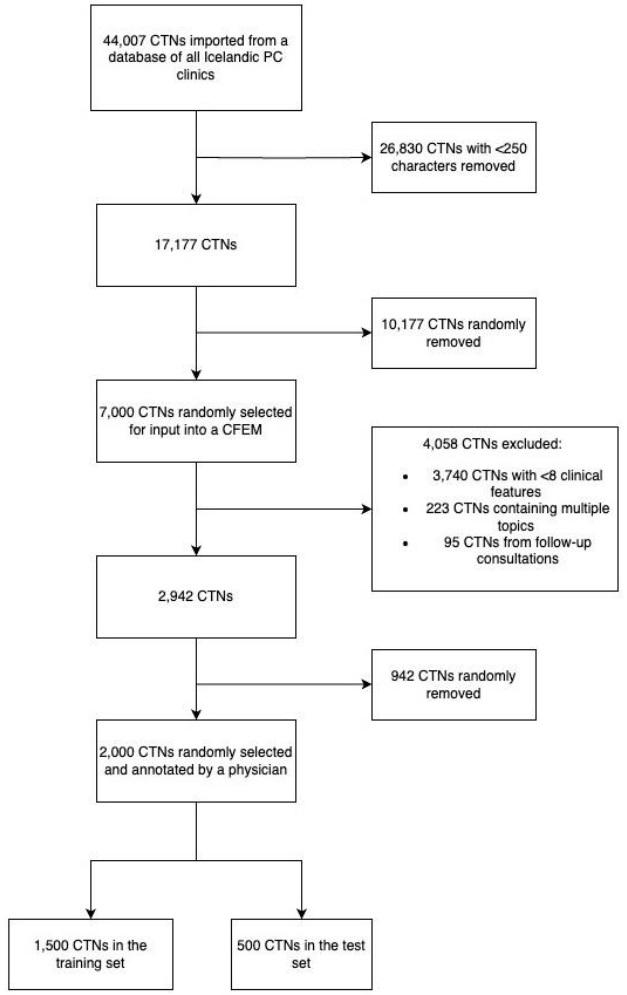


Figure 8. The flowchart in the second study.

Table 4 shows the demographics of the training set and both test sets in the second study. The mean population age in test set 2 is slightly lower than in the training and test set 1, likely because of different distributions of ICD codes.

Table 4. The demographic and ICD code distribution in the training and both test sets in study II.

Variable	Training set	Test set 1	Test set 2
Total size	1500	500	664
Female (%)	66.1	62.6	61.1
Mean age (min-max)	54 (18-93)	55 (19-91)	45 (18-92)
ICD-10 codes (%)			
J00 (Acute nasopharyngitis)	20.2	19.2	0
J15 (Bacterial pneumonia)	0.4	0.2	0
J20 (Acute bronchitis)	46.5	48.2	0
J44 (Chronic obstructive pulmonary disease)	11.1	9.8	0
J45 (Asthma)	21.8	22.6	0
J10 (Influenza)	0	0	0.8
J11 (Influenza)	0	0	99.2

The outcome distribution in the training and test datasets can be seen in Table 5. The distribution is similar between the training and test set 1, which is unsurprising since they come from the same groups of diagnoses. Patients in test set 2 were diagnosed with influenza (J10 and J11) and, therefore, have a different outcome distribution. Antibiotic prescribing is much lower, which one would expect given the viral nature of influenza. Since influenza patients often have significantly raised CRP values, the mean CRP being higher in test set 2 compared to the other two datasets is not surprising. There are also slightly higher re-evaluation rates in PC and ED within seven days. This fits well with influenza patients, who often experience more severe symptoms than patients in the other two datasets.

Table 5. The outcome distribution in the training dataset, test sets 1 and 2.

Patient outcomes	Training set	Test set 1	Test set 2
CRP value, mean (range), mg/dL	20.7 (3-183)	18.8 (3-84)	28.2 (3-178)
Antibiotics prescribed (%)	49.0	50.8	15.5
CXR referrals (%)	6.6	6.8	3.5
CXRs signs of pneumonia (% of referrals)	12.6	9.6	9.1
CXR incidentalomas (% of referrals)	1.6	3.2	4.3
PC re-evaluation within 7 days (%)	8.2	6.8	10.0
ED re-evaluation within 7 days (%)	0.2	0.6	0.8

4.2.2 Outcome comparison between the LR and HR groups

Table 6 compares the outcome rates in the HR and LR groups across both test sets. In test set 1, a significant difference was observed between the HR and LR groups in CRP values and antibiotic treatments. However, in test set 2, only antibiotic treatment showed a substantial difference between the groups. Since test set 2 includes only patients diagnosed with influenza, the similarity in CRP values is expected, as influenza patients typically exhibit higher CRP values than many of the diagnoses in test set 1.

Table 6. The comparison of outcome rates in the test sets between the LR and HR groups. CRP, C-reactive protein; PC, primary care; ED, emergency department; CXR, chest X-ray.

Outcomes	Low-Risk group	High-Risk group	P value
Test set 1			
CRP value, mean, mg/dL	7.9	23	<.05
Antibiotics prescribed	66 (40%)	188 (56.1%)	<.05
PC re-evaluation	9 (5.4%)	36 (10.7%)	0.07
ED re-evaluation	1 (0.6%)	3 (0.9%)	0.67
CXRs ordered	6 (3.6%)	25 (7.4%)	0.07
Positive CXRs (b)	0 (0.0%)	3 (0.9%)	0.3
Test set 2			
CRP value, mean, mg/dL	29.4	27.4	1
Antibiotics prescribed	33 (10.9%)	68 (18.7%)	<.05
PC re-evaluation	30 (9.9%)	35 (9.6%)	0.9
ED re-evaluation	1 (0.3%)	4 (1.2%)	0.38
CXRs ordered	8 (2.7%)	15 (4.1%)	0.4
Positive CXRs (b)	0 (0.0%)	2 (0.6%)	0.5

4.2.3 Outcome distribution in the 25-repeated nested CV of the training dataset

When training the RSTM, we performed 25 repeats of a 4-fold nested CV on the training dataset. We analyzed each run statistically and calculated each outcome's mean and 95% confidence intervals. Figures 9-15 display the patient count and outcome distribution results.

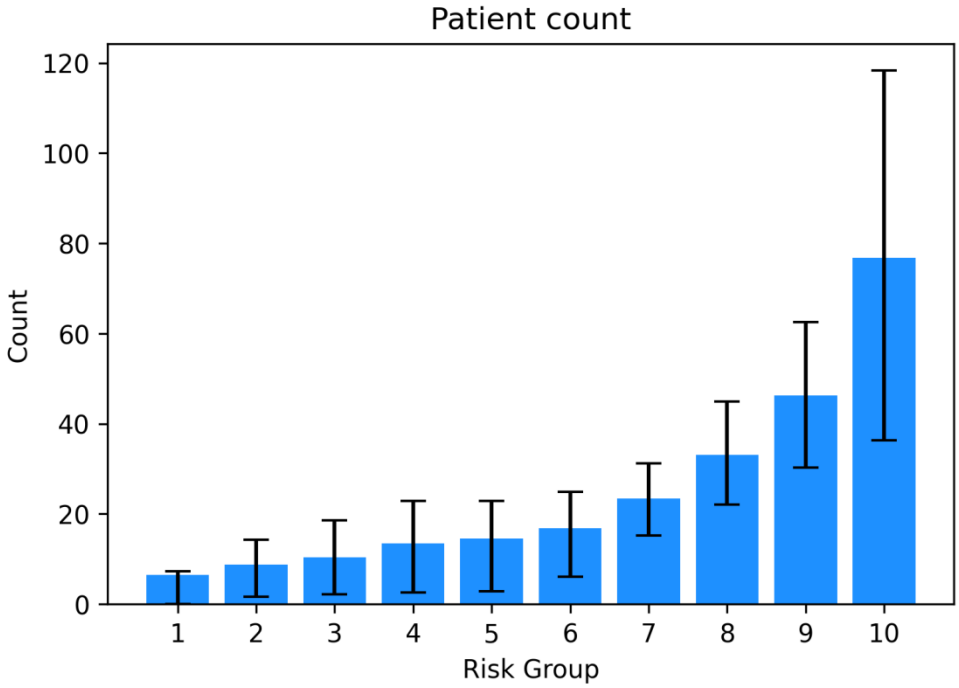


Figure 9. The patient count distribution, with 95% confidence intervals, by risk groups observed during 25 repeats of a 4-fold nested CV of the training dataset. Groups 1-5 are low-risk groups and 6-10 are high-risk groups

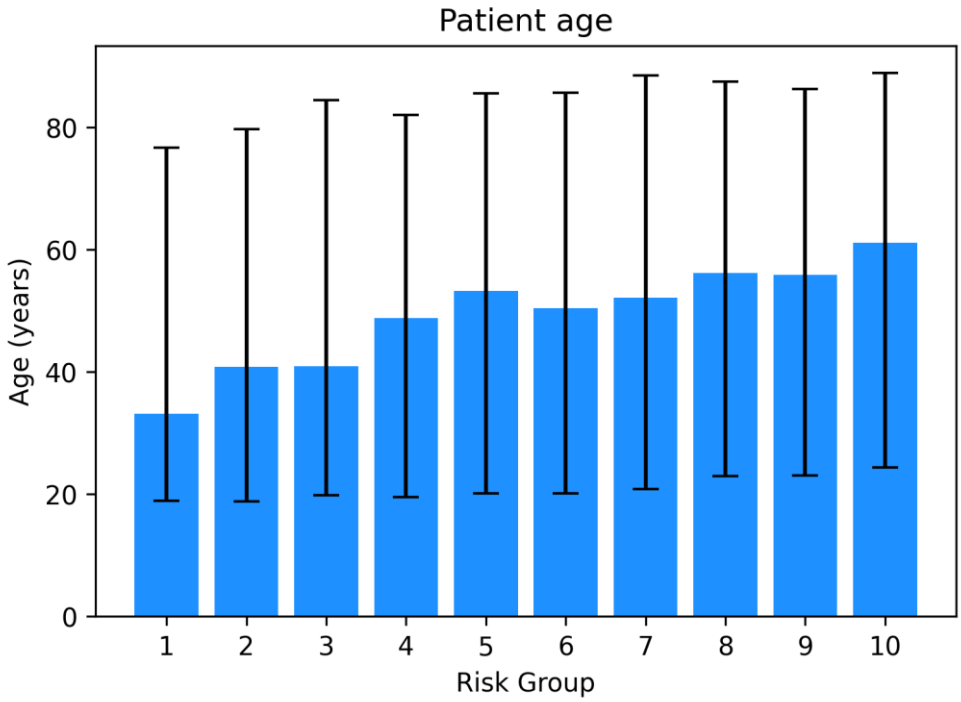


Figure 10. The patient age distribution, with 95% confidence intervals, by risk groups observed during 25 repeats of a 4-fold nested CV of the training dataset. Groups 1-5 are low-risk groups and 6-10 are high-risk groups

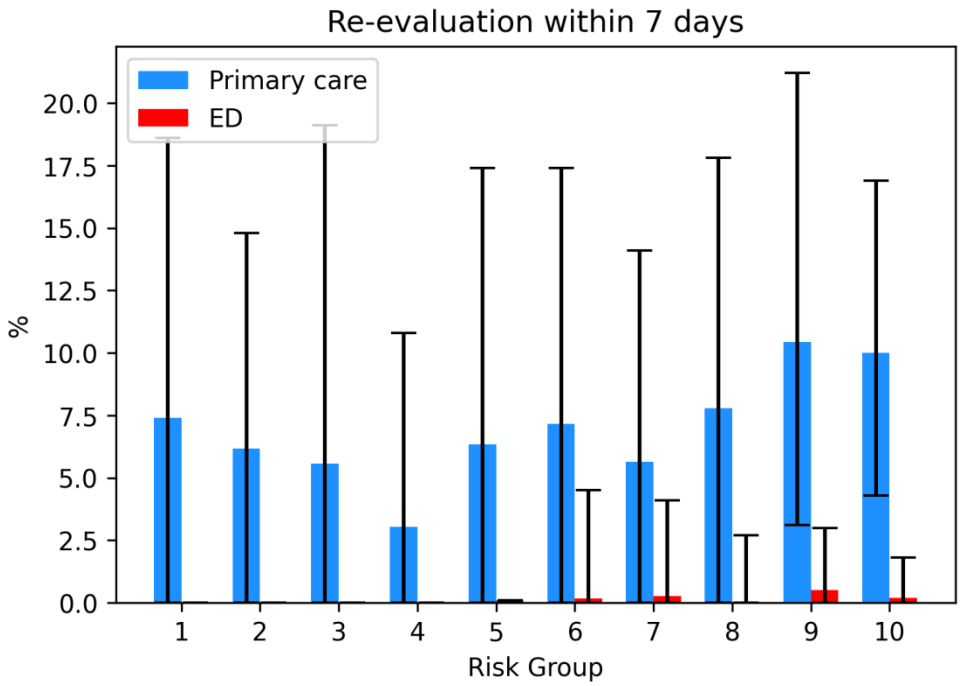


Figure 11. The mean re-evaluation rates within 7 days in PC and ED, with 95% confidence intervals, by risk groups observed during 25 repeats of a 4-fold nested CV of the training dataset. Groups 1-5 are low-risk groups and 6-10 are high-risk groups. PC, primary care; ED, emergency department.

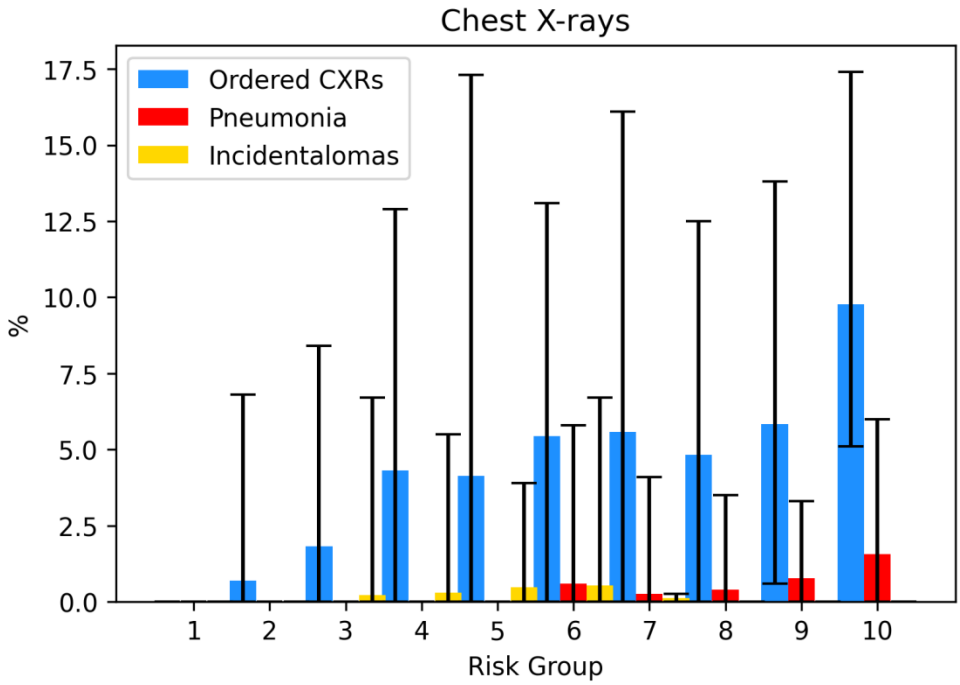


Figure 12. The mean CXR referral, pneumonia, and incidentaloma rates, with 95% confidence intervals, by risk groups observed during 25 repeats of a 4-fold nested CV of the training dataset. Groups 1-5 are low-risk groups and 6-10 are high-risk groups. CXR, Chest X-ray;

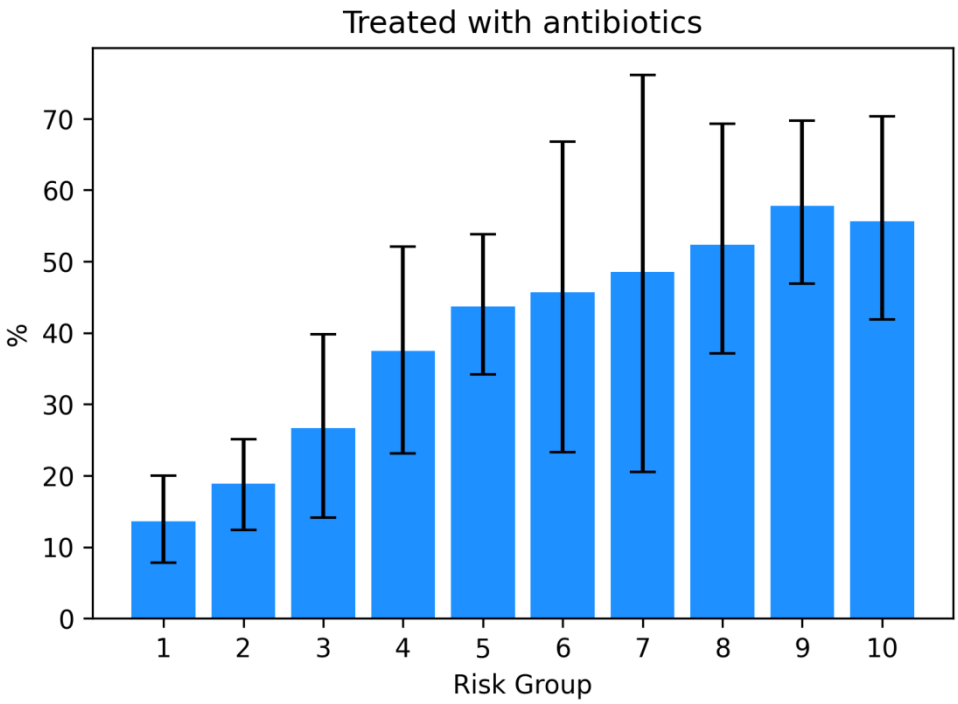


Figure 13. The mean antibiotic treatment rates, with 95% confidence intervals, by risk groups observed during 25 repeats of a 4-fold nested CV of the training dataset. Groups 1-5 are low-risk groups and 6-10 are high-risk groups.

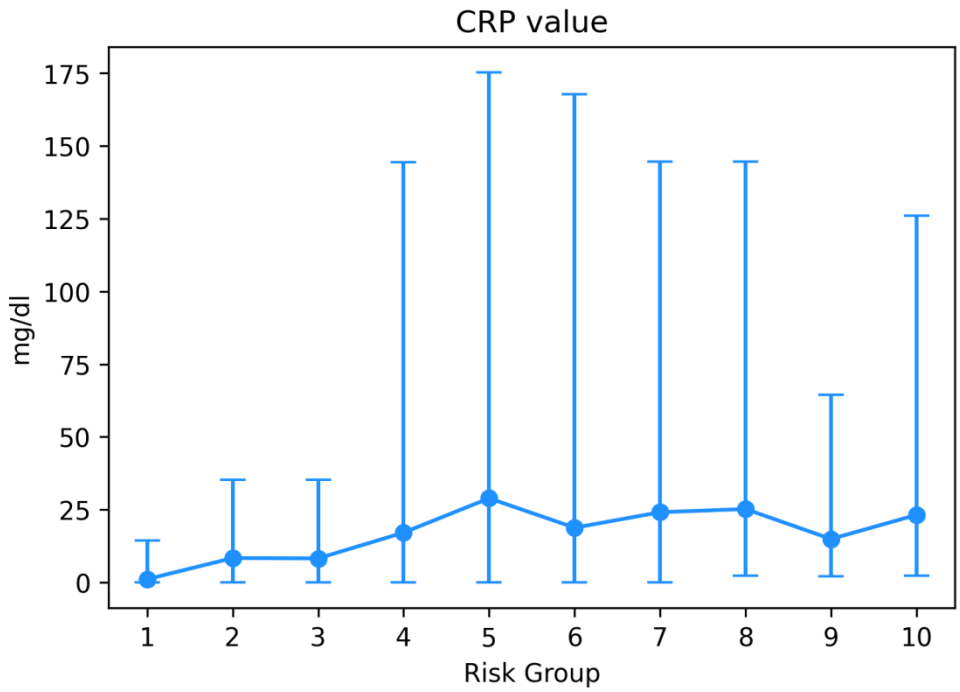


Figure 14. The mean CRP values, with 95% confidence intervals, by risk groups observed during 25 repeats of a 4-fold nested CV of the training dataset. Groups 1-5 are low-risk groups and 6-10 are high-risk groups.

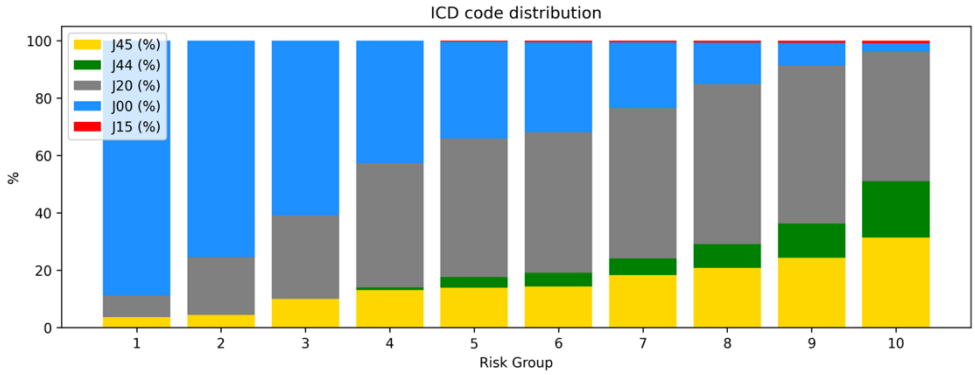


Figure 15. The mean ICD code distribution observed during 25 repeats of a 4-fold nested CV of the training dataset. Groups 1-5 are low-risk groups and 6-10 are high risk groups.

Age and patient count increase, approaching the higher risk groups. The LR groups contain 52% of incidentalomas, 9% of CXR referrals, and no CXRs positive for pneumonia. No patients went to the ED within 7 days in groups 1-4. All outcomes display an increasing percentage approaching higher-risk groups. Of the selected ICD codes, J00 represents patients with the mildest symptoms, dominating the lower-risk groups and diminishing approaching the higher-risk groups. J20 takes up the second most space in the lower risk groups but increases in the middle but then shrinks again in groups 9 and 10. J45 is more diagnosed in the lower risk groups than J44, which starts to increase after risk group 5, and both increase from there, approaching risk group 10. There were only patients diagnosed with J15 in groups 6-10.

4.2.4 The outcome distributions in test set 1

After training the model, we gave the model test set 1 as input and analyzed the corresponding output statistically. Figures 16-22 show the patient count and outcome distribution in test set 1.

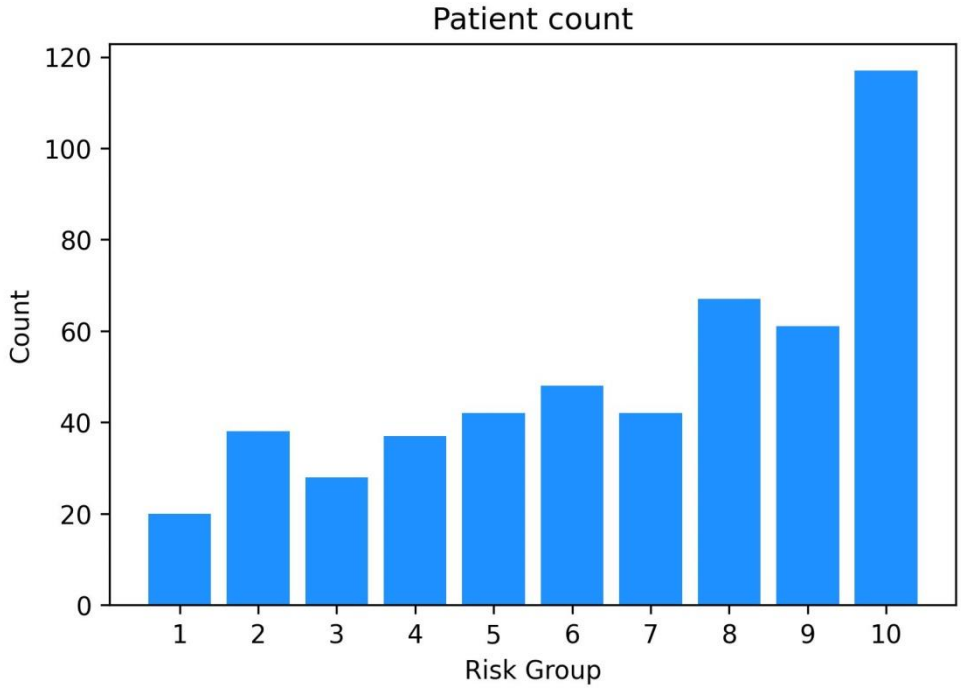


Figure 16. Patient count distribution by risk groups observed in test set 1. Groups 1-5 are low-risk groups and 6-10 are high-risk groups.

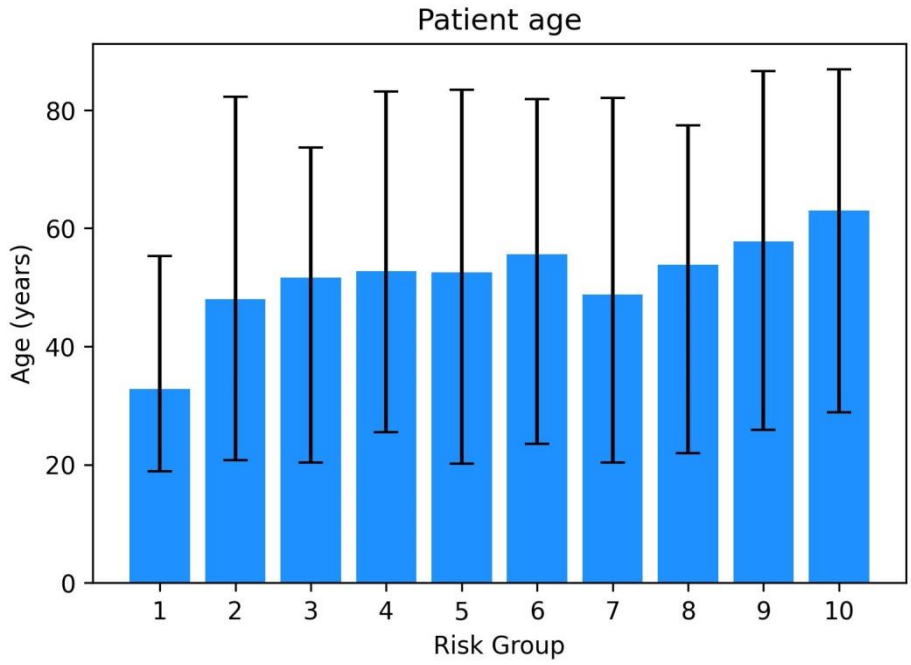


Figure 17. Patient age distribution by risk groups observed in test set 1 with 95% confidence intervals. Groups 1-5 are low-risk groups and 6-10 are high-risk groups.

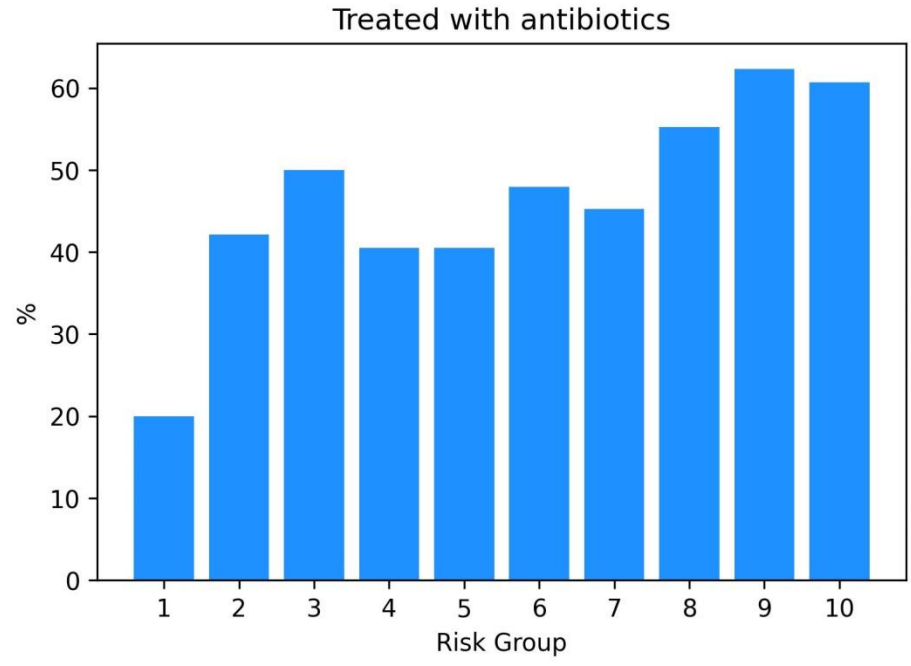


Figure 18. Distribution of antibiotic treatments by risk groups observed in test set 1. Groups 1-5 are low-risk groups and 6-10 are high-risk groups.

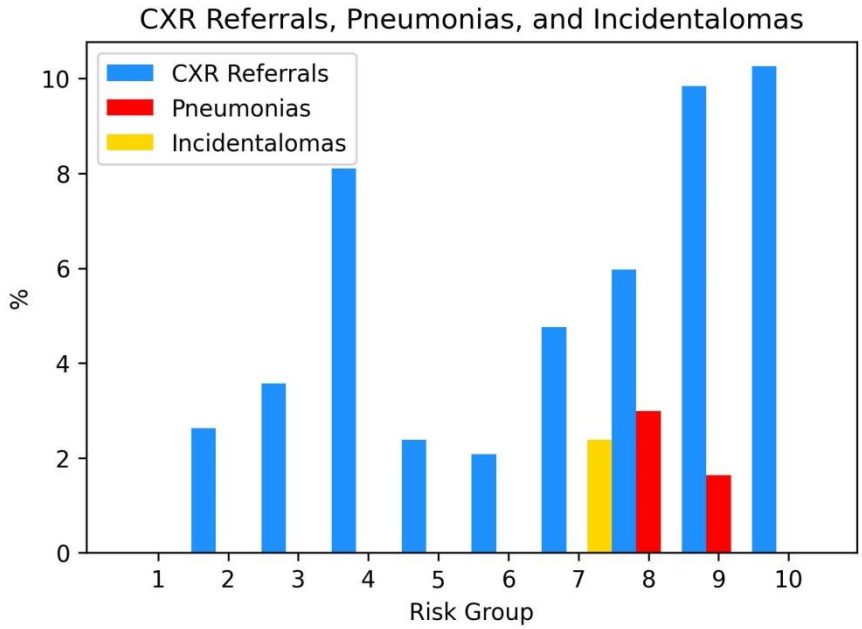


Figure 19. The distribution of CXR referrals, pneumonia, and incidentalomas by risk group was observed in test set 1. Groups 1-5 are low-risk groups and 6-10 are high-risk groups. CXR, Chest-X-Ray.

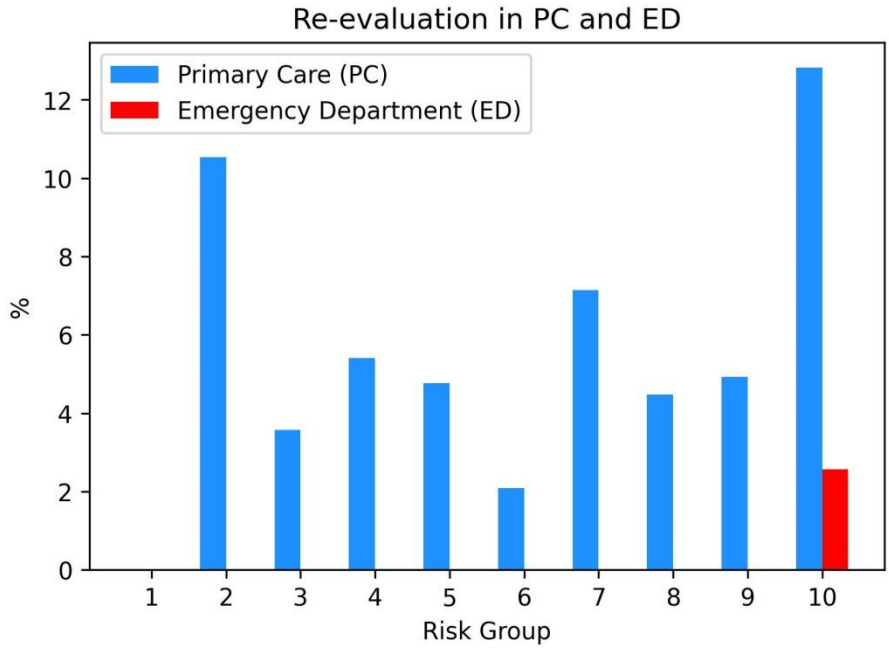


Figure 20. The PC and ED re-evaluation rates observed by risk groups in test set 1. Groups 1-5 are low-risk groups and 6-10 are high-risk groups. PC, Primary Care; ED, Emergency Department.

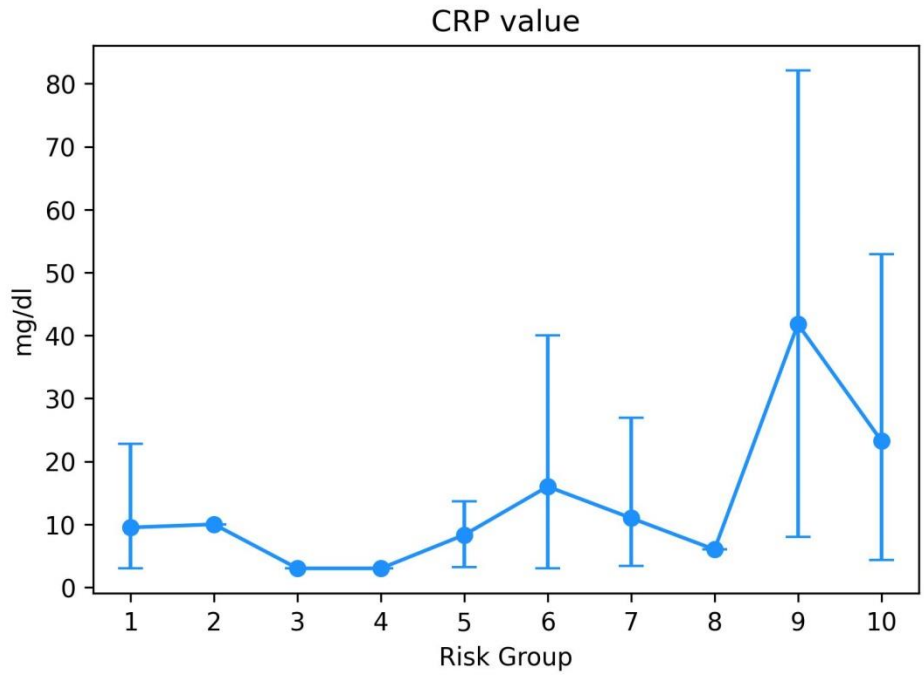
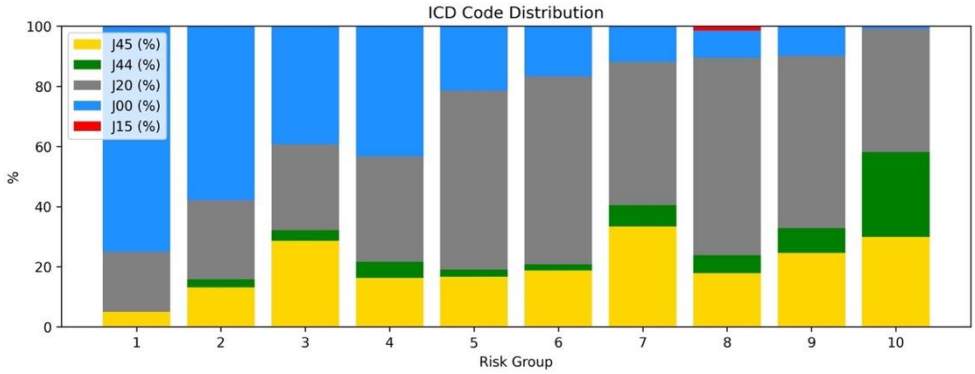


Figure 21. The mean CRP value distribution by risk group observed in test set 1, with 95% confidence intervals. Groups 1-5 are low-risk groups and 6-10 are high-risk groups. CRP, C-Reactive Protein.

Figure 22. The ICD code distribution by risk group observed in test set 1. Groups 1-5 are low-risk groups and 6-10 are high risk groups.



The age and patient count increase, approaching higher-risk groups. There were no pneumonia diagnoses, CXRs positive for pneumonia, or ED evaluations in the LR groups. Antibiotic treatment incidence increases from group 1 to 10. The CRP value is low in groups 1-8, rising sharply in groups 9 and 10. The distribution of ICD codes follows a similar pattern to that described in Section 4.2.1. Code J00 is prevalent in lower-risk groups, gradually decreasing as we approach risk group 10. J20, initially the second most common in lower-risk groups, peaks in the middle and then declines

nearing risk group 10. J45 shows a steady increase from risk group 1 through to risk group 10. J44 first appears in risk group 2, increases progressively, and is most prominent in risk group 10. Notably, J15 is present only in risk group 8.

4.2.5 The outcome distributions in test set 2

We input test set 2 and analyzed the corresponding output statistically. Figures 23-28 show the patient count and outcome distribution in test set 2.

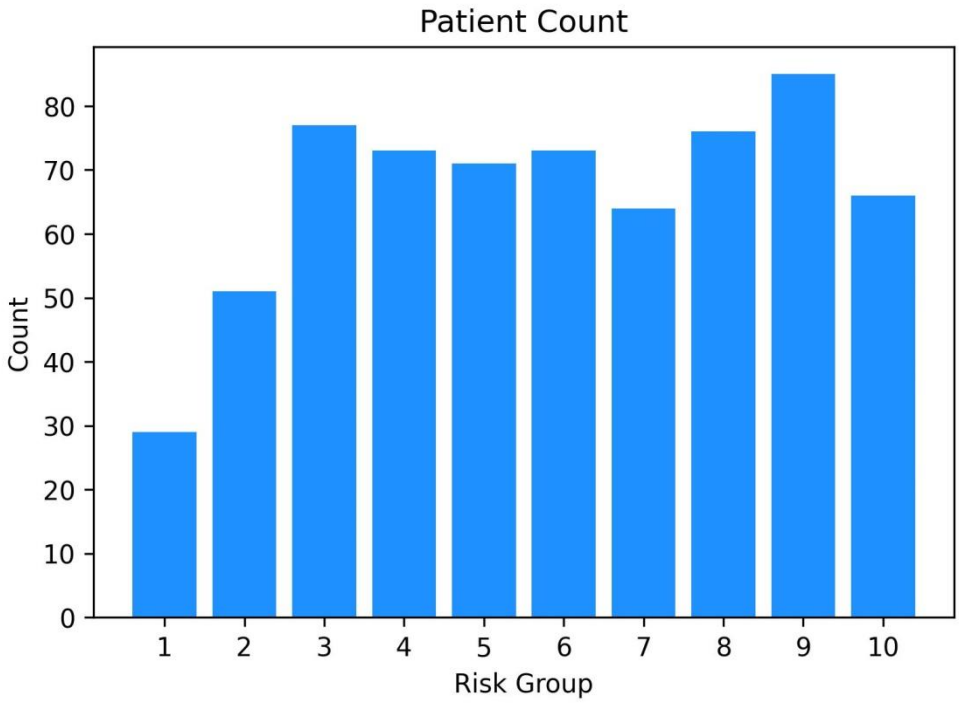


Figure 23. The patient count distribution by risk groups observed in test set 2. Groups 1-5 are low-risk groups and 6-10 are high-risk groups.

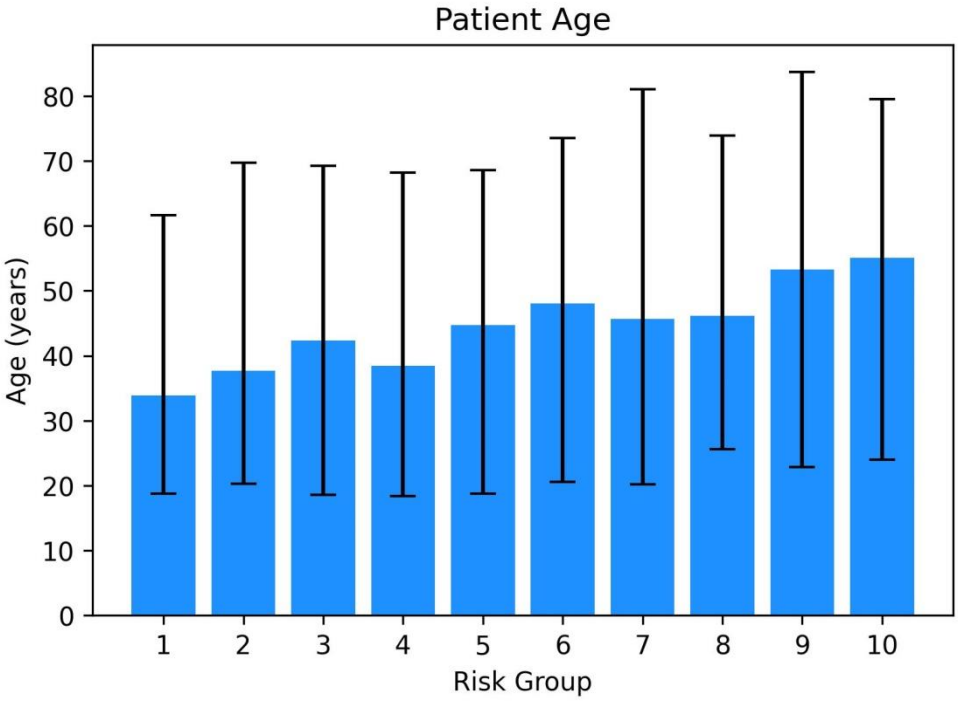


Figure 24. The patient age distribution by risk groups observed in test set 2. Groups 1-5 are low-risk groups and 6-10 are high-risk groups.

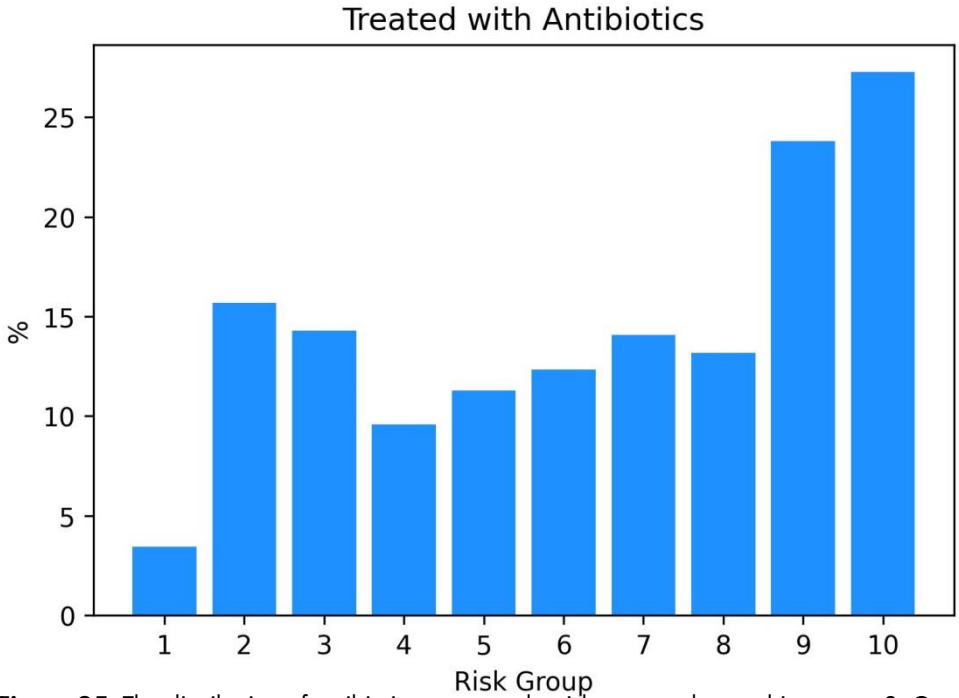


Figure 25. The distribution of antibiotic treatments by risk groups observed in test set 2. Groups 1-5 are low-risk groups and 6-10 are high risk groups.

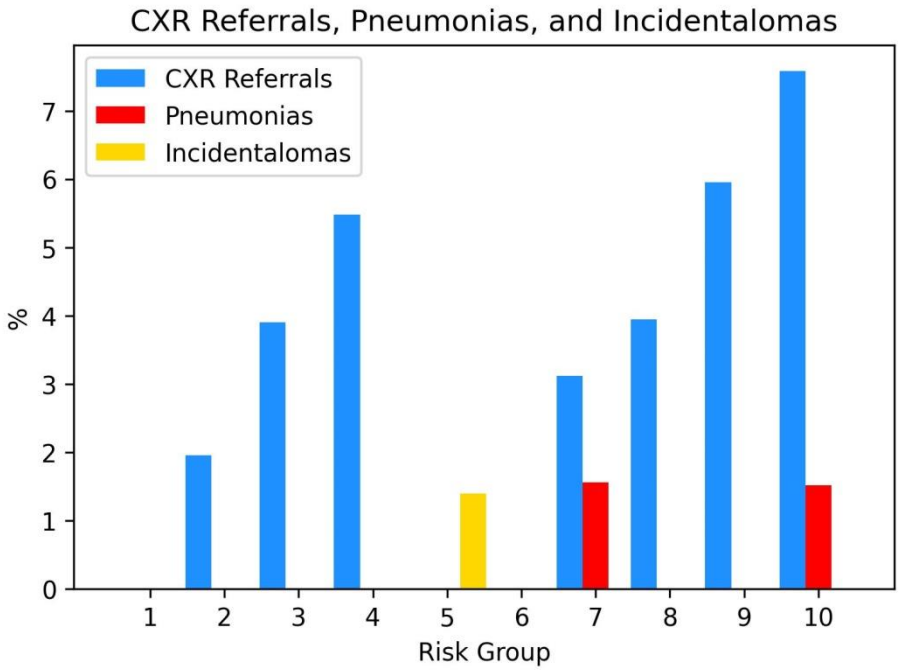


Figure 26. The distribution of CXR referrals, pneumonia, and incidentalomas by risk group observed in test set 2. Groups 1-5 are low-risk groups and 6-10 are high-risk groups. CXR, Chest-X-Ray.

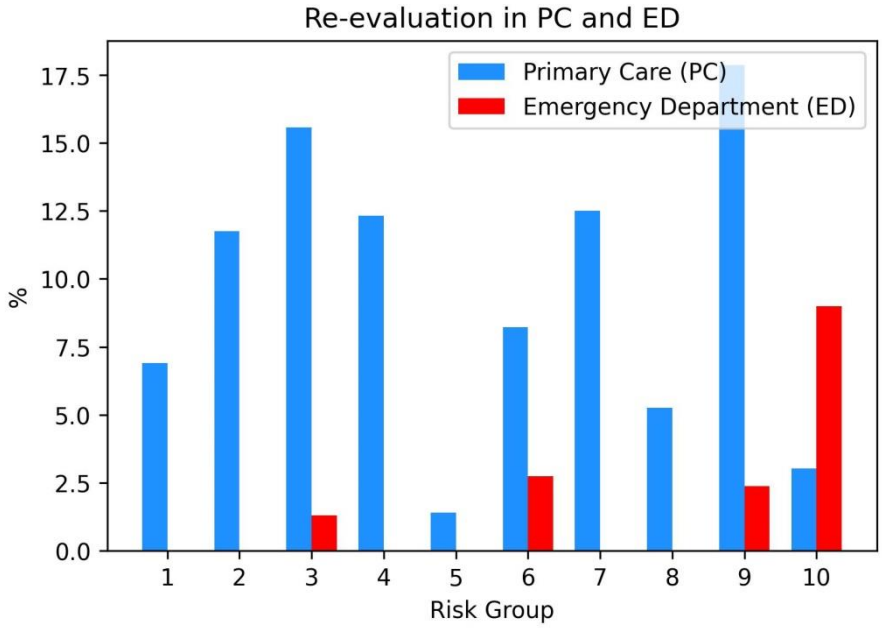


Figure 27. The PC and ED re-evaluation rates observed by risk groups in test set 2. Groups 1-5 are low-risk groups and 6-10 are high-risk groups. PC, Primary Care; ED, Emergency Department.

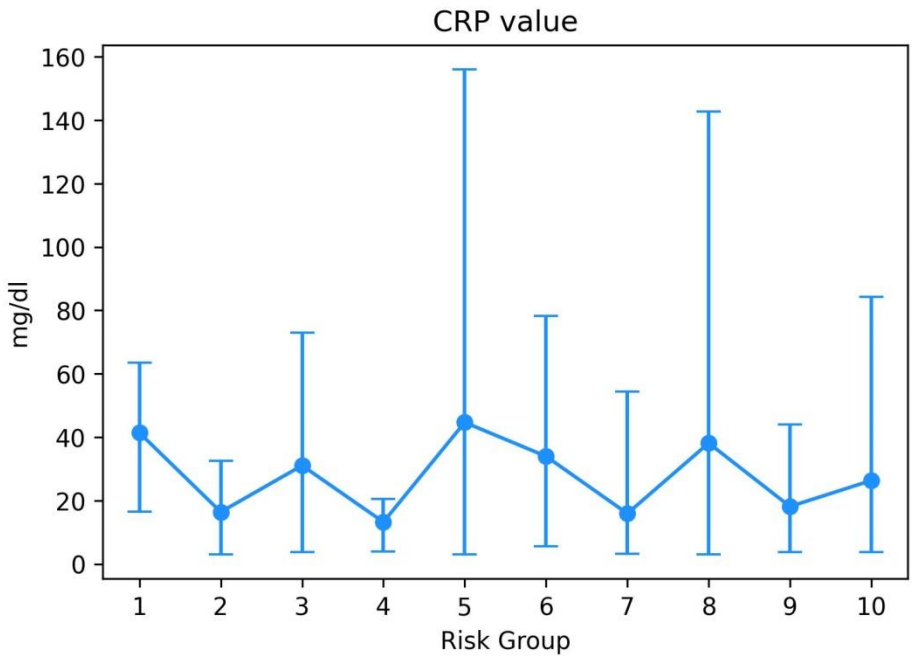


Figure 28. The mean CRP value distribution, with 95% confidence intervals, by risk group observed in test set 2. Groups 1-5 are low-risk groups and 6-10 are high-risk groups. CRP, C-Reactive Protein.

In the second test set, we observe many of the same trends as in the first. The number of patients and their ages is rising, nearing the HR groups. Antibiotic treatment shows a similar pattern. No CXRs indicating pneumonia are found in the LR risk groups; however, 35% of the CXR referrals and a single incidentaloma are within these LR groups. Unlike the first test set, the re-evaluation rate for PC does not exhibit a clear upward trend toward the HR groups, with only one ED evaluation recorded in the LR test set. Additionally, the CRP values demonstrate a more subdued upward trend compared to test set 1, which is anticipated since influenza patients generally have higher CRP values than those in the initial test set. No ICD code distribution analysis was conducted in the second test set since only two ICD codes were selected, which both signified influenza diagnoses.

4.2.6 Risk score comparison with Diehr et al. and Anthonisen et al.

Numerous studies have aimed to establish diagnostic rules for pneumonia. However, all but one incorporate variables from patient examinations or blood tests, rendering them incomparable to the RSTM. The only comparable study was conducted by Diehr et al.[83] We, therefore, calculated the Diehr risk score for all patients in our population and then compared the risk score of the RSTM to the normalized predictive score of the diagnostic rule that Diehr et al. created in test set 1. The results can be seen in figure 29. There is a strong correlation between the score from the RSTM and that of the Diehr score.

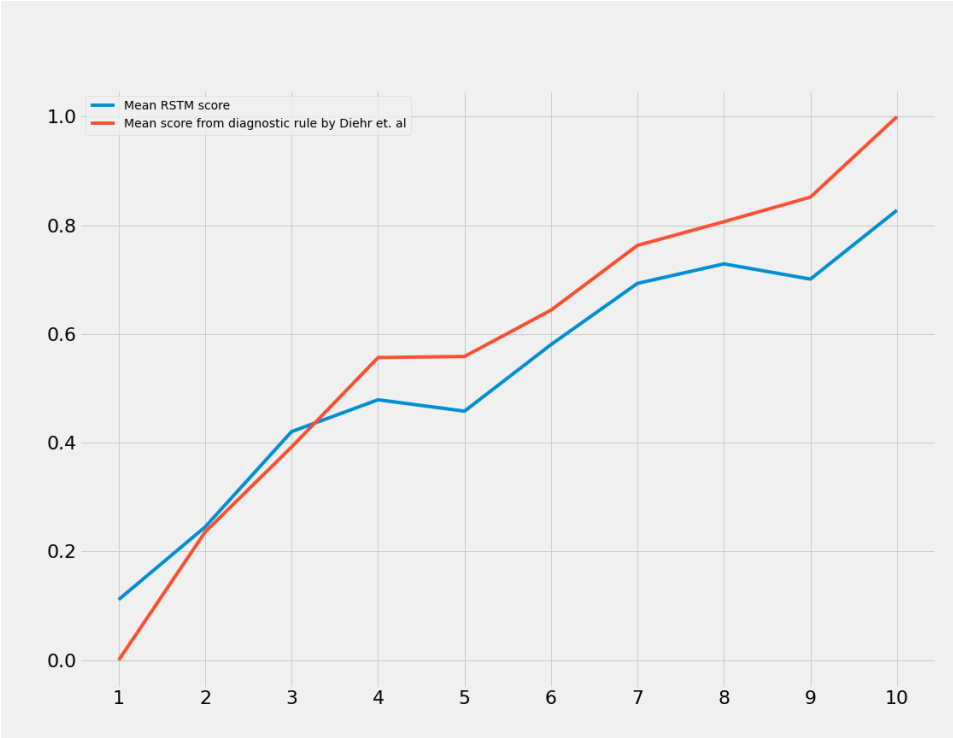


Figure 29. The mean score of the RSTM compared to the normalized score of the diagnostic rule for pneumonia created by Diehr et. al in by risk group. Groups 1-5 are low-risk groups and 6-10 are high-risk groups

In a study conducted by Anthonisen et al.[84], the researchers created and validated a trifecta of symptoms to predict whether a patient with a COPD exacerbation should be treated with antibiotics. They determined that patients exhibiting two or more of the three symptoms—increased dyspnea, sputum, and sputum purulence—benefited from antibiotic treatment. This suggests that COPD patients with a growing number of these symptoms are more likely to experience a bacterial exacerbation. We analyzed the COPD patients in test set 1 and calculated the Anthonisen risk score for

each risk group. The findings are displayed in Figure 30, which illustrates a rising number of patients displaying more positive cardinal symptoms nearing the HR groups. Only patients in risk group 10 should have been prescribed antibiotics according to their diagnostic rule.

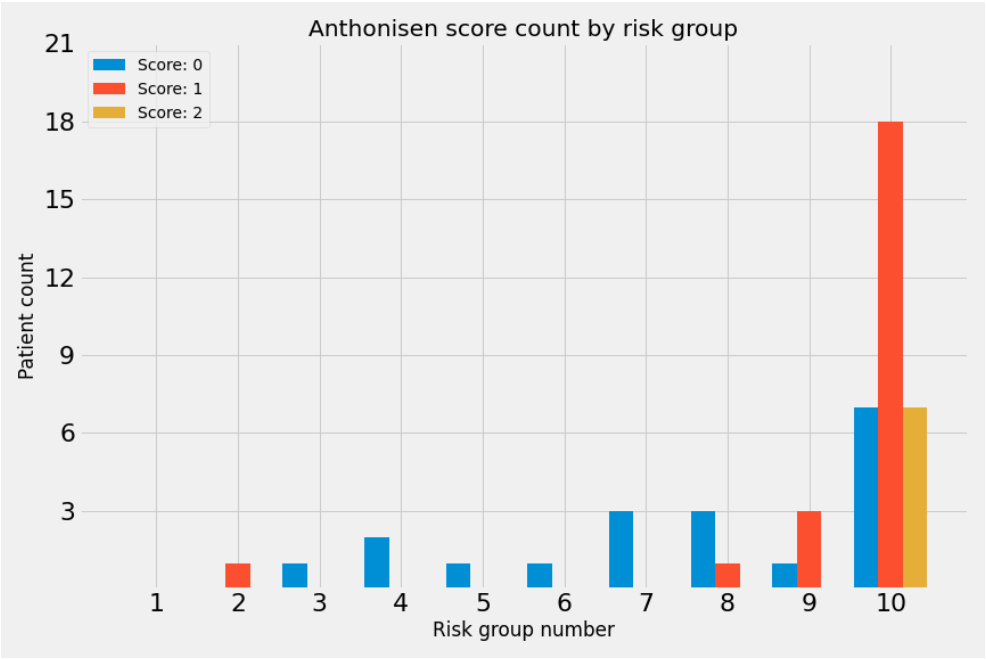


Figure 30. The Anthonisen risk score by risk group in test set 1. Patients with a score of 2 or higher should receive antibiotic treatment. No patient in test set 1 had a score of 3. Groups 1-5 are low-risk groups and 6-10 are high-risk groups

4.3 Study III

The purpose of study III was to evaluate the performance of the ML model trained in study II prospectively. The model in study III was modified slightly after study II was completed, as described in section 3.4.1.3.

4.3.1 Descriptive statistics

554 contacts from 512 patients were recorded and recruited in a single primary care clinic in Iceland. Twelve patients denied participating for various reasons and one patient could not participate because of language barriers. 302 patients with 337 patient contacts were filtered out based on presenting complaints, leaving 197 patients with 204 patient contacts. The flowchart in study III is displayed in Figure 31.

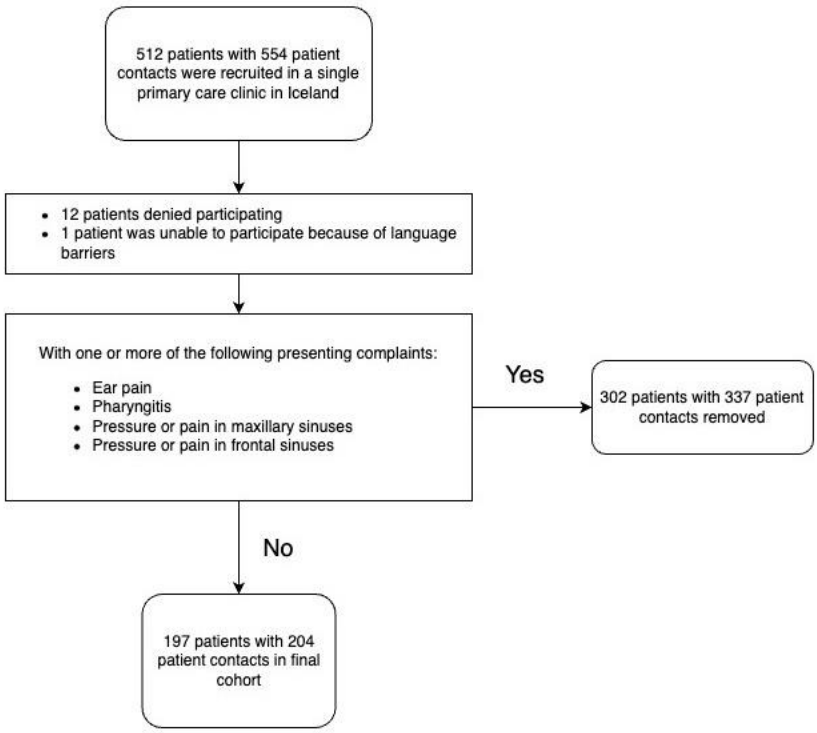


Figure 31. The flowchart in the third study.

Table 7 provides an overview of the demographics of the whole and final population after patients have been removed based on primary complaints. This methodology aimed to increase the number of patients appropriate for the RSTM in the final population. The RSTM was trained on patients diagnosed with J00, J20, J15, J44 and J45. The removal increases proportionally the number of patients suitable for the model while reducing the number of unsuitable patients. Patients with ICD codes such as B34, R05, J45, J20, J18, J22, and J40 are suitable and increased, while patients with J02 and J01 are unsuitable and decrease. J00 is suitable but decreases as well. Table 8 compares the outcome rates in the whole and final population.

Table 7. Comparison of the demographics of the filtered group to those of the entire population.

Variable	Filtered population	Whole population
Patient contacts, count	204	554
Mean age (min-max)	42.3 (18 - 88)	45.2 (18-88)
Female (%)	46.6	48.8
ICD-10 codes (%)		
B34 (Viral infection, unspecified)	25.7	20.8
R05 (Cough)	17.9	11.4
J45 (Asthma)	8.7	5.1
J20 (Acute bronchitis)	7.3	4.7
J18 (Pneumonia)	5.1	2.4
J32 (Chronic sinusitis)	3.7	2.6
J22 (Unspecified acute lower respiratory infection)	3.7	2.5
J40 (Bronchitis, not specified as acute or chronic)	3.2	5.2
J02 (Acute pharyngitis)	2.8	6.1
R06 (Abnormalities of breathing)	2.3	1.4
No ICD code	2.3	2.6
J01 (Acute sinusitis)	1.8	4.4
J00 (Acute nasopharyngitis)	1.4	2.8

Table 8. An overview of the outcomes in the filtered compared to the whole population.

Variable	Filtered population	Whole population
CRP value, mean (95% CI), mg/dL	31.53 (0.7 - 220.6)	28.69 (1 - 201.5)
Antibiotics prescribed, count (%)	51 (25.0)	178 (32.9)
Re-evaluation in PC within 7 days, count (%)	18 (8.8)	31 (5.7)
Evaluation in ED within 7 days, count (%)	5 (2.5)	5 (0.9)
ED referral, count (%)	4 (2.0)	5 (0.9)
Pneumonia ICD code, count (%)	14 (6.7)	20 (3.7)
CXR referrals, count (%)	19 (9.3)	24 (4.4)
CXRs pneumonia, count (%)	9 (4.4)	10 (1.9)
CXRs incidentaloma, count (%)	2 (1.0)	4 (0.7)
CT thorax referrals, count (%)	4 (2.0)	8 (1.5)
Lung cancer diagnoses, count (%)	1 (0.5)	2 (0.4)

4.3.2 Patient count and outcome distribution

Images 32-36 show the patient count distribution and outcome distributions for selected outcomes. Generally, the outcome distributions show increasing proportional values as

we approach higher risk groups. Groups 2-5 and 8-9 have relatively few patients, causing variability in their proportional outcome values. We analyzed the CRP values in each group, and the mean CRP value with 95% CIs is displayed in Figure 36. The patient diagnosed with pneumonia, who had a negative CXR in risk group 2, had a CRP of 112, causing the spike in risk group 2. One patient, diagnosed with J01 (Acute Sinusitis) in risk group 1, had a CRP of 81, causing the raised CI in group 1. Four patients had a CRP value above 130 and were all diagnosed with pneumonia, and three of them had positive CXRs as well. They were all in the HR group.

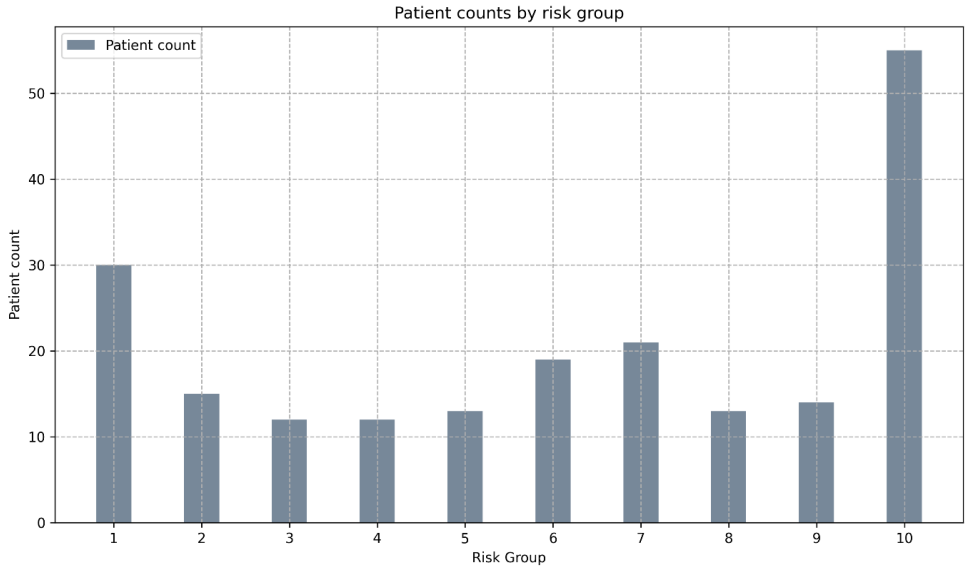


Figure 32. The patient counts distribution by risk group. Groups 1-5 are low-risk groups and 6-10 are high-risk groups

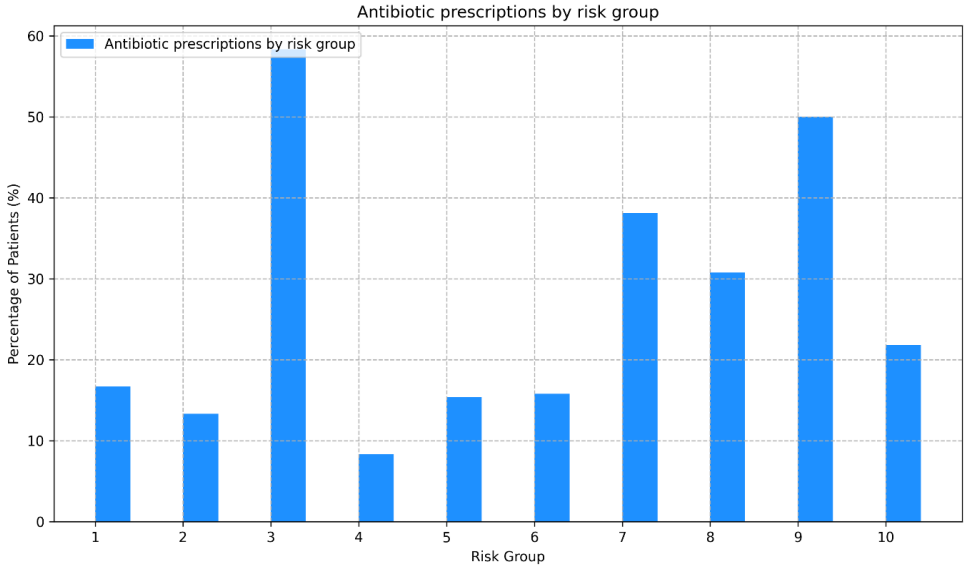


Figure 33. The percentage of antibiotic prescriptions by risk group. Groups 1-5 are low-risk groups and 6-10 are high-risk groups

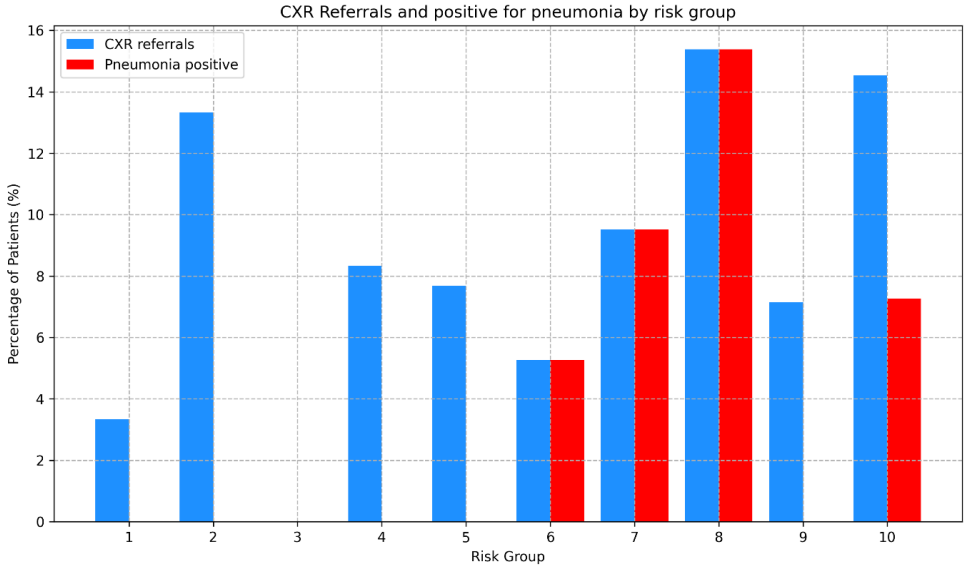


Figure 34. The percentage of CXR referrals and CXRs positive for pneumonia by risk group. Groups 1-5 are low-risk groups and 6-10 are high-risk groups. CXR, Chest X-Ray.

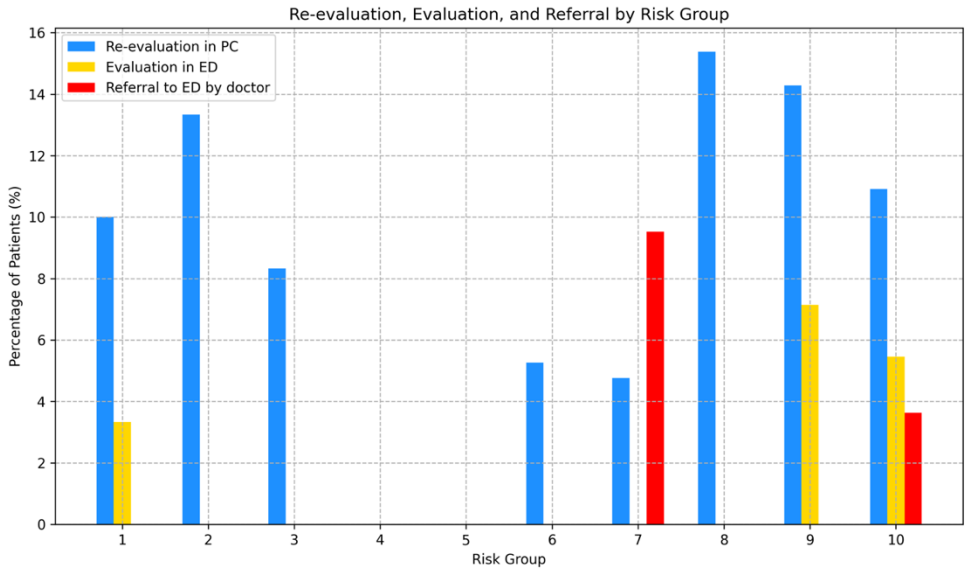


Figure 35. The percentage of patients who sought re-evaluation in ED or PC or were referred to the ED by the physician by risk group. Groups 1-5 are low-risk groups and 6-10 are high-risk groups. PC, Primary Care; ED, Emergency Department.

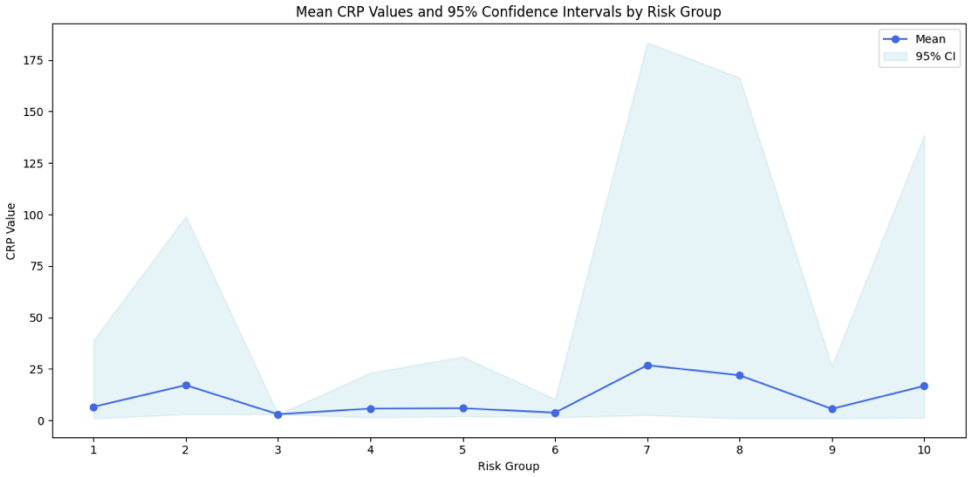


Figure 36. The mean CRP values with 95% confidence intervals by risk group. Groups 1-5 are low-risk groups and 6-10 are high-risk groups . CRP, C-Reactive Protein.

4.3.3 Comparison of outcomes in HR and LR groups

A comparison of the outcomes in the HR and LR groups is displayed in Table 9 with p-values and Odds Ratio (OR). 40% of the patients are in the LR groups. The mean CRP value is higher in the HR groups but not statistically significant. Patients in the HR groups were 50% more likely to receive a CRP measurement and 49% more likely to be prescribed antibiotics. They were also 38% more likely to be re-evaluated in PC within 7 days. No patients from the LR groups were sent to the ED, whereas patients in the HR groups had odds of referral to the ED that were 6.26 times higher, calculated with Haldane-Anscombe correction. 26% of the CXR referrals were in the LR groups, and all were normal. Patients in the HR groups had twice the odds of being referred to a CXR.

Table 9. Comparison of clinical outcomes between low-risk and high-risk groups**.** Statistically significant values are marked in bold, with significant findings in pneumonia diagnoses via ICD codes (p=0.009, OR=9.66) and CXR-confirmed pneumonia (p=0.03, OR=19.00).

*Corrected with Haldane-Anscombe correction

**Subsequent CXR was normal

***Includes only patients who were referred for a CXR by a physician

****Includes only patients who were referred for a CT by a physician

Outcome	Low-Risk Group	High Risk-Group	P-value	Odds Ratio
Count	82	122	N/A	N/A
Mean CRP values (95% CIs)	19.91 (0.5 - 95.0)	34.04 (1 - 228.6)	0.55	N/A
No. CRP measurements (%)	23 (29.6)	45 (36.9)	0.23	1.5
Antibiotics prescribed (%)	17 (20.7)	34 (27.9)	0.25	1.5
PC re-evaluation (%)	6 (7.3)	12 (9.8)	0.62	1.4
ED evaluation (%)	1 (1.2)	4 (3.3)	0.65	2.8*
ED referral (%)	0 (0)	4 (3.3)	0.15	6.3*
CXR referrals (%)	5 (6.1)	14 (11.5)	0.22	2.0
CXR incidentalomas (%)	1 (1.2)	1 (0.8)	0.47	0.3***
CXR pneumonia (%)	0 (0)	9 (7.4)	**0.03**	19.0***
Pneumonia ICD code (%)	1 (1.5)**	13 (9.6)	**0.009**	9.7
CT referrals (%)	1 (1.2)	3 (2.5)	0.65	2.0
Lung cancer diagnoses (%)	0 (0)	1 (0.8)	1	1.8****

Two incidentalomas were found: one in the LR group and one in the HR group. Both patients had a CT of the thorax as a follow-up, and neither incidentaloma had clinical significance. In the HR groups, 9 CXRs indicated pneumonia, accounting for 64% of the CXR referrals. Using the Haldane-Anscombe correction, OR is notably high at 19, with a p-value of 0.03, indicating statistical significance. The HR groups had 13 pneumonia diagnoses, while the LR groups had one, resulting in an odds ratio of 9.66 with a statistically significant p-value of 0.009. The single patient in the LR group diagnosed with pneumonia underwent a CXR as part of the study, which turned out to be normal. Likewise, two patients from the HR group also had normal CXRs as part of the study. Four patients were referred to a CT image of thorax, one in the LR groups and three in the HR groups, resulting in an OR of 2.04 for the HR groups. One CT image resulted in a lung cancer diagnosis in the HR group; the rest was normal.

4.3.4 Pneumonia analysis

When evaluating an ML model that stratifies patients in primary care with respiratory symptoms into risk groups, ensuring that pneumonia risk is low in the low-risk groups is paramount. Figure 37 shows the patient count in the population who were either diagnosed with pneumonia or had a positive CXR for pneumonia by risk group. All patients with positive CXRs for pneumonia were in groups 6-10. In risk group 2, one patient was diagnosed with pneumonia, but a subsequent CXR was negative. In group 9, two patients received pneumonia diagnoses despite negative CXR results.

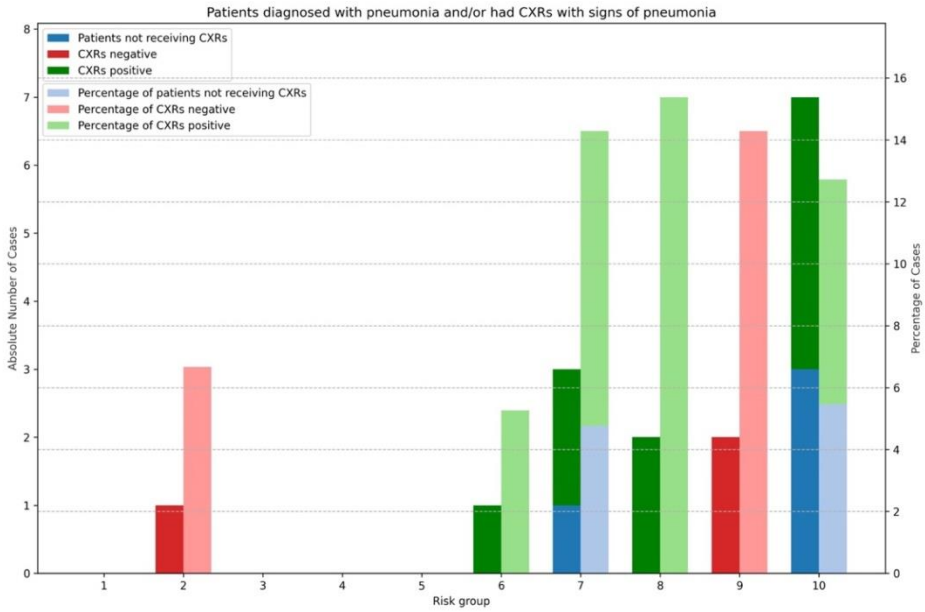


Figure 37. Distribution of pneumonia diagnoses and positive CXR results for pneumonia across risk groups**.** The left y-axis shows the absolute number of cases, while the right y-axis displays the percentage of cases. Groups 1-5 are low-risk groups and 6-10 are high-risk groups.

4.3.5 Outcomes in the removed patient's group

We also examined the patients excluded from the final analysis due to presenting complaints. Six patients were diagnosed with pneumonia, with two belonging to the LR groups, resulting in three pneumonia diagnoses among the LR groups in total. Two of these patients had a CXR, one as part of the study, the other referred by the diagnosing physician, and both CXRs were negative. One of the three patients did not have a CXR after being diagnosed with pneumonia. Figure 38 is identical to Figure 37 but includes all patients. Among the excluded patients, one was diagnosed with lung cancer in risk group 10. That brings the total number of lung cancer diagnoses to two patients, both in risk group 10.

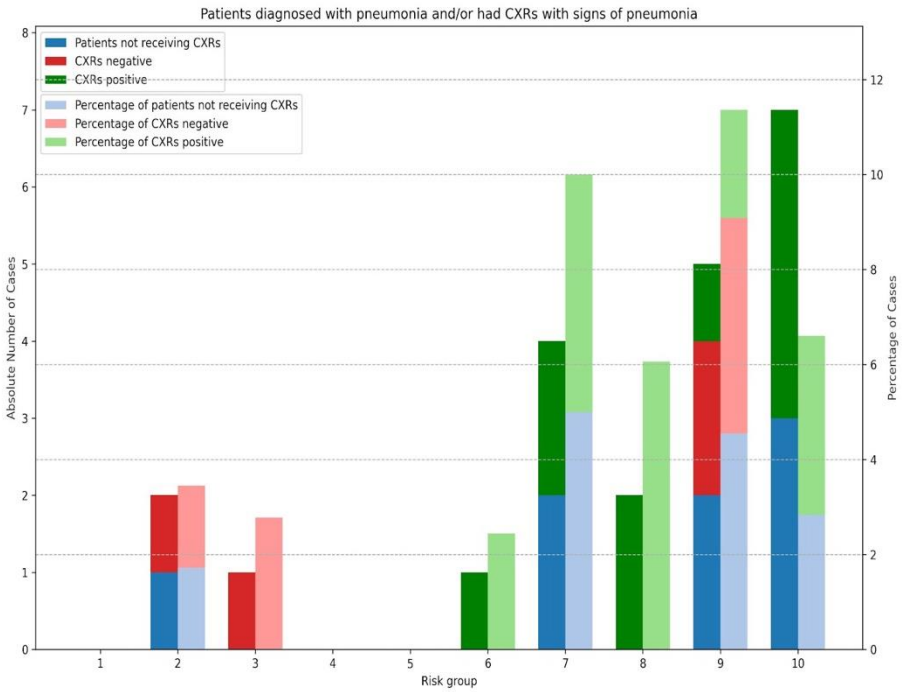


Figure 38. Distribution of pneumonia diagnoses and positive CXR results for pneumonia across risk groups for the whole population**.** The left y-axis shows the absolute number of cases, while the right y-axis displays the percentage of cases. Groups 1-5 are low-risk groups and 6-10 are high-risk groups.

4.3.6 Risk score comparison to Diehr et al.

We also compared the risk score of the patients in this study to that of the pneumonia diagnostic rule from Diehr et al. as we did in the second study. The result of this comparison is displayed in Figure 39. The correlation appears not to be as strong as in the second study. The Pearson correlation coefficient is 0.4424 which indicates a moderate positive correlation. One of the reasons for a weaker correlation is that we removed patients with pharyngitis symptoms. In the Diehr score symptoms of pharyngitis add -2 to the score.

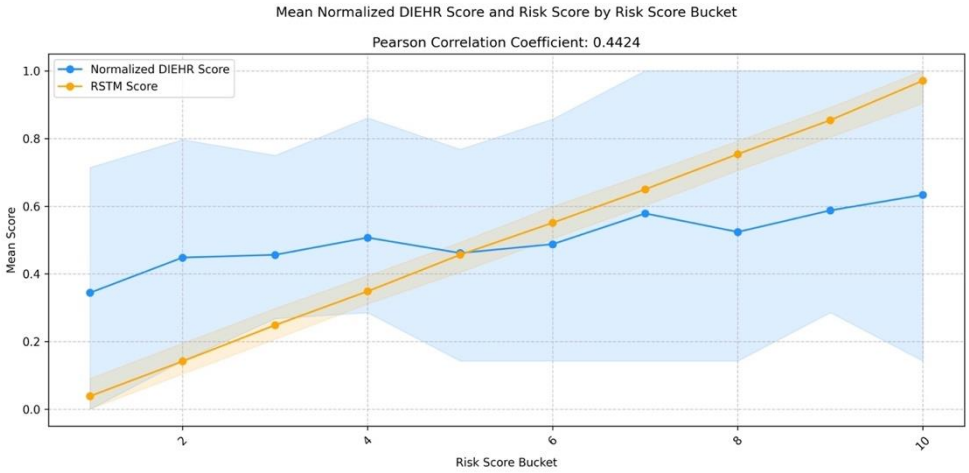


Figure 39. The correlation between the RSTM risk score and the normalized Diehr score with 95% CIs.

5 Discussion

This thesis uses methods of ML, a sub-category of AI, for diagnostic and outcome prediction in PC. The studies presented in this thesis show that the models trained on patient EHR data has the potential to support diagnostic processes within PC. They appear to reach similar, if not superior, diagnostic performance when it comes to diagnosing primary headaches as physicians. Furthermore, they seem to use similar input features as humans do when evaluating patients despite receiving a broad base of features as potential predictors. As the first study demonstrated, the model achieved a higher score than the physicians on all metrics except specificity. The MCC metric includes all four quadrants of the confusion matrix, so it can be considered the most critical metric, and the ML classifier outperforms all physicians on the MCC. The ML classifier struggles the most with distinguishing between TTH and MA-. The most likely explanation is that a subset of our patients has episodic TTH, where the symptomatic presentation is a mix of MA- and TTH features. Studies have shown that physicians also struggle to diagnose these patients correctly, often misdiagnosing them as having a simple TTH or MA-[85,86]. The thesis further illustrates that an ML model trained on patient EHR data, the RSTM, can triage patients safely and effectively. The RSTM appears to outperform doctors in certain cases, including two instances where physicians might have misdiagnosed pneumonia. In these cases, the model accurately classified the patients as LR, and both had negative CXRs, pointing to the absence of pneumonia. The third patient diagnosed with pneumonia in the LR groups did not have a CXR.

There is still much more to explore. While the RSTM can manage a subset of respiratory infections, ideally, we want to be able to triage patients with all types of respiratory diseases as well as other common non-infectious diagnostic groups in primary care. Theoretically, it should be possible to triage all patient groups similarly to how the RSTM triages patients with respiratory symptoms. It remains to be seen if multiple models need to be developed, each tailored to a single diagnostic group, or if a model trained on patient data from a broader selection of ICD codes can be created. Using chief complaints to identify patients unsuitable for the RSTM is an innovative yet unvalidated methodology. This method proportionally increases the number of patients in the final population suitable for the RSTM. If we want to be able to validate ML models in clinical settings prospectively, trained on patient data from a subgroup of patients, we need to be able to find out beforehand which patients are suitable for the model, and which are not. This emphasizes the need for effective prospective validation methods within specific patient subgroups when working with ML models in clinical settings and methods to segregate patients fitting for a particular ML model. As ML models are applied to diverse diagnostic groups with large cohorts and a set of

broad and complex outcomes, the challenges of interpreting the results and analyzing their inner workings increases significantly.

Choosing the proper ICD codes for training ML models is not simple. As patients experiencing sinusitis, pharyngitis, and otitis present with a vastly different clinical presentation than the patients diagnosed with the diagnoses selected for training the RSTM, we excluded these infections since including them would make it harder to analyze the outcome distributions. The selected group of diagnoses differs in severity on the same axis, starting with J00 as the least severe, followed by J20, J45, J44, and J15, each encompassing the possibility of a progressively worse outcome. Adding one of the other infections to the selected diagnoses would introduce a new axis to the model, and the outcome distributions would no longer be on a singular axis. The study design outlined in the last two studies allowed us to achieve a clearer and more robust validation of the RSTM. If we had included other groups of respiratory infections, the outcome distributions would have been less clear, making it harder to evaluate the RSTM based on outcomes. In the prospective study, doctors used a much broader cohort of diagnoses than the ones we selected to train the RSTM. Symptomatic diagnoses were a large part of the diagnoses used by the physicians in the study. The assignment of ICD codes by doctors is often subjective, and their accuracy and performance metrics can vary significantly based on the underlying diagnoses[87]. This is especially true for symptomatic ICD codes such as R05, often used when patients present with cough and mild respiratory infection symptoms, and unspecific ones such as B34.9, which signifies an unspecified viral infection and can include many other viral etiologies than respiratory viruses. Some patients diagnosed with J00 could have been diagnosed with J20 and vice versa. The same can be stated for J45 and J20. Thus, when selecting the ICD codes, they can contain patients with a much broader range of conditions than the selected ICD codes indicate. This could be the reason why the RSTM was able to triage influenza patients correctly since the training data captures the whole spectrum of respiratory symptoms, even though we did not include all ICD codes for numerous respiratory infections.

The clinical guidelines for antibiotic treatment for the group of patients in the RSTM training data are clear: patients with bacterial infection, mainly pneumonia, are to be prescribed antibiotics, but if a viral etiology is suspected, no antibiotics should be prescribed. The main problem is knowing how to differentiate between patients who truly have a bacterial infection and those who have a viral infection. Therefore, we have a group of diagnoses where patients with pneumonia and some severe exacerbations of COPD are to be treated with antibiotics, but the rest have no benefit of antibiotics treatment. In contrast, distinguishing between viral and bacterial sinusitis is more complicated.

When selecting the initial group of ICD codes, we did consider starting with other ICD codes than we finally chose. After careful consideration, the ICD codes we chose were

the most feasible ones because of several factors. The major categories of respiratory tract infections in primary care include pharyngitis, otitis, sinusitis, and upper and lower respiratory infections, which do not belong to the previously mentioned groups. Our decision considered the current availability of clear clinical guidelines, and the extent to which these guidelines rely solely on symptoms, the incidence of ICD codes in adult populations, along with the ease and precision of differentiating outcomes where it is preferable to classify patients as LR or otherwise HR. Considering sinusitis, the gold standard for distinguishing between bacterial or viral infection is a sinus puncture followed by aspiration and bacterial culture. However, due to its invasive nature, this procedure is rarely utilized in primary care[88]. Likewise, training a triage model for patients with pharyngitis did not appear to be a good starting point. The most common bacterial cause, Group A Streptococcus (GAS)[89], is more prevalent in pediatric populations where carrier status is also high[90]. This makes it even more complicated to differentiate between GAS carriers and patients with viral pharyngitis. Similarly, the prevalence of otitis media is significantly higher in pediatric popluations than adults[91,92].

Adding additional groups of respiratory infections to the RSTM may vary in difficulty. Adding other diagnostic groups will depend on the structure of clinical guidelines for each infection group. It is easier to add groups of respiratory tract infections to the RSTM, where clinical guidelines are mainly based on symptomatic histories, than to train the RSTM on additional patient data from sinusitis patients. If we look at sinusitis for example, treatment guidelines primarily rely on symptomatic history, which can be obtained without a patient examination, like the variables in the RSTM. According to the Swedish Strategic Programme for the Rational use of Antimicrobial Agents (STRAMA) guidelines[93], the patient's symptoms steer the diagnosis and treatment. Duration of symptoms, colored nasal discharge, biphasic symptom course, and subjective measures of pain guide if the patient should see a doctor. The doctor visit might involve a CRP measurement or an X-ray referral. Subsequently, antibiotics should only be prescribed based on the patient's reported symptoms and diagnostic test results. Thus, the RSTM could be enhanced to determine whether patients with sinusitis require examination in primary care when symptoms are mild and short in duration. Training the RSTM on additional patient data will also likely be difficult since we do not know how the coefficients would change if with added patient data.

5.1 Clinical scores

A clinical score is a quantitative or qualitative tool used in healthcare to assess a patient's condition, risk, or likelihood of having a specific disease or outcome. It often combines multiple clinical parameters, symptoms, or test results into a single value or classification to support decision-making in diagnosis, prognosis, or treatment. Other terms than clinical score are used interchangeably, such as prediction algorithm, clinical decision rules, risk score, or scoring tool[94] but we will stick to the term clinical score. The emphasis on subjective evaluation in clinical settings has resulted in a great

proliferation of clinical scoring algorithms. A Pubmed search carried out to identify clinical scoring systems showed 250,000 publications since 1965 and an almost exponential growth in the number of publications since 1995. Perhaps this growth is related to the progress in ML from the 1990s as was discussed in the Introduction. Many well-known scores are widely used in clinical settings, but in recent years, the increase in poorly validated scores has been much larger[95].

However, the question arises of whether clinical scores translate to real improvement in patient outcomes. The literature is ambivalent, but it probably depends on how well they have been validated. For example, a well-known and validated clinical score, the modified CENTOR score[96] has been shown to reduce antibiotic prescriptions for pharyngitis in GPs[97]. The strongest evidence for clinical scores appears to be in cases where certain conditions can be excluded, such as in the case of the CHA_2DS_2-VASc score for stroke risk[98] among patients with atrial fibrillation, multiple known clinical scores for appendicitis risk[99] and the Wells Criteria for deep venous thrombosis[100]. Our findings, in the case of the RSTM, are very similar. It might, perhaps, be most effective as a tool of exclusion within a diagnostic group where clearer guidelines on diagnosis, use of diagnostics tests and treatments are needed. Likewise, within that group, the proportion of patients with symptoms that do not need a clinical visit according to healthcare professionals exceeds 70% in some studies[101]. Furthermore, unnecessary antibiotic prescribing is high[102–105]. The extrapolation of clinical risk scores to other populations must also be considered. A risk score developed in Iceland, a country with a relatively small homogeneous population, would have to be validated in different countries before it could be used. For example, the coronary risk score derived from the Framingham study has overestimated the cardiovascular risk in British[106,107] populations. As has already mentioned, designing clinical scores to make them accessible is of immence importance. If tools such as high-quality clinical guidelines and clinical scores are inaccessible, the usefulness of them decreases, and the clinical worker needs more time to retrieve information needed for the clinical rule. Lack of time among healthcare personnel in clinical settings, exacerbates this issue further[108].

Our choice of using CXRs for improving detection of pneumonia increases the quality of the third study, but higher-quality diagnostic images could also have been used. In general, pneumonia is divided into two groups based on where the patient is thought to have contracted the infection, i.e., Community-Acquired Pneumonia (CAP) as opposed to Hospital Acquired Pneumonia (HAP). The clinical guidelines and outcomes of these two groups differ significantly, and the clinical guidelines for CAP lack robust evidence[109]. The pneumonia patients in our study consisted solely of CAP managed in the community. Clinical guidelines from the British Thoracic Society on CXR referrals, when CAP is suspected, do not advise referral if the diagnosis of pneumonia is evident based on clinical signs and symptoms. Other diagnostic methods, such as CTs or ultrasound, could have been used, which appear to have higher sensitivity and specificity for CAP than CXRs[110]. The main downside of CTs is their higher cost and

radiation exposure. Imaging diagnostics are not infallible, and there is some uncertainty in CAP diagnosis, especially for CXRs to rule out the diagnosis of pneumonia[111]. CXRs are still the gold standard for pneumonia diagnosis[112]., so that was our imaging method of choice. There is also the possibility that some patients with negative CXRs would have developed signs of pneumonia on their CXR if given more time. The patients who were diagnosed with pneumonia and had negative CXRs were all treated with antibiotics, so we don't know what would have happened if they would not have been treated.

The most well-known pneumonia score, CURB-65, was excluded from our risk score comparison due to the unavailability of Blood Urea Nitrogen (BUN) measurements. Additionally, no patients had respiratory rates exceeding 30 or exhibited symptoms of confusion. Consequently, we determined that CURB-65 would not serve as a suitable marker for comparison.

5.2 Varying importance of clinical outcomes

The analyzed outcomes serve as varying indicators of a patient's true condition and from different perspectives. Some are subjective from the patients view and others from the doctor's view. Others are more objective. For instance, a positive CXR or elevated CRP levels more reliably indicate severe symptoms compared to a follow-up in primary care within a week. Follow-up rates reveal the patient's perception of their symptoms, whereas CRP provides an objective assessment linked to illness severity; additionally, a radiologist unfamiliar with the patient can more accurately interpret images to detect pneumonia than a GP diagnosing a patient based on clinical symptoms only. A GP may prescribe antibiotics for several reasons, including patient expectations, defensive medicine, or legitimate clinical needs. However, a GP will only refer patients to the ED when necessary. A patient seeking evaluation in ED indicates his own perception of more severe symptoms than when seeking re-evaluation in PC. Thus, this outcome reflects the patient's own perception of symptom severity but it can also mean other things. In Iceland, the waiting times for evaluation in PC can be long, and sometimes patients seek evaluation in the ED because of easier access but that is probably a rarity. ICD codes represent the symptom severity of the patient from the view of the doctor. However, as discussed in the beginning of this chapter, ICD codes are subjective, erroneous and biased and dependent on the doctor seeing the patient.

5.3 Findings from other studies

To the best of our knowledge, no studies exist identical to the studies presented in this thesis. The one that comes closest is the one we have already cited, which initially sparked the idea of trying the methodology described in this thesis[79]. There have been numerous attempts to create diagnostic rules and simpler models for pneumonia diagnosis. In section 4.2.6, we compared the risk score of the RSTM to the diagnostic rule for pneumonia created by Diehr et al. We also calculated the Anthonisen score for

the COPD patients, from Anthonisen et al., in our population. The scores from both diagnostic rules aligned well with the risk score of the RSTM. Many online symptom checkers also exist with variable degrees of accuracy, but research indicates that such tools generally have poor diagnostic performance[113–115]. Still, they are pretty popular with patients as an estimated 46% of patients check their symptoms online before seeing a healthcare professional, and the accuracy of the information they receive is low[116]. Offering scientifically based, high-quality information specific to everyone concerned about health-related issues could markedly impact the rate at which patients seek medical attention, with significant downstream effects on healthcare service quality and costs. In healthcare systems, where central triage agencies are increasingly established to influence patient flow with sub-acute problems, such as the 111 service in the UK and 1700 service in Iceland, high-quality CTMs could be instrumental.

5.4 Clinical implications

If the RSTM continues to demonstrate similar performance in future studies, the impact on patient flow and resource allocation within healthcare could be substantial. It would enable multiple new use cases that are currently not feasible. For example, it may allow the identification of patients with respiratory infections experiencing mild symptoms who only require symptomatic treatment before seeking primary care attention. Directing these patients through alternative pathways, rather than requiring a physical consultation, would reduce patient volume and alleviate pressure on the already strained healthcare system. The RSTM would fit well into central triage services which have been established in large western healthcare systems, such as the 111 service in the UK and the 1700 service in Iceland. Patients needing healthcare are asked to contact a centralized phone numbers or chat services to get advice, triage, and pointers on where they should go to get help for their present healthcare problems. Such services often employ nurses or other healthcare workers to triage patients through phone which is quite challenging to do accurately. Research has shown that nurse phone triage assisted by a CDSS system appears higher quality than GP triage[117] and same day consultations decrease significantly as a result of telephone triage[118]. If it is possible to develop scientifically validated ML models for patient triage, triage staff would get tools to evaluate patients much more accurately which might have significant effects on reduced consultations on already crowded primary care clinics. For patients that do consult with a healthcare worker, the RSTM can provide probabilities on the most important outcomes for patients experiencing certain respiratory symptoms. Allowing a doctor to know immediately when not to prescribe antibiotics or refer for a CXR is of high value in a high tempo environment where doctors often see 20-40 patients daily. Not to mention out-of- hours services where the volume of patients with respiratory symptoms is often much higher. The RSTM could provide physicians with a more standardized decision algorithm for when to refer adult patients for CXRs in cases of suspected pneumonia. Although some clinical guidelines exist for CXR referral, they are often vague and based on subjective estimations of pneumonia risk. This

subjectivity diminishes their utility, as it relies on variable physician judgment. The RSTM could enhance guideline adherence and ensure more consistent care by offering an objective risk assessment, immediately accessible to the physician as the patient enters the examination room.

The overuse of antibiotics for patients with respiratory tract infections in PC is a well-known problem. Increased antimicrobial resistance leads to an increased incidence of severe infections, complications, and longer duration of hospital stays. Methicillin-Resistant Staphylococcus Aureus (MRSA), a single type of resistant bacteria, causes more deaths each year in the United States than emphysema, HIV/AIDS, Parkinson's disease, and homicide combined[119]. About 70-80% of all human antibiotics are prescribed in PC[120]. About half of these are for patients with respiratory infections[121]. Studies indicate that, on average, general practitioners prescribe antibiotics in 38% of consultations for colds and upper respiratory infections, exceeding significantly what is warranted based on clinical symptoms[122]. Research also indicates that 50-75% of antibiotic prescriptions for acute cough and bronchitis are not justified and probably unnecessary[123]. Patients with acute cough and bronchitis are similar to the ones appropriate for the RSTM. In the third study, 33% of antibiotics were prescribed to patients in the LR groups. We have two patients diagnosed with pneumonia, and both had a negative CXR. We do not know the actual probability of a patient truly having pneumonia in the LR risk groups, as we did not have any radiologically positive cases of pneumonia in the LR groups. As the LR group had 82 patients, the highest possible value is 1,2% compared to 7,4% in the HR groups. Considering these numbers, it is interesting to contemplate if we might be able to eliminate antibiotic prescriptions in the LR group if the RSTM displays similar performance in a larger prospective study. The main question is if a risk benefit analysis shows the RSTM to achieve superior risk profiles compared to the traditional treatment pathways used today.

5.5 Strengths and limitations

This thesis, along with the studies underpinning it, has both strengths and limitations. The first study benefits from a large dataset of patients as well as benchmarking the performance of the model against the performance of GPs and GP trainees. The selection of a few ICD codes strengthens and limits the study. It includes all the most common primary headache diagnoses while allowing for easier interpretation of results. The current diagnostic selection makes the result less applicable to real clinical settings, and a broader approach is needed and preferably with patient outcomes as well. The fact that the quality of CTNs varies also limits the study, but to respond to this, we removed low-text quality CTNs as previously described. In that process, 84% of the CTNs were removed, creating a bias towards more complicated clinical cases being overrepresented in the final dataset. As physicians are more likely to register positive symptoms than negative, unmentioned CFs were assigned a binary value of 0, and a

CF was only annotated as positive when it was considered abnormal. We trained the classifier with missing CFs marked with a special value, but the results were the same.

The annotation by a single physician limits the first two studies. Annotation by multiple physicians, with cross-referencing, would have improved the quality of the study. Additionally, having physicians randomly sampling the annotations for auditing would have increased the quality. The models trained in these two studies use CTNs as source of training data. CTNs are written by human physicians and contain biases and errors from these humans which means that models trained on such data become erroneous and biased as well.

The second study also benefitted from an even larger dataset than the first study. With two separate large test datasets and outcomes to measure the RSTMs performance, the results are likelier to be more applicable to real clinical settings. Using a second test set from a patient distribution outside the training data distribution also tests the generalizability of the RSTM. The study was a retrospective one and suffers from the downsides and limitations of such types of studies. As in study I, removing CTNs information scarce CTNs likely creates biases towards more severe clinical cases.

The fact that humans create the data from which the ML models learn significantly influences the results of all studies. CTNs can be thought of as human-labeled data created by a large cohort of doctors. This fact should support the model's learning from a broad dataset of labeled diagnostic information. Studies have shown that individual errors and biases decrease with increased numbers of annotators[124], leading to more robust models. Using ML models as triage tools based on symptomatic history for patients with respiratory symptoms in primary care also appears promising. The third study further points to that conclusion, but a more extensive prospective study for validation is needed.

6 Conclusions

This thesis explores the possibility of using EHR data as training resources for ML models. It evaluates the ML model's performance both retroactively and in a limited prospective manner. This is a small but vital advancement toward guaranteeing that models such as the RSTM can be securely and effectively utilized in clinical settings. We conclude that ML models trained on EHR data have the potential to have a substantial positive effect on resource utilization within healthcare and are likely to improve patient outcomes simultaneously.

To advance this work further, more extensive prospective studies are necessary to determine better the likelihood of conditions such as pneumonia and lung cancer. Expanding the scope of research to include a broader range of outcomes and exploring severe manifestations of other respiratory tract infections, such as pharyngitis, otitis, and sinusitis, is equally important. Given the limitations of current filtering methods for primary complaints, these studies will help evaluate how the RSTM triages patients with more severe symptoms in these categories and if more robust filtering methods are needed.

Another significant unanswered question is whether this methodology could be broadly applied to train ML models to accurately predict outcomes for other diagnostic groups in PC. Conditions like back pain, which primarily require conservative management but overuse of tertiary healthcare services and imaging diagnostics is common, could benefit from such advancements. Questions regarding the communication between humans and web interfaces integrated with ML models must also be explored.

References

1. Rahimi SA, Légaré F, Sharma G, et al. Application of Artificial Intelligence in Community-Based Primary Health Care: Systematic Scoping Review and Critical Appraisal. *J Med Internet Res*. 2021;23(9):e29839. doi:10.2196/29839

2. Kueper JK, Terry AL, Zwarenstein M, Lizotte DJ. Artificial Intelligence and Primary Care Research: A Scoping Review. *Ann Fam Med*. 2020;18(3):250-258. doi:10.1370/afm.2518

3. McCorduck P. *Machines Who Think*. 2nd ed. AK Peters Ltd.; 2004.

4. Deep Blue | IBM. Accessed November 14, 2024. https://www.ibm.com/history/deep-blue

5. Breiman L. Random Forests. *Mach Learn*. 2001;*45*(1):5-32. doi:10.1023/A:1010933404324

6. Hochreiter S, Schmidhuber J. Long Short-Term Memory. *Neural Comput*. 1997;9(8):1735-1780. doi:10.1162/neco.1997.9.8.1735

7. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118. doi:10.1038/nature21056

8. 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. doi:10.1038/s41591-018-0268-3

9. Lee EJ, Kim YH, Kim N, Kang DW. Deep into the Brain: Artificial Intelligence in Stroke Imaging. *J Stroke*. 2017;19(3):277-285. doi:10.5853/jos.2017.02054

10. 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. doi:10.1038/s41591-019-0447-x

11. 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. doi:10.1016/j.compbiomed.2018.05.011

12. Cheng JZ, Ni D, Chou YH, et al. Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. *Sci Rep*. 2016;6:24454. doi:10.1038/srep24454

13. Papers with Code - Natural Language Processing. Accessed November 13, 2024. https://paperswithcode.com/area/natural-language-processing

14. Cramer JS. The origins of logistic regression. Published online 2002.

15. Papers with Code - ImageNet Benchmark (Image Classification). Accessed February 3, 2024. https://paperswithcode.com/sota/image-classification-on-imagenet

16. Papers with Code - GLUE Dataset. Accessed February 3, 2024. https://paperswithcode.com/dataset/glue

17. Papers with Code - Speech Recognition. Accessed February 3, 2024. https://paperswithcode.com/task/speech-recognition

18. Dalianis H. *Clinical Text Mining: Secondary Use of Electronic Patient Records*. Springer; 2018. doi:10.1007/978-3-319-78503-5

19. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. Published online December 5, 2017. Accessed May 24, 2023. http://arxiv.org/abs/1706.03762

20. GLUE Benchmark. Accessed February 3, 2024. https://gluebenchmark.com/

21. Kalliamvakou E. Research: quantifying GitHub Copilot's impact on developer productivity and happiness. The GitHub Blog. September 7, 2022. Accessed January 26, 2024. https://github.blog/2022-09-07-research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness/

22. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2

23. Sharing Google's Med-PaLM 2 medical large language model, or LLM. Google Cloud Blog. Accessed January 26, 2024. https://cloud.google.com/blog/topics/healthcare-life-sciences/sharing-google-med-palm-2-medical-large-language-model

24. Introducing ChatGPT. Accessed January 26, 2024. https://openai.com/blog/chatgpt

25. Gilson A, Safranek CW, Huang T, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med Educ*. 2023;9(1):e45312. doi:10.2196/45312

26. Li J, Dada A, Puladi B, Kleesiek J, Egger J. ChatGPT in healthcare: A taxonomy and systematic review. *Comput Methods Programs Biomed*. 2024;245:108013. doi:10.1016/j.cmpb.2024.108013

27. Ji Z, Lee N, Frieske R, et al. Survey of Hallucination in Natural Language Generation. Published online July 14, 2024. Accessed November 15, 2024. http://arxiv.org/abs/2202.03629

28. Meta M. The Llama 3 Herd of Models | Research - AI at Meta. Accessed August 27, 2024. https://ai.meta.com/research/publications/the-llama-3-herd-of-models/

29. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality | LMSYS Org. Accessed August 27, 2024. https://lmsys.org/blog/2023-03-30-vicuna

30. OpenAI O. GPT-4. Accessed August 27, 2024. https://openai.com/index/gpt-4-research/

31. Wu S, Koo M, Blum L, et al. Benchmarking Open-Source Large Language Models, GPT-4 and Claude 2 on Multiple-Choice Questions in Nephrology. *NEJM AI*. 2024;0(0):AIdbp2300092. doi:10.1056/AIdbp2300092

32. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *Npj Digit Med*. 2023;6(1):1-6. doi:10.1038/s41746-023-00873-0

33. Mattu JL Julia Angwin,Lauren Kirchner,Surya. How We Analyzed the COMPAS Recidivism Algorithm. ProPublica. Accessed October 31, 2024. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

34. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. Published online October 25, 2019. doi:10.1126/science.aax2342

35. What Is Bias in Machine Learning? [Real-World Examples]. Scalable Path. Accessed October 31, 2024. https://www.scalablepath.com/machine-learning/bias-machine-learning

36. Hamilton J. Constraint-Driven Innovation.

37. Hvað notar meðalheimili á Íslandi mörg vött af rafmagni? Vísindavefurinn. Accessed November 14, 2024. http://www.visindavefur.is/svar.php?id=55099

38. Number of ChatGPT Users (Nov 2024). Exploding Topics. March 30, 2023. Accessed November 14, 2024. https://explodingtopics.com/blog/chatgpt-users

39. Dendral. In: *Wikipedia*. ; 2023. Accessed January 28, 2024. https://en.wikipedia.org/w/index.php?title=Dendral&oldid=1157836583

40. van Melle W. MYCIN: a knowledge-based consultation program for infectious disease diagnosis. *Int J Man-Mach Stud*. 1978;10(3):313-322. doi:10.1016/S0020-7373(78)80049-2

41. CADUCEUS: An Experimental Expert System for Medical Diagnosis. In: *The AI Business: Commercial Uses of Artificial Intelligence*. MIT Press; 1986:67-80. Accessed February 3, 2024. https://ieeexplore.ieee.org/document/6284803

42. Shortliffe E, Scott A, Bischoff M, Campbell A, van Melle B, Jacobs C. *ONCOCIN: An Expert System for Oncology Protocol Management*.; 1981:881.

43. Carpenter CR, Keim SM, Upadhye S, Nguyen HB. Risk Stratification of the Potentially Septic Patient in the Emergency Department: The Mortality in the Emergency Department Sepsis (MEDS) Score. *J Emerg Med*. 2009;37(3):319-327. doi:10.1016/j.jemermed.2009.03.016

44. Chen CC, Chong CF, Liu YL, Chen KC, Wang TL. Risk stratification of severe sepsis patients in the emergency department. *Emerg Med J*. 2006;23(4):281-285. doi:10.1136/emj.2004.020933

45. Griffin SJ, Little PS, Hales CN, Kinmonth AL, Wareham NJ. Diabetes risk score: towards earlier detection of Type 2 diabetes in general practice. *Diabetes Metab Res Rev*. 2000;16(3):164-171. doi:10.1002/1520-7560(200005/06)16:3<164::AID-DMRR103>3.0.CO;2-R

46. Lindström J, Tuomilehto J. The Diabetes Risk Score: A practical tool to predict type 2 diabetes risk. *Diabetes Care*. 2003;26(3):725-731. doi:10.2337/diacare.26.3.725

47. D'Agostino RB, Vasan RS, Pencina MJ, et al. General Cardiovascular Risk Profile for Use in Primary Care. *Circulation*. 2008;117(6):743-753. doi:10.1161/CIRCULATIONAHA.107.699579

48. Beswick A, Brindle P. Risk scoring in the assessment of cardiovascular risk. *Curr Opin Lipidol*. 2006;17(4):375-386. doi:10.1097/01.mol.0000236362.56216.44

49. Mesko B, Görög M. A short guide for medical professionals in the era of artificial intelligence. *Npj Digit Med*. 2020;3. doi:10.1038/s41746-020-00333-z

50. Walton OB IV, Garoon RB, Weng CY, et al. Evaluation of Automated Teleretinal Screening Program for Diabetic Retinopathy. *JAMA Ophthalmol*. 2016;134(2):204-209. doi:10.1001/jamaophthalmol.2015.5083

51. Senders JT, Arnaout O, Karhade AV, et al. Natural and Artificial Intelligence in Neurosurgery: A Systematic Review. *Neurosurgery*. 2018;83(2):181. doi:10.1093/neuros/nyx384

52. Zheng B, Yoon SW, Lam SS. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Syst Appl*. 2014;41(4, Part 1):1476-1482. doi:10.1016/j.eswa.2013.08.044

53. 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. doi:10.1118/1.4948498

54. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat Med*. 2021;27(4):582-584. doi:10.1038/s41591-021-01312-x

55. Potnis KC, Ross JS, Aneja S, Gross CP, Richman IB. Artificial Intelligence in Breast Cancer Screening: Evaluation of FDA Device Regulation and Future Recommendations. *JAMA Intern Med*. 2022;182(12):1306-1312. doi:10.1001/jamainternmed.2022.4969

56. Chouffani El Fassi S, Abdullah A, Fang Y, et al. Not all AI health tools with regulatory authorization are clinically validated. *Nat Med*. Published online August 26, 2024:1-3. doi:10.1038/s41591-024-03203-3

57. Health C for D and R. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. *FDA*. Published online October 5, 2022. Accessed June 6, 2023. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

58. Serra-Burriel M, Locher L, Vokinger KN. Development Pipeline and Geographic Representation of Trials for Artificial Intelligence/Machine Learning—Enabled Medical Devices (2010 to 2023). *NEJM AI*. 2023;1(1):AIp2300038. doi:10.1056/AIpc2300038

59. Medscape. "I Cry but No One Cares": Physician Burnout & Depression Report 2023. Medscape. 2023. Accessed August 21, 2023. https://www.medscape.com/slideshow/2023-lifestyle-burnout-6016058

60. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med*. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838

61. Kim J, Chen ML, Rezaei SJ, et al. Perspectives on Artificial Intelligence—Generated Responses to Patient Messages. *JAMA Netw Open*. 2024;7(10):e2438535. doi:10.1001/jamanetworkopen.2024.38535

62. Schwieger A, Angst K, de Bardeci M, et al. Large language models can support generation of standardized discharge summaries - A retrospective study utilizing ChatGPT-4 and electronic health records. *Int J Med Inf*. 2024;192:105654. doi:10.1016/j.ijmedinf.2024.105654

63. Leek JT, Peng RD. What is the question? *Science*. 2015;347(6228):1314-1315. doi:10.1126/science.aaa6146

64. Risk Calculator. Accessed January 13, 2023. http://risk.hjarta.is/risk_calculator/

65. Framingham Heart Study. Accessed November 15, 2024. https://www.framinghamheartstudy.org/

66. MDCalc - Medical calculators, equations, scores, and guidelines. MDCalc. Accessed January 14, 2023. https://www.mdcalc.com/

67. Health C for D and R. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. *FDA*. Published online August 7, 2024. Accessed November 20, 2024. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

68. 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. doi:10.1126/scitranslmed.aab3719

69. Thiel SW, Rosini JM, Shannon W, Doherty JA, Micek ST, Kollef MH. Early prediction of septic shock in hospitalized patients. *J Hosp Med*. 2010;5(1):19-25. doi:10.1002/jhm.530

70. Rahimian F, Salimi-Khorshidi G, Payberah AH, et al. Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. *PLoS Med*. 2018;15(11):e1002695. doi:10.1371/journal.pmed.1002695

71. Shahidi F, Rennert-May E, D'Souza AG, Crocker A, Faris P, Leal J. Machine learning risk estimation and prediction of death in continuing care facilities using administrative data. *Sci Rep*. 2023;13(1):17708. doi:10.1038/s41598-023-43943-9

72. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*. 2018;1:18. doi:10.1038/s41746-018-0029-1

73. Levin S, Toerper M, Hamrock E, et al. Machine-Learning-Based Electronic Triage More Accurately Differentiates Patients With Respect to Clinical Outcomes Compared With the Emergency Severity Index. *Ann Emerg Med*. 2018;71(5):565-574.e2. doi:10.1016/j.annemergmed.2017.08.005

74. Reverberi C, Rigon T, Solari A, et al. Experimental evidence of effective human–AI collaboration in medical decision-making. *Sci Rep*. 2022;12:14952. doi:10.1038/s41598-022-18751-2

75. Zhai C, Wibowo S, Li LD. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learn Environ*. 2024;11(1):28. doi:10.1186/s40561-024-00316-7

76. Yan W, Li J, Mi C, et al. Does global positioning system-based navigation dependency make your sense of direction poor? A psychological assessment and eye-tracking study. *Front Psychol*. 2022;13. doi:10.3389/fpsyg.2022.983019

77. Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J Exp Psychol Gen*. 2015;144(1):114-126. doi:10.1037/xge0000033

78. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3(11):e745-e750. doi:10.1016/S2589-7500(21)00208-9

79. 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. doi:10.1038/s41591-018-0335-9

80. Hagstofan: Íbúar landsins voru 383.726 í byrjun árs 2024. Hagstofa Íslands. Accessed December 18, 2024. https://hagstofa.is/utgafur/frettasafn/mannfjoldi/mannfjoldinn-1-januar-2024/

81. Hlynsson HD, Ellertsson S, Daðason JF, Sigurdsson EL, Loftsson H. Semi-self-supervised Automated ICD Coding. Published online May 20, 2022. Accessed May 23, 2022. http://arxiv.org/abs/2205.10088

82. The Python Language Reference. Python documentation. Accessed December 8, 2024. https://docs.python.org/3/reference/index.html

83. 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. doi:10.1016/0021-9681(84)90149-8

84. 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. doi:10.7326/0003-4819-106-2-196

85. Fumal A, Schoenen J. Tension-type headache: current research and clinical management. *Lancet Neurol*. 2008;7(1):70-83. doi:10.1016/S1474-4422(07)70325-3

86. Kaniecki RG. Migraine and tension-type headache: an assessment of challenges in diagnosis. *Neurology*. 2002;58(9 Suppl 6):S15-20. doi:10.1212/wnl.58.9_suppl_6.s15

87. Moore HC, Lehmann D, de Klerk N, et al. How Accurate Are International Classification of Diseases-10 Diagnosis Codes in Detecting Influenza and Pertussis Hospitalizations in Children? *J Pediatr Infect Dis Soc*. 2014;3(3):255-260. doi:10.1093/jpids/pit036

88. Bjerrum L, Gahrn-Hansen B, Munck AP. C-reactive protein measurement in general practice may lead to lower antibiotic prescribing for sinusitis. *Br J Gen Pract*. 2004;54(506):659.

89. Harberger S, Goldin J, Graber M. Bacterial Pharyngitis. In: *StatPearls*. StatPearls Publishing; 2024. Accessed November 19, 2024. http://www.ncbi.nlm.nih.gov/books/NBK559007/

90. Shaikh N, Leonard E, Martin JM. Prevalence of streptococcal pharyngitis and streptococcal carriage in children: a meta-analysis. *Pediatrics*. 2010;126(3):e557-564. doi:10.1542/peds.2009-2648

91. Daly KA. Epidemiology of Otitis Media. *Otolaryngol Clin North Am*. 1991;24(4):775-786. doi:10.1016/S0030-6665(20)31089-6

92. Daly KA, Giebink GS. Clinical epidemiology of otitis media. *Pediatr Infect Dis J*. 2000;19(5):S31.

93. Jones GP. The pharmacological treatment of rhinosinusitis.

94. Cowley LE, Farewell DM, Maguire S, Kemp AM. Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. *Diagn Progn Res*. 2019;3:16. doi:10.1186/s41512-019-0060-y

95. Challener DW, Prokop LJ, Abu-Saleh O. The Proliferation of Reports on Clinical Scoring Systems: Issues About Uptake and Clinical Utility. *JAMA*. 2019;321(24):2405-2406. doi:10.1001/jama.2019.5284

96. McIsaac WJ, White D, Tannenbaum D, Low DE. A clinical score to reduce unnecessary antibiotic use in patients with sore throat. *CMAJ Can Med Assoc J*. 1998;158(1):75-83.

97. Patel C, Green BD, Batt JM, et al. Antibiotic prescribing for tonsillopharyngitis in a general practice setting: Can the use of Modified Centor Criteria reduce antibiotic prescribing? *Aust J Gen Pract*. 2019;48(6):395-401. doi:10.31128/AJGP-08-18-4685

98. Potpara TS, Polovina MM, Licina MM, Marinkovic JM, Prostran MS, Lip GYH. Reliable identification of "truly low" thromboembolic risk in patients initially diagnosed with "lone" atrial fibrillation: the Belgrade atrial fibrillation study. *Circ Arrhythm Electrophysiol*. 2012;5(2):319-326. doi:10.1161/CIRCEP.111.966713

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

100. Wells criteria for DVT is a reliable clinical tool to assess the risk of deep venous thrombosis in trauma patients - PMC. Accessed November 21, 2024. https://pmc.ncbi.nlm.nih.gov/articles/PMC4898382/

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

102. Fleming-Dutra KE, Hersh AL, Shapiro DJ, et al. Prevalence of Inappropriate Antibiotic Prescriptions Among US Ambulatory Care Visits, 2010-2011. *JAMA*. 2016;315(17):1864-1873. doi:10.1001/jama.2016.4151

103. Harris AM, Hicks LA, Qaseem A. Appropriate Antibiotic Use for Acute Respiratory Tract Infection in Adults: Advice for High-Value Care From the American College of Physicians and the Centers for Disease Control and Prevention. *Ann Intern Med*. 2016;164(6):425-434. doi:10.7326/M15-1840

104. Derek A. Misurski Rp, David A. Lipson MD, Arun K. Changolkar P. Inappropriate Antibiotic Prescribing in Managed Care Subjects With Influenza. 2011;17. Accessed January 21, 2023. https://www.ajmc.com/view/ajmc_11sep_misurski_601to608

105. Fu M, Gong Z, Zhu Y, et al. Inappropriate antibiotic prescribing in primary healthcare facilities in China: a nationwide survey, 2017-2019. *Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis*. Published online November 25, 2022:S1198-743X(22)00587-0. doi:10.1016/j.cmi.2022.11.015

106. Brindle P, Jonathan E, Lampe F, et al. Predictive accuracy of the Framingham coronary risk score in British men:prospective cohort study. *BMJ*. 2003;327(7426):1267. doi:10.1136/bmj.327.7426.1267

107. Tunstall-Pedoe H, Woodward M, SIGN group on risk estimation. By neglecting deprivation, cardiovascular risk scoring will exacerbate social gradients in disease. *Heart Br Card Soc*. 2006;92(3):307-310. doi:10.1136/hrt.2005.077289

108. 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. doi:10.1001/jama.282.15.1458

109. Armitage K, Woodhead M. New guidelines for the management of adult community-acquired pneumonia. *Curr Opin Infect Dis*. 2007;20(2):170-176. doi:10.1097/QCO.0b013e3280803d70

110. Karimi E. Comparing Sensitivity of Ultrasonography and Plain Chest Radiography in Detection of Pneumonia; a Diagnostic Value Study. *Arch Acad Emerg Med*. 2019;7(1):e8.

111. Makhnevich A, Sinvani L, Cohen SL, et al. The Clinical Utility of Chest Radiography for Identifying Pneumonia: Accounting for Diagnostic Uncertainty in Radiology Reports. *Am J Roentgenol*. 2019;213(6):1207-1212. doi:10.2214/AJR.19.21521

112. Moberg AB, Taléus U, Garvin P, Fransson SG, Falk M. Community-acquired pneumonia in primary care: clinical assessment and the usability of chest radiography. *Scand J Prim Health Care*. 2016;34(1):21. doi:10.3109/02813432.2015.1132889

113. Chambers D, Cantrell AJ, Johnson M, et al. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. *BMJ Open*. 2019;9(8):e027743. doi:10.1136/bmjopen-2018-027743

114. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ*. 2015;351:h3480. doi:10.1136/bmj.h3480

115. Wallace W, Chan C, Chidambaram S, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. *NPJ Digit Med*. 2022;5(1):118. doi:10.1038/s41746-022-00667-w

116. Martin SS, Quaye E, Schultz S, et al. A randomized controlled trial of online symptom searching to inform patient generated differential diagnoses. *Npj Digit Med*. 2019;2(1):1-6. doi:10.1038/s41746-019-0183-0

117. 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. doi:10.1186/s12875-020-01122-z

118. Campbell JL, Fletcher E, Britten N, et al. Telephone triage for management of same-day consultation requests in general practice (the ESTEEM trial): a cluster-randomised controlled trial and cost-consequence analysis. *The Lancet*. 2014;384(9957):1859-1868. doi:10.1016/S0140-6736(14)61058-8

119. Guidos R J. Combating Antimicrobial Resistance: Policy Recommendations to Save Lives. *Clin Infect Dis Off Publ Infect Dis Soc Am*. 2011;52(Suppl 5):S397-S428. doi:10.1093/cid/cir153

120. English surveillance programme for antimicrobial utilisation and resistance (ESPAUR) report.

121. Dolk FCK, Pouwels KB, Smith DRM, Robotham JV, Smieszek T. Antibiotics in primary care in England: which antibiotics are prescribed and for which conditions? *J Antimicrob Chemother*. 2018;73(suppl_2):ii2-ii10. doi:10.1093/jac/dkx504

122. 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. doi:10.1136/bmjopen-2014-006245

123. Pouwels KB, Dolk FCK, Smith DRM, Robotham JV, Smieszek T. Actual versus 'ideal' antibiotic prescribing for common conditions in English primary care. *J Antimicrob Chemother*. 2018;73(Suppl 2):19-26. doi:10.1093/jac/dkx502

124. Artstein R, Poesio M. Bias decreases in proportion to the number of annotators. Published online January 1, 2005.

Original publications

Paper I

Scandinavian Journal of Primary Health Care

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ipri20

Artificial intelligence in the GPs office: a retrospective study on diagnostic accuracy

Steindor Ellertsson, Hrafn Loftsson & Emil L. Sigurdsson

To cite this article: Steindor Ellertsson, Hrafn Loftsson & Emil L. Sigurdsson (2021) Artificial intelligence in the GPs office: a retrospective study on diagnostic accuracy, Scandinavian Journal of Primary Health Care, 39:4, 448-458, DOI: 10.1080/02813432.2021.1973255

To link to this article: https://doi.org/10.1080/02813432.2021.1973255

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 29 Sep 2021.

Submit your article to this journal

Article views: 1643

View related articles

View Crossmark data

Citing articles: 2 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=ipri20

SCANDINAVIAN JOURNAL OF PRIMARY HEALTH CARE
2021, VOL. 39, NO. 4, 448–458
https://doi.org/10.1080/02813432.2021.1973255

Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE

OPEN ACCESS Check for updates

Artificial intelligence in the GPs office: a retrospective study on diagnostic accuracy

Steindor Ellertsson[a], Hrafn Loftsson[b] and Emil L. Sigurdsson[a,c,d]

[a]Primary Health Care Service of the Capital Area, Reykjavik, Iceland; [b]Department of Computer Science, Reykjavik University, Reykjavik, Iceland; [c]Development Centre for Primary Health Care in Iceland, Reykjavik, Iceland; [d]Department of Family Medicine, University of Iceland, Reykjavik, Iceland

ABSTRACT

Objective: Machine learning (ML) is expected to play an increasing role within primary health care (PHC) in coming years. No peer-reviewed studies exist that evaluate the diagnostic accuracy of ML models compared to general practitioners (GPs). The aim of this study was to evaluate the diagnostic accuracy of an ML classifier on primary headache diagnoses in PHC, compare its performance to GPs, and examine the most impactful signs and symptoms when making a prediction.

Design: A retrospective study on diagnostic accuracy, using electronic health records from the database of the Primary Health Care Service of the Capital Area (PHCCA) in Iceland.

Setting: Fifteen primary health care centers of the PHCCA.

Subjects: All patients that consulted a physician, from 1 January 2006 to 30 April 2020, and received one of the selected diagnoses.

Main outcome measures: Sensitivity, Specificity, Positive Predictive Value, Matthews Correlation Coefficient, Receiver Operating Characteristic (ROC) curve, and Area under the ROC curve (AUROC) score for primary headache diagnoses, as well as Shapley Additive Explanations (SHAP) values of the ML classifier.

Results: The classifier outperformed the GPs on all metrics except specificity. The SHAP values indicate that the classifier uses the same signs and symptoms (features) as a physician would, when distinguishing between headache diagnoses.

Conclusion: In a retrospective comparison, the diagnostic accuracy of the ML classifier for primary headache diagnoses is superior to GPs. According to SHAP values, the ML classifier relies on the same signs and symptoms as a physician when making a diagnostic prediction.

KEYPOINTS

- Little is known about the diagnostic accuracy of machine learning (ML) in the context of primary health care, despite its considerable potential to aid in clinical work. This novel research sheds light on the diagnostic accuracy of ML in a clinical context, as well as the interpretation of its predictions. If the vast potential of ML is to be utilized in primary health care, its performance, safety, and inner workings need to be understood by clinicians.

ARTICLE HISTORY
Received 18 February 2021
Accepted 29 June 2021

KEYWORDS
General practice; artificial intelligence; primary headache disorders; electronic health records; statistical data interpretation; computer-assisted diagnosis; Iceland

Introduction

On a typical day, general practitioners (GPs) make multiple decisions when diagnosing and treating patients. They have limited access to immediate imaging diagnostics and tests and rely more on the patient's history and clinical examination than the second and tertiary stages of healthcare. To establish a diagnosis, a GP starts with the chief complaint, makes a hypothesis with a perceptual list of differential diagnoses, and asks the patient a series of targeted questions to include or exclude diagnoses. The GP then performs a clinical examination to confirm further or refute diagnoses while deciding if further diagnostic tests are needed. When the GP has reached a diagnostic conclusion, with a reasonable degree of certainty, he makes the diagnosis. In light of this description, a part of a GP's role requires him to act as a classifier of diseases. Physicians store their reasoning process in a free text format, often referred to

CONTACT Steindor Ellertsson steindor.ellertsson@gmail.com Primary Health Care Service of the Capital Area, Grenimelur 44, 107, Reykjavik, Iceland

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

as the clinical text note (CTN), which is a part of a patient's electronic health record (EHR). We conjecture that the CTNs can be used to develop predictive models for any step in the clinical decision-making process.

With the increased adoption of EHRs and thus, data availability, methods based on machine learning (ML) have become a focus of research in health care. ML has been shown to be a powerful tool in the medical diagnostic process with broad future possibilities in general practice [1]. ML has matched or exceeded human performance in multiple visual diagnostic tasks, such as diagnosing diabetic retinopathy from fundoscopy images [2,3], diagnosing radiology images [4,5], diagnosing skin lesions [6,7], and interpreting electrocardiograms [8,9]. However, the use and research of ML in a medical context is still in its early stages [10], with surveys showing less than 10% of healthcare providers using ML to assist in daily clinical work in 2019 in Britain [11].

The literature on peer-reviewed studies comparing the diagnostic accuracy of ML classifiers to GPs is lacking. The aim of this study was to explore the performance of an ML classifier on a common clinical problem in general practice. The null hypothesis put forward was that an ML classifier could not match or outperform the diagnostic accuracy of GPs on primary headache diagnoses. Headaches account for up to five percent [12] of all clinical problems in general practice (the majority of which are primary headaches) and are often a tough clinical problem to diagnose correctly. Primary headache diagnoses are made predominantly in primary care [12], and since the input data is extracted from CTNs, written by GPs, we chose GPs as benchmark. The four most common primary headache diagnoses were selected: tension-type headache (TTH), migraine with aura (MA+), migraine without aura (MA-), and cluster headache (CH). Interpretability is essential in medical settings [13], and the results of the examination of the intricacies of diagnostic ML models are scarcely published in the medical literature. Therefore, the impact of each clinical feature to the model's predictions was scrutinized with Shapley Additive Explanations (SHAP), which are widely used within the field of AI research to interpret ML models [14].

Materials and methods

We conducted a retrospective study and obtained 15,024 CTNs from 13,682 patient visits to one of 15 clinics of the Primary Health Care of the Capital Area (PHCCA), the largest provider of primary health care (PHC) in Iceland. As an inclusion criterion, we selected every PHC consultation where one of the four headache diagnoses was made, from 1 January 2006 to 30 April 2020. A large portion of the dataset included CTNs that contained little clinical information. Such information-scarce notes were filtered out by creating a medical keyword dictionary, containing the relevant symptoms a physician asks about when making each diagnosis. The dictionary contained words that physicians use to describe different headache symptoms in text format, for example, nausea, phonophobia, vomiting, etc. and was created by manually reading a subsample of the dataset (around 1% for each class) and selecting the right keywords. The dictionary was then applied to the whole dataset, filtering out CTNs without the chosen keywords, resembling methodology used in a similar research study [15]. Simultaneously, 93 duplicate CTNs were removed, leaving us with 2,563 information-rich CTNs, or roughly 17% of the original dataset. These CTNs had an accompanying headache diagnosis that falls into one of the four headache diagnosis categories mentioned above.

Eight hundred randomly selected CTNs were then manually annotated by a physician. The annotation method was inspired by a group of researchers who applied similar annotations on medical text [15] and its purpose is to assign binary and numerical values to clinical features, representing the existence or lack of specific signs and symptoms in the CTNs, as they are textually referenced in the CTNs. As an example of annotation, consider a typical, frequently seen sentence in a CTNs: 'The patient experienced photophobia but no phonophobia'. The questions, 'Does the patient have photophobia?' and 'Does the patient have phonophobia?' are annotated as two features, with binary values of 1 and 0. Numerical value features were assigned a value in a specific range, for example, a feature for blood pressure (e.g. with the value 134/78) or a feature for temperature (e.g. with the value 37.1 °C). This annotation process was repeated for every clinical feature referenced in the 800 CTNs. Every sign and symptom were annotated, including those unrelated to headache diagnoses to reduce the annotator's possible bias who, importantly, was also blind to the diagnoses. Where possible, a feature was only annotated as positive if the feature was considered abnormal. To give an example, a question about tactile sensation in a patient's lower extremities, was phrased as 'Does the patient have signs of reduced and/or asymmetric tactile sensation in the lower extremities?', instead of 'Does the patient have

normal and/or symmetric tactile sensation in the lower extremities?' Therefore, if a patient had abnormal tactile sensation, the binary value of 1 was assigned, and 0 if it was normal.

The annotation resulted in a dataset of 8,595 question-answer pairs (where a clinical feature is the question and the corresponding value is the answer), resulting in 254 different clinical features. This dataset was randomly split into training, validation, and test sets. The split was 75% (training), 12.5% (validation), and 12.5% (test) or 600, 100, and 100 CTNs in each, respectively. Thereafter, the ML classifier was trained on the training set, and the impact of each input feature examined using SHAP values. SHAP values are calculated by knocking out every feature of the input data, one at a time, and measuring how the predictions of the classifier change. SHAP values approximate the contribution of each feature to the model predictions, effectively revealing which features the ML classifier considers most important when making a prediction. A randomized grid search was performed to optimize the hyperparameters of the ML classifier before training, creating a four-class multi-classifier (one class for each of the four headache diagnoses) of type Random Forest. The validation set was used to further finetune the hyperparameters of the ML classifier. Once training was finished, the classifier's performance was compared to two groups of physicians: three GP specialists and three physician trainees in GP. The comparison was carried out using the test set, which had been set aside for this purpose, and to provide an unbiased evaluation of the final model fit on the training set. The physician validation process was performed via a custom web interface where the physicians reviewed the annotated features and their value for each CTN and subsequently selected one of the four headache diagnoses they found to be appropriate. Thus, the performance comparison was performed on the original diagnoses, to see how the ML classifier and the physicians compare when given the same set of information. A 10-fold cross-validation scheme was used to validate the results and to calculate 95% confidence intervals.

Statistical analysis

The metrics used for the performance assessment were the following: sensitivity, specificity, Positive Predictive Value (PPV), Matthews Correlation Coefficient (MCC), Receiver Operating Characteristics (ROC) curve, and Area under the ROC curve (AUROC score). A confusion matrix is a statistical presentation of the performance of a classifier and consists of four quadrants. Each quadrant contains a value count for one of four performance variables: true-positive count (TP), false-positive count (FP), true-negative count (TN) and false-negative count (FN). The MCC is a useful metric for imbalanced datasets since it only produces a high score if the prediction obtained good results in all of the four quadrants of the confusion matrix. It ranges from -1 to 1, where 1 represents a perfect prediction. The MCC is defined as follows:

$$\mathrm{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(FN + FP)(TN + FN)}}$$

We calculated the MCC for each diagnosis as a dichotomous outcome. We obtained an AUROC score by plotting the true-positive rate (TPR) against the false-positive rate (FPR) for each diagnosis at varying thresholds and calculating the area under the curve. The ROC curve was plotted for each diagnosis as a dichotomous outcome. All means are weighted if not stated otherwise. All data analyses were done in Python (version 3.6), in which the ML classifier was trained and validated with the scikit-learn library (version 0.22.1) [16].

Results

The flow of participants is shown in Figure 1. A total of 12,368 notes were deemed as information-scarce and filtered out. 4,913 positive features and 3,682 negative features were annotated for a total of 8,595 features, resulting in 1.33 positive features for every negative one and 10.74 features in each CTN on average. Table 1 shows the demographic distribution of the training, validation and test datasets, which all have similar demographics and have women as the majority of patients. Table 2 shows the results and comparison of the classifier and physicians on the test set. The classifier outperforms all physicians for all metrics, except for specificity. The classifier's lower specificity for TTH draws its weighted average down as it achieves equal or higher specificity compared to the physicians on the other three diagnoses. The classifier achieves the best performance in the cases of TTH, MA+, and CH. It has the lowest performance for MA- as it, in some cases, struggles to discern between MA- and TTH.

Figure 2 shows the top-20 most impactful features by SHAP values for each diagnosis. Figure 2(a) displays the features with a significant effect on a positive CH prediction. These features include CH's autonomic

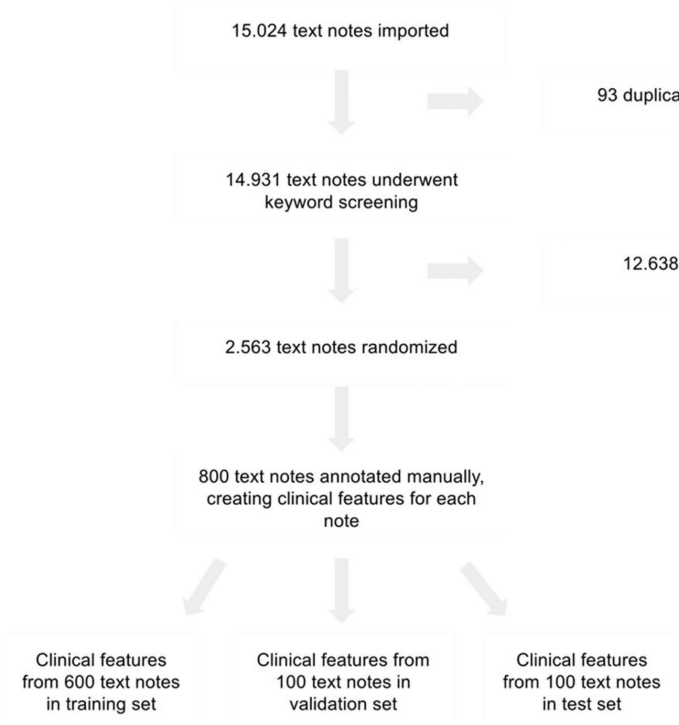


Figure 1. The inclusion and filtering process of the CTNs. The clinical features were created by annotating the CTNs, which were then split into training, validation and test sets.

Table 1. The demographic comparison of the training, validation and test set.

	Training set	Validation set	Test set
Total size	600	100	100
Female	434 (72.3%)	65 (65%)	72 (72%)
Mean age (min–max)	34.56 (6–90)	33.53 (8–78)	31.29 (8–77)

symptoms: ptosis, conjunctivitis, lacrimation, runny/stuffy nose accompanying a unilateral headache located around a single eye. Myalgia symptoms reduce CH's diagnostic likelihood, as does visual disturbance, often accompanying MA+. In Figure 2(b), we see that the two most impactful features of MA + are prodromal symptoms and visual disturbance, both the hallmarks of an aura. Unilaterality and photophobia positively impact a MA + prediction, but less since MA- also typically has these features. Aura can present itself as limb numbness, which also positively affects the prediction. In some CTNs, only the word aura was found, without a more detailed description of the presentation. In the case of MA-, displayed in Figure 2(c), we see a different SHAP profile than for MA+. The classifier uses prodromal symptoms and visual disturbance to distinguish between MA + and MA-. The SHAP profile describes a unilateral headache, with accompanying nausea, vomiting, photophobia, phonophobia, and without symptoms of myalgia. This description fits well with MA-. Figure 2(d) shows the SHAP values for TTH, where we see a pattern of absence of symptoms. Only myalgia symptoms affect the prediction positively.

Figure 3 shows the ROC curve for the test set compared to the physicians. The ROC curve plots the true-positive rate (TPR) against the false-positive rate (FPR) and the performance of a classifier increases as the curve moves up and to the left. A perfect classifier covers the whole plot, achieving an AUROC score of 1. The mean results of the specialists and trainee physicians are located lower and to the right relative to the plotted line of the ML classifier on each plot, meaning the physicians achieve a lower ratio of TPR and FPR than the classifier on all diagnoses.

Table 2. The performance metrics for the classifier and physicians on the test set.

	ML classifier			
	Sensitivity	Specificity	PPV	MCC
Cluster headache	0.83	1.00	1.00	0.91
Migraine with aura	0.92	0.99	0.92	0.90
Migraine without aura	0.67	0.99	0.86	0.73
Tension headache	0.99	0.85	0.95	0.87
Weighted average	0.95	0.88	0.94	0.86
	GP Specialist 1			
Cluster headache	0.83	1.00	1.00	0.91
Migraine with aura	0.92	0.95	0.73	0.79
Migraine without aura	0.80	0.92	0.53	0.61
Tension headache	0.89	0.96	0.98	0.79
Weighted average	0.88	0.95	0.88	0.77
	GP Specialist 2			
Cluster headache	0.67	1.00	1.00	0.81
Migraine with aura	0.58	0.96	0.70	0.59
Migraine without aura	0.90	0.87	0.45	0.58
Tension headache	0.90	0.95	0.98	0.79
Weighted average	0.86	0.94	0.85	0.73
	GP Specialist 3			
Cluster headache	0.83	1.00	1.00	0.91
Migraine with aura	0.67	0.99	0.89	0.74
Migraine without aura	0.50	0.95	0.56	0.48
Tension headache	0.97	0.72	0.91	0.75
Weighted average	0.90	0.78	0.88	0.73
	GP Trainee 1			
Cluster headache	0.83	1.00	1.00	0.91
Migraine with aura	0.92	0.99	0.92	0.90
Migraine without aura	0.70	0.93	0.54	0.56
Tension headache	0.93	0.88	0.96	0.80
Weighted average	0.89	0.91	0.90	0.78
	GP Trainee 2			
Cluster headache	0.83	0.95	0.56	0.65
Migraine with aura	0.42	0.99	0.83	0.55
Migraine without aura	0.80	0.79	0.31	0.41
Tension headache	0.81	0.95	0.98	0.64
Weighted average	0.78	0.91	0.76	0.58
	GP Trainee 3			
Cluster headache	0.83	0.98	0.71	0.75
Migraine with aura	0.50	0.99	0.86	0.62
Migraine without aura	0.70	0.90	0.44	0.49
Tension headache	0.94	0.90	0.97	0.82
Weighted average	0.87	0.91	0.86	0.75

PPV stands for Positive Predictive Value and MCC for Matthews Correlation Coefficient.

Discussion

Statement of principal findings

The ML classifier outperforms the physicians on all metrics except specificity. The MCC is arguably the most important metric, being the only one taking all four quadrants of the confusion matrix into account. The ROC curves in Figure 3 also shows a superior performance of the ML classifier compared to the physicians mean scores. In some cases, it struggles with discerning between TTH and MA- and the most probable explanation is that some patients in our dataset have an episodic TTH (ETTH), which often presents itself with mixed features of MA- and TTH. Physicians often misdiagnose these patients as having either a simple TTH or MA- [17,18], depending on the presentation of their symptoms. Additionally, the case count of MA- is relatively low. Interestingly, the classifier performs better in the case of CH, despite CH having the lowest case count. The difference is that CH has more distinguishing clinical features from the other three diagnoses, illustrating an important point: ML models are sensitive to the quality of the input data. If the training data contains errors and biases, the models learned will be erroneous and biased in the same manner. However, ML methods allow us to review the training data and correct for biases and diagnostic errors [19], which is impossible to do for physicians. As the quality of the training data increases, so will the performance of the ML models.

The SHAP profile for each diagnosis fits well with the symptoms and signs a physician would need to evaluate when suspecting these diagnoses. The SHAP profile of CH and the migraine diagnoses show features that affect the model positively, while in the case of TTH, the absence of symptoms is the main theme which fits well with the ill-defined clinical motif of TTH [17].

Strengths and weaknesses of the study

Our study benefits from a large dataset and benchmarking against GPs, which is essential when validating clinical prediction models. The selection of a few diagnoses strengthens and limits the study, as the results are easier to interpret but lack a broader diagnostic application. The small sample size of non-TTH diagnoses, and the fact that the diagnoses are obtained from a single physician when training the ML classifier, poses a limitation. In future studies, it would be optimal to have multiple physicians agreeing on the diagnoses. Another limitation is that many physicians store only a part of their reasoning process in the CTNs. To counter this, roughly 84% of the CTNs were filtered out, leaving only information-rich CTNs before splitting the dataset into training, validation and test sets. A large dataset also reduces the effect of this limitation, since it is unlikely that all physicians leave out the exact same symptoms. However, the filtering does introduce a bias toward more complex clinical cases being included. A prospective study is needed to elucidate these limitations further. Since unmentioned clinical features were assigned a binary value of 0, and as physicians are more likely to register positive symptoms than negative, a clinical feature

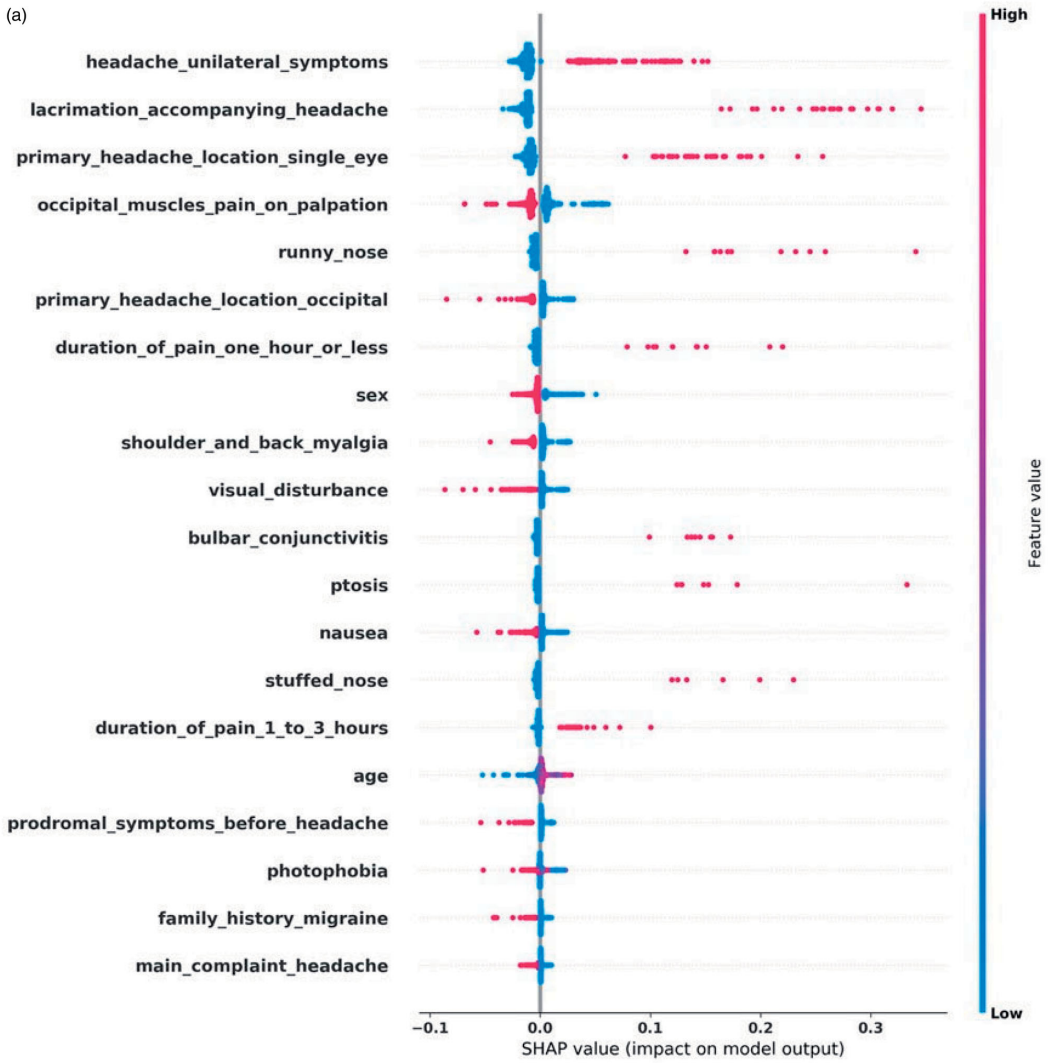


Figure 2. (a) Shapley Additive Explanations (SHAP) for the most impactful input features for a CH prediction. Each feature's name is on the left. The dots' color represents the feature's value, where blue color is lower than red. A blue dot means that a feature was negative in a single CTN, that is, it was not present. Red means the opposite. On the X-axis, the impact on the prediction is plotted, where higher score leads to increased probability of outputting a positive prediction.

was only annotated as positive if it was considered abnormal, as mentioned before. The ML classifier was also trained with missing features marked specifically but the results were the same. Again, a prospective study could explore this possible limitation further. The annotation process also limits the study with a single GP-trainee annotating the dataset as resources did not allow for additional annotators. Having GPs auditing random subsamples of annotations would have increased the quality of the annotations.

Findings in relation to other studies

We did not find any peer-reviewed papers on the evaluation of diagnostic ML classifiers in PHC settings in the literature, highlighting the importance of further

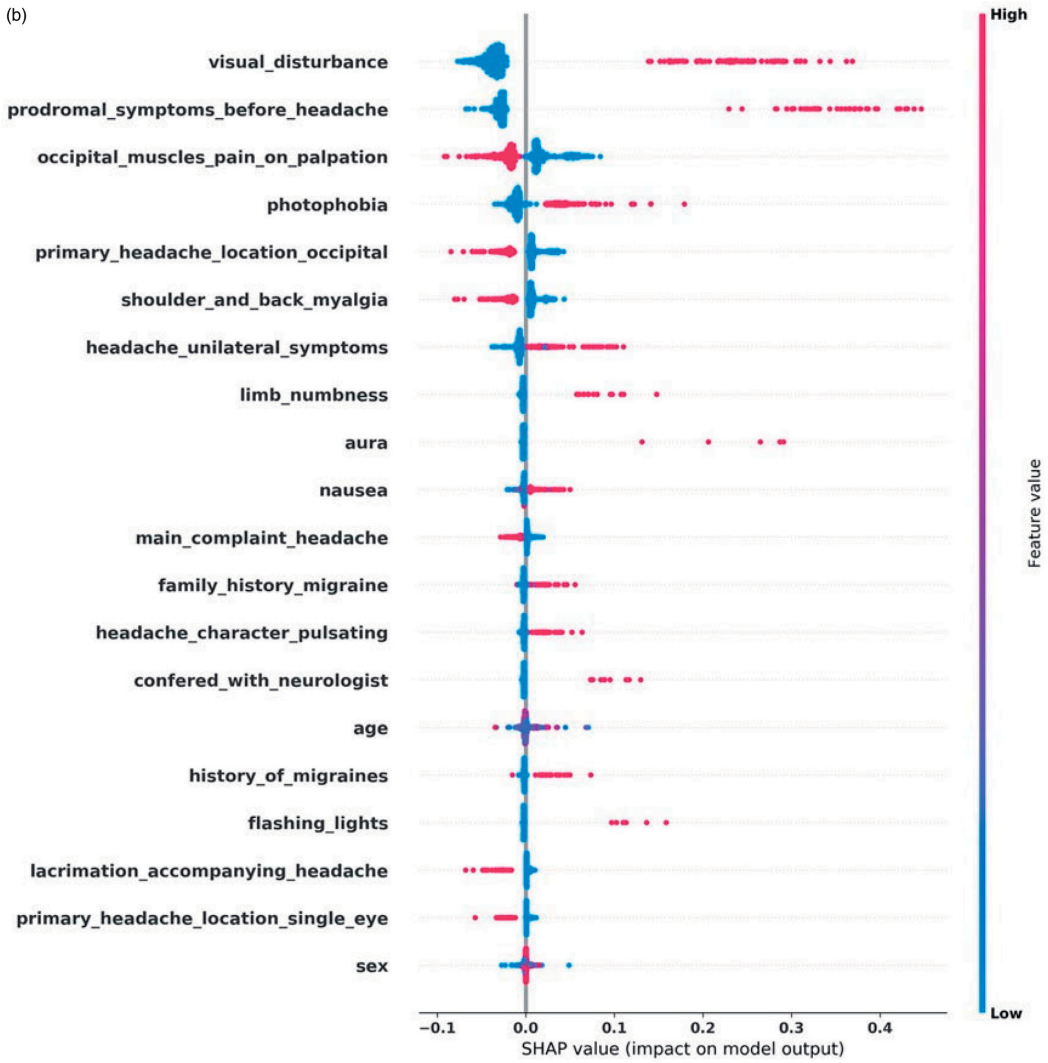


Figure 2. (b) Shapley Additive Explanations (SHAP) for the most impactful input features for an MA+prediction. Each feature's name is on the left. The dots' color represents the feature's value, where blue color is lower than red. A blue dot means that a feature was negative in a single CTN, that is, it was not present. Red means the opposite. On the X-axis, the impact on the prediction is plotted, where higher score leads to increased probability of outputting a positive prediction.

research. Studies examining online symptom checkers (OSC) are the nearest comparison, as an ML classifier, as described above, would be implemented similarly. An audit study evaluated 23 symptom checkers across 770 standardized patient evaluations [20], reporting a significant variance in the OSCs' diagnostic performance, which was not close to matching the diagnostic performance of human physicians. Babylon Health, a PHC provider in Great Britain, has developed an OSC and reported comparable diagnostic accuracy to

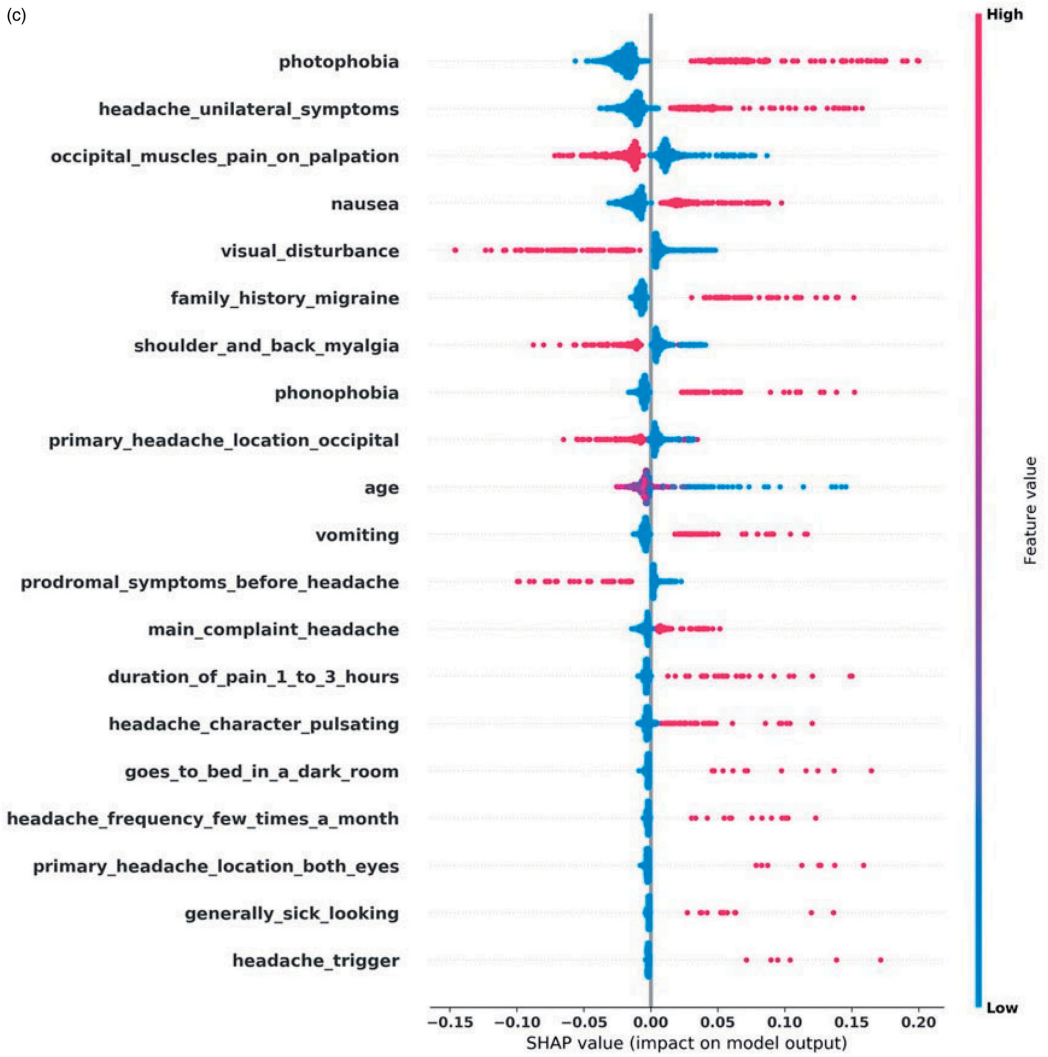


Figure 2. (c) Shapley Additive Explanations (SHAP) for the most impactful input features for an MA- prediction. Each feature's name is on the left. The dots' color represents the feature's value, where blue color is lower than red. A blue dot means that a feature was negative in a single CTN, that is, it was not present. Red means the opposite. On the X-axis, the impact on the prediction is plotted, where higher score leads to increased probability of outputting a positive prediction.

human physicians in a non-peer-reviewed paper [21]. The difference in the architecture of different OSCs complicates comparison, with many not disclosing their inner workings and research has shown that the design of computerized diagnostic system can have significant impact on their effectiveness [22].

Meaning of the study

This research is a step toward developing accurate, safe, and interpretable ML classifiers that can positively impact PHC in multiple ways. The interpretability of clinical AI models is of essence, if clinicians are to integrate AI solutions into their workflow. If the

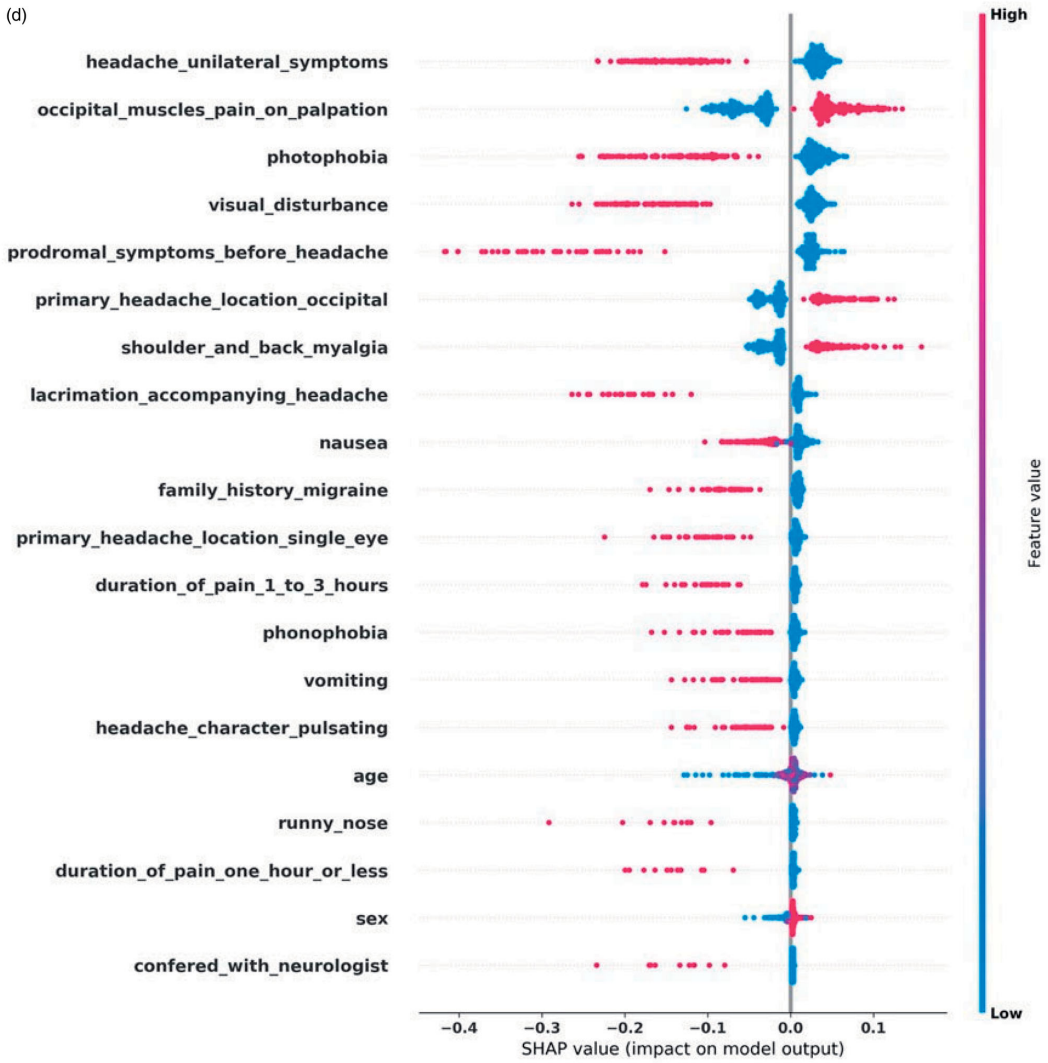


Figure 2. (d) Shapley Additive Explanations (SHAP) for the most impactful input features for a TTH prediction. Each feature's name is on the left. The dots' color represents the feature's value, where blue color is lower than red. A blue dot means that a feature was negative in a single CTN, that is, it was not present. Red means the opposite. On the X-axis, the impact on the prediction is plotted, where higher score leads to increased probability of outputting a positive prediction.

results are generalizable, any arbitrary clinical prediction model could be developed, implemented as an online service. Such a service could potentially allow for pre-screening of patients, enable self-care and lead to a more correct use of the health care system, that is, patients using expensive emergency services less, and managing self-isolating symptoms at home [23,24].

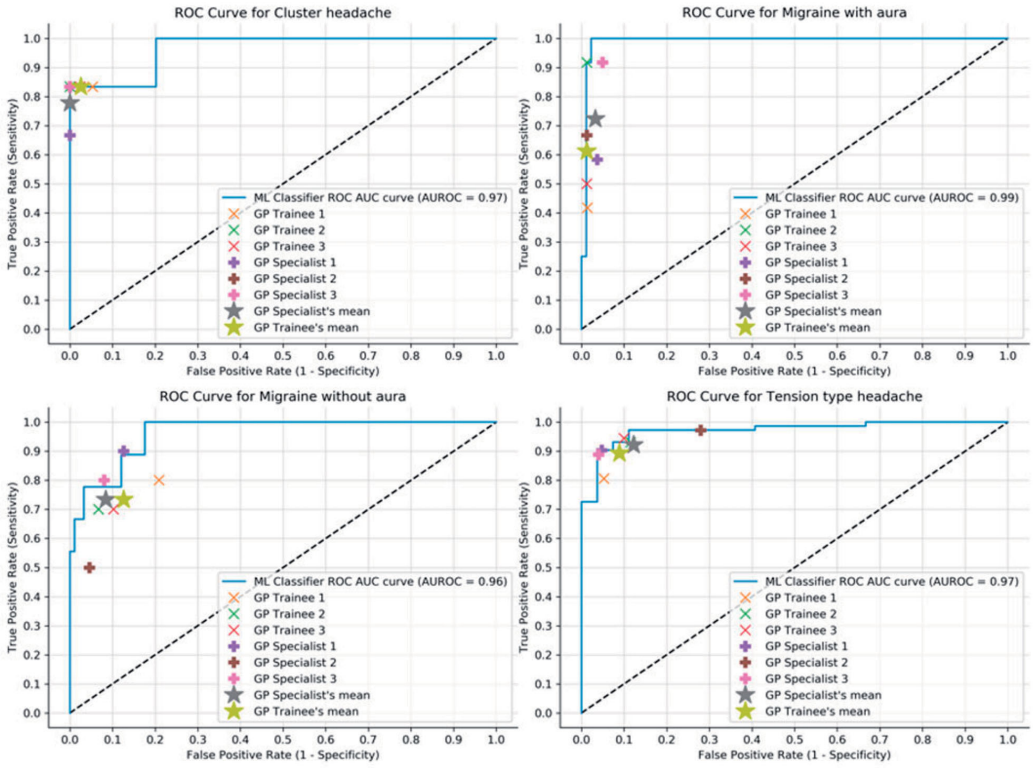


Figure 3. The ROC curve for the ML classifier, plotted for each diagnosis on the test set. The AUROC score is a measure of classifier performance with a maximum value of 1. It is calculated as the area under the ROC curve. The performance of each physician is plotted, as is the mean performance of the trainee physicians and specialists.

Acknowledgements

The authors would like to thank Dr. Hordur Bjornsson and Dr. Gudmundur Haukur Jorgenssen, PhD, for contributing to the research.

Ethical approval

Authorization for this study was given by The National Bioethics Committee in Iceland in October 2019 (reference number VSN-19-153).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was funded by the Scientific fund of the Icelandic College of General Practice.

References

[1] Buch VH, Ahmed I, Maruthappu M. Artificial intelligence in medicine: current trends and future possibilities. Br J Gen Pract. 2018;68(668):143–144.

[2] 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.

[3] 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.

[4] Cheng J-Z, Ni D, Chou Y-H, et al. Computer-Aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. Sci Rep. 2016;6:24454.

[5] 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.

[6] Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–118.

[7] Liu Y, Jain A, Eng C, et al. A deep learning system for differential diagnosis of skin diseases. Nat Med. 2020; 26(6):900–908.
[8] 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.
[9] Raghunath S, Ulloa Cerna AE, Jing L, et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. Nat Med. 2020; 26(6):886–891.
[10] Kueper JK, Terry AL, Zwarenstein M, et al. Artificial intelligence and primary care research: a scoping review. Ann Fam Med. 2020;18(3):250–258.
[11] Mistry P. Artificial intelligence in primary care. Br J Gen Pract. 2019;69(686):422–423.
[12] Frese T, Druckrey H, Sandholzer H. Headache in general practice: Frequency, management, and results of encounter. Biondi-Zoccai G, editor. Int Sch Res Notices. 2014;2014:169428.
[13] Price WN. Big data and black-box medical algorithms. Sci Transl Med. 2018;10(471) :1-5
[14] Lundberg S, Lee S-I. A unified approach to interpreting model predictions. arXiv. 2017:170507874. https://arxiv.org/abs/1705.07874
[15] 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.
[16] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–2830.
[17] Fumal A, Schoenen J. Tension-type headache: current research and clinical management. Lancet Neurol. 2008;7(1):70–83.
[18] Kaniecki RG. Migraine and tension-type headache: an assessment of challenges in diagnosis. Neurology. 2002;58(9 Suppl 6):S15–S20.
[19] Singh H, Schiff GD, Graber ML, et al. The global burden of diagnostic errors in primary care. BMJ Qual Saf. 2017;26(6):484–494.
[20] Semigran HL, Linder JA, Gidengil C, et al. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ. 2015;351:h3480.
[21] Razzaki S, Baker A, Perov Y, et al. A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis [Internet]. Available from: https://www.researchgate.net/publication/326056790_A_comparative_study_of_artificial_intelligence_and_human_doctors_for_the_purpose_of_triage_and_diagnosis.
[22] Van de Velde S, Heselmans A, Delvaux N, et al. A systematic review of trials evaluating success factors of interventions with computerised clinical decision support. Implement Sci. 2018;13(1):114.
[23] Madan A et al. Webgp: the virtual general practice – PDF free download [Internet]; 2021 [cited 2021 Jan 30]. Available from: https://docplayer.net/2513124-Webgp-the-virtual-general-practice.html.
[24] Middleton K, Butt M, Hammerla N, et al. Sorting out symptoms: design and evaluation of the 'babylon check' automated triage system. ArXiv160602041 Cs. 2016. https://arxiv.org/abs/1606.02041

Paper II

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

MORE ONLINE
www.annfammed.org

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, J11, 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

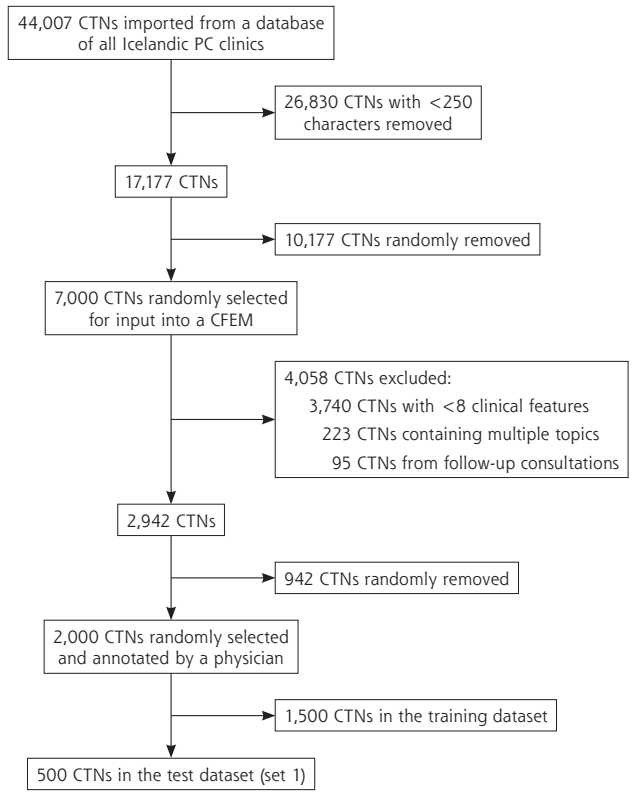
For each risk group, we examined the following outcomes: mean C-reactive protein (CRP) value, *ICD-10* code distribution, the proportion of patients re-evaluated in primary care and emergency departments within 7 days, the proportion of patients referred for a CXR, CXRs with signs of pneumonia and incidentalomas, 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 (18-93)	55 (19-91)	45 (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 (3-183)	18.8 (3-84)	28.2 (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

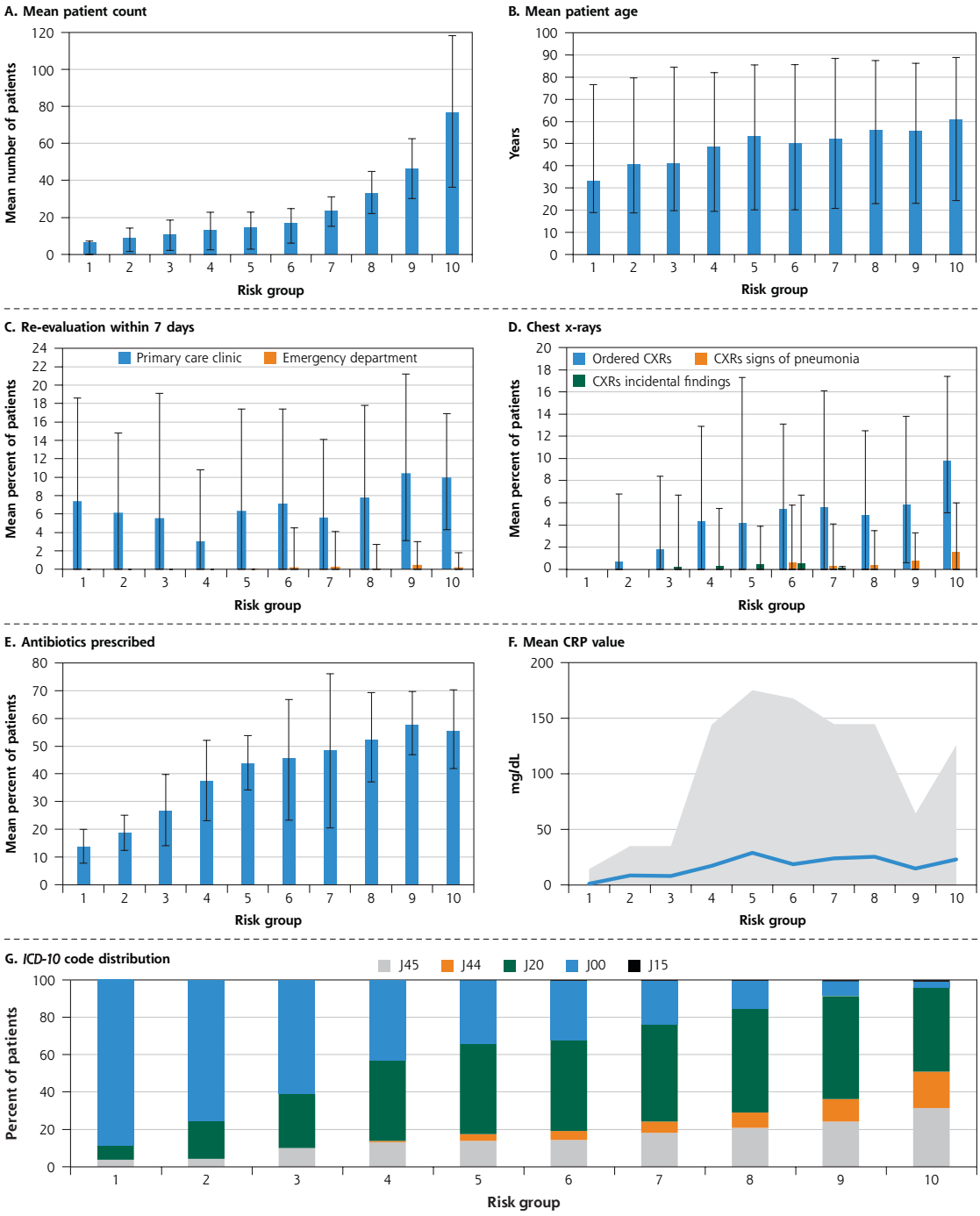
Outcomes[a]	Low-Risk Group	High-Risk Group	Sensitivity	Specificity	PPV	NPV	*P* Value
Test set 1							
Mean CRP values	7.9	23.0	...	...	...	...	<.05
Antibiotics prescribed	66 (40%)	188 (56.1%)	0.74	0.40	0.56	0.60	<.005
PC re-evaluation	9 (5.4%)	36 (10.7%)	0.80	0.34	0.11	0.95	.07
ED re-evaluation	1 (0.6%)	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.

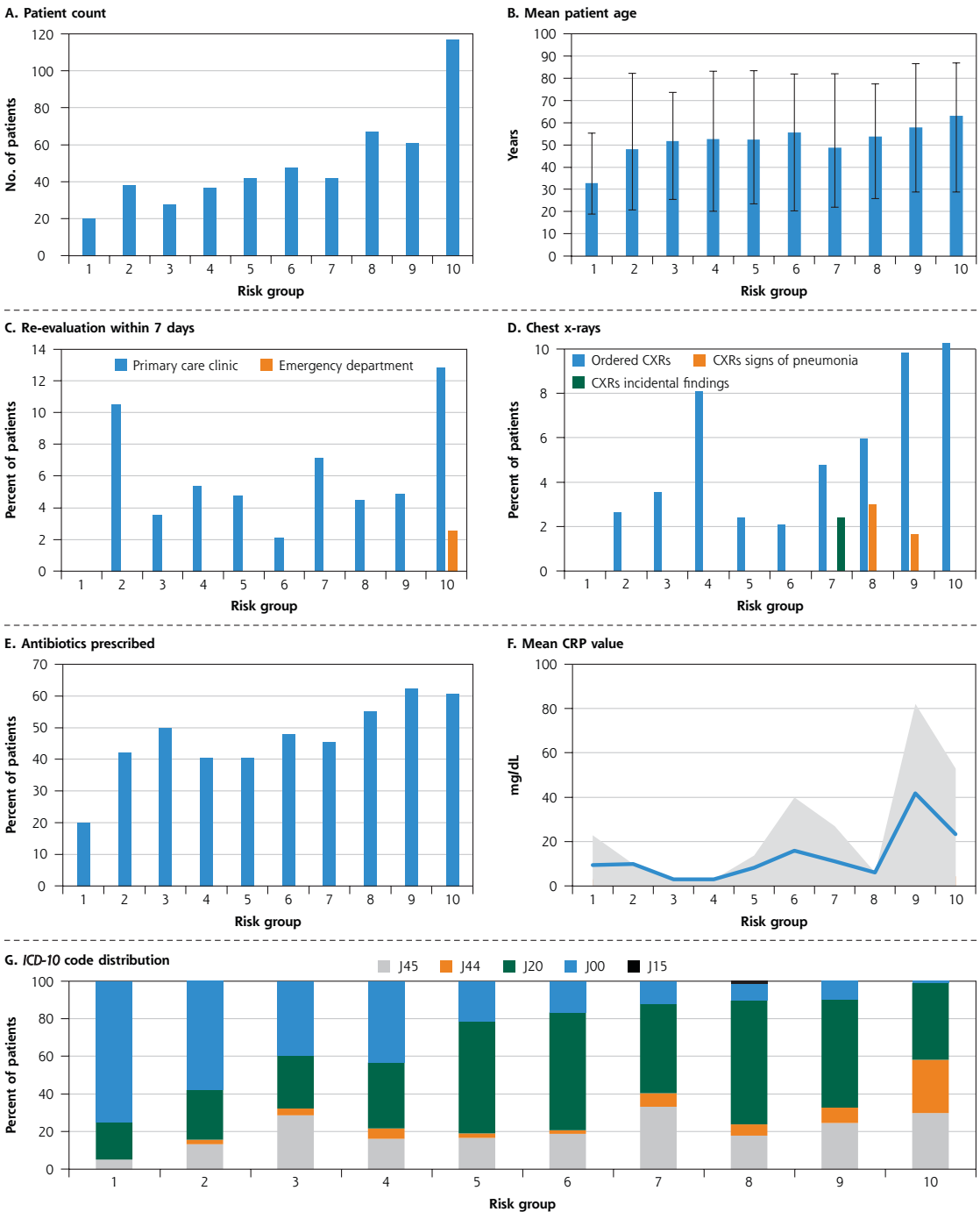
Figure 2. The outcome distribution in the cross-validated data set.



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.

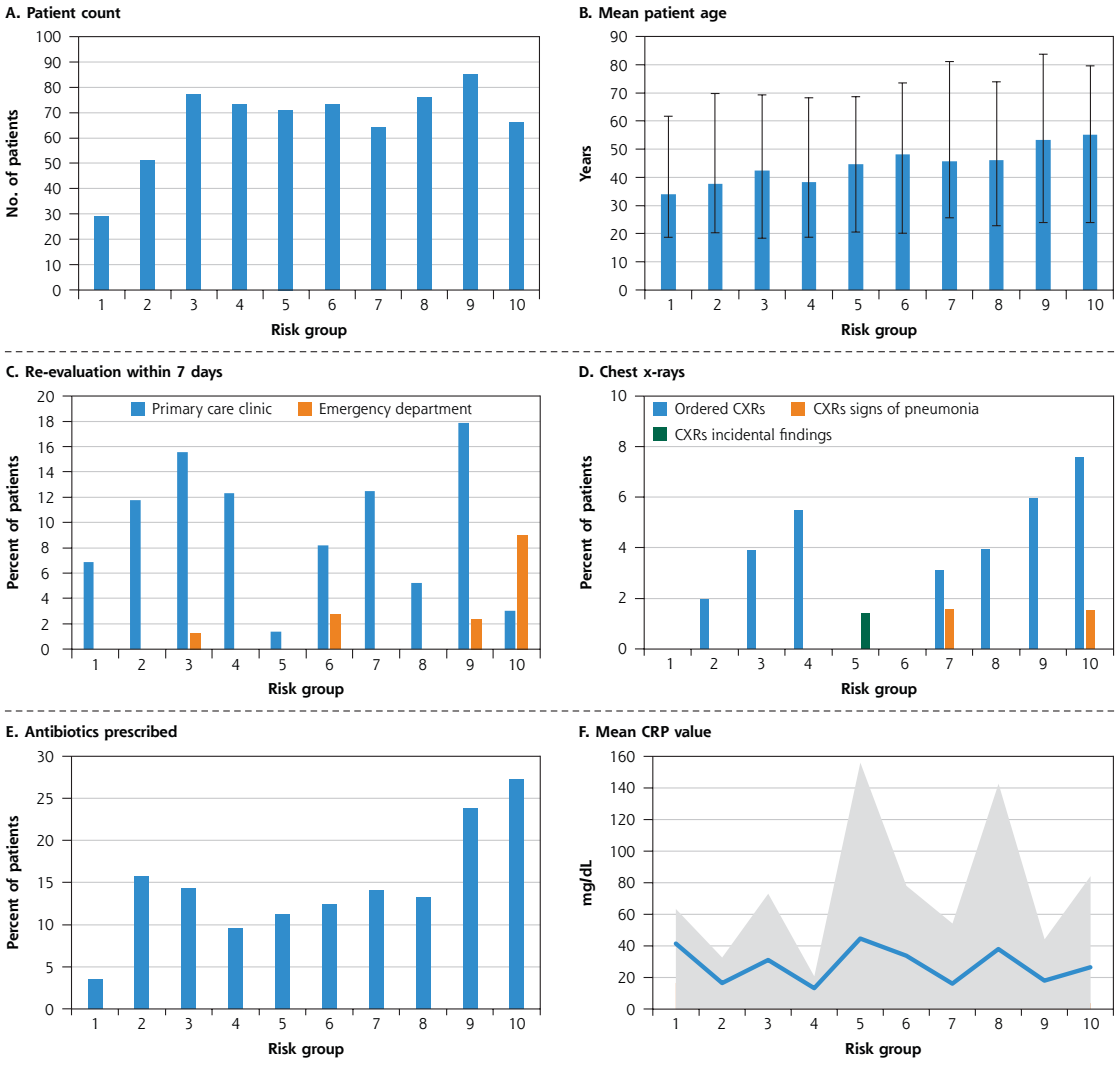
Figure 3. The outcome distribution in test set 1.



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.



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

Submitted September 9, 2022; submitted, revised, December 9, 2022; accepted January 12, 2023.

Funding support: The Scientific fund of the Icelandic College of General Practice funded this research.

Ethical approval: The National Bioethics Committee in Iceland authorized this study in November 2020 (reference number VSN-20-198).

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.queensroadpartnership.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. 10.3109/02813439008994964
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

Paper III

Can Artificial Intelligence Save Resources in Primary Care? A Prospective Study

Steindór Oddur Ellertsson, MD
Hrafn Loftsson, PhD
Emil L. Sigurdsson MD, PhD

1. Primary Health Care of the Capital Area, Iceland
2. Department of Computer Science, Reykjavik University
3. Development Center for Primary Health Care in Iceland
4. Department of Family Medicine, University of Iceland

How this fits in: Artificial Intelligence (AI) has shown promise in numerous diagnostic processes within healthcare, but its application in primary care is still underexplored. Most existing studies depend on retrospective data and compare AI performance to human decisions, which is not an ideal benchmark. This study aims to prospectively evaluate the performance and safety of an AI model in primary care by analyzing model predictions in context with real patient outcomes.

Funding support: The Scientific Fund of the Icelandic College of General Practice funded this research.

Ethical approval: The National Bioethics Committee in Iceland authorized this study in November 2021 (reference number VSN-21-204).

Acknowledgments: We extend our gratitude to the staff of Centrum Medical Clinic in Reykjavik for their assistance in recruiting participants for this study. We also thank Dr. Guðmundur Jörgensen for his valuable proofreading and insightful scientific discussions on the study's topic.

Abstract

Background: Respiratory symptoms are the most frequent primary complaint in primary care. While many of these symptoms resolve without intervention, they can also signal serious illnesses. With rising physician workloads and healthcare expenses, triaging patients before face-to-face consultations could be beneficial, potentially directing low-risk patients toward alternative communication methods. In prior work, we developed a machine learning (ML) model that demonstrated promising efficacy in identifying patients with mild symptoms, utilizing only pre-visit data. The main aim of this study was to prospectively validate the ML model by examining patient outcomes relative to the model's risk score.

Design and Setting: Individuals 18 years or older presenting with respiratory symptoms at a single primary care clinic in Iceland's capital area were recruited. The final cohort was filtered based on presenting complaints to ensure it was appropriate for the model.

Method: Participants provided information about their symptoms via a web interface before consulting with a general practitioner, allowing the model to calculate a risk score. Based on the model's risk scores, we split patients into ten risk groups, gathered a range of patient outcomes, and analyzed them within each group.

Results: Patients in groups 1-5 had no severe outcomes such as immediate emergency department referral, lung cancer diagnosis, or chest x-ray positive for pneumonia. Two patients were diagnosed with lung cancer, both in the highest-risk group.

Conclusions: The model appears to be safe in this small population. A larger prospective study is needed to confirm our findings.

Keywords: primary care; respiratory tract infections; triage; clinical decision support system; prospective studies

Introduction

Artificial intelligence (AI) has recently made significant advancements within medicine, especially in machine learning (ML). The advent of Deep Learning (DL) has fueled advances within the field of ML in the last decade, with DL models showing promise in various diagnostic tasks within multiple fields of medicine(1–6). These models are typically applied to a single diagnostic task within a controlled environment and are usually evaluated retrospectively. Researchers have increasingly shifted towards prospective studies and randomized controlled trials(7), but more high-quality studies on these models within clinical settings are still needed(8,9). Limited evidence exists of the benefits of using ML models in patient-operated triage tools in primary care(10). Patients use the internet heavily to seek healthcare information(11), and online symptom checkers are no exception(12). However, the quality of online symptom checkers and information varies greatly, underscoring the need to develop high-quality information and triage tools for healthcare(13). Symptom checkers using ML models appear to be the most promising(14).

ML algorithms are under-researched in Primary Care (PC), a field whose complexity likely deters more extensive research and application. Compared to other subfields of medicine, PC is characterized by high levels of uncertainty(15,16) and a multitude of decision-making variables. In Western countries, PC serves as the primary point of contact for most patients, positioning it as a critical area for intervention. Consequently, applying ML in PC offers a substantial opportunity to significantly enhance healthcare outcomes. This underscores the urgent need to develop robust and safe ML models tailored to the intricacies of PC.

Workload has been increasing in PC in recent decades(17,18), paralleled by an increase in the frequency of patient contacts(19). The number of General Practitioners (GPs) in the workforce is declining(19), resulting in decreased accessibility to GPs and diminished patient satisfaction(20). Burnout among healthcare workers is a growing problem(21–23), creating a

spiral where fewer workers tend to a growing amount of patients. Moreover, the overprescribing of antibiotics, particularly for treating respiratory infections(24), is a well-known problem in PC(25), and up to 50% of antibiotic prescribing is considered unnecessary(26). Antibiotic overuse has been a driving force behind the global surge in antibiotic resistance, posing substantial health risks(27–29), especially in patients with respiratory tract infections(30,31). Additionally, the overuse of imaging diagnostics represents another significant challenge within healthcare(32–34).

In previous work, we developed an ML model that showed high promise in forecasting the risk of a lower respiratory tract infection (LRTI) in patients who come to PC with respiratory symptoms(35). The model only uses symptoms as predictors, so it can calculate patient risk scores while they are still at home. We call this model a Respiratory Symptom Triage Model (RSTM). In this paper, we present the results from a study seeking to confirm these retrospective results prospectively.

Method

Participants were recruited at a single clinic at the Primary Health Care of the Capital Area (PHCCA) in Iceland in collaboration with staff at the clinic. To ensure broad coverage of patient recruitment, all patients exhibiting respiratory symptoms were invited to participate. This approach is crucial because it is impossible to ascertain in advance whether patients' symptoms are appropriate for the model, and we used presenting complaints to decide which patients were appropriate for the model, similar to what we had done when training the model.

The determination of whether patients' symptoms were appropriate for the model was made after they completed the questionnaire based on an analysis of their presenting complaints. Following the selection of presenting complaints, patients were asked to list them in order of severity by organizing them in a ranked list. We have not seen this methodology

described in the literature before, hence making it invalidated. We excluded patients with presenting complaints indicative of sinusitis, pharyngitis, and ear infection to include only patients with respiratory symptoms that indicated a common cold or a LRTI. The RSTM was not trained on patients with the excluded infections, so its predictive accuracy for outcomes in these cases remains unvalidated. To assess the performance of the RSTM in patients with infections similar to those in its training dataset, we employed the filtering method to increase the likelihood of including similar patients. The filtering methodology is explained in detail in the appendix.

Before the study commenced, a question regarding blood in sputum/hemoptysis was incorporated into the questionnaire, as this feature was not included in the model after training. Hemoptysis, being a common symptom of lung cancer, requires prompt medical evaluation in clinical practice. Therefore, patients who reported hemoptysis were assigned an additional 0.5 to their risk score, leading to a high-risk classification. We presented participants with two categories of questions: those requiring a response of "yes," "no," or "I don't know," and those necessitating the input of specific values, e.g., temperature. A full list of questions and input features of the model can be found in the appendix. The web interface was available in Icelandic and English.

The outcomes measured were mean C-reactive protein (CRP) value, International Classification of Diseases (ICD) code distribution, the proportion of patients re-evaluated in PC within seven days, the proportion of patients referred to emergency department (ED) immediately, the proportion of patients referred to a chest X-ray (CXR), CXRs with signs of pneumonia, tumors, and incidentalomas, the proportion of patients receiving antibiotic prescriptions, the proportion of patients referred for Computed Tomography (CT) of thorax, and the proportion of patients referred for CT thorax with pneumonia, cancer, and incidentalomas. Additionally, if a patient was diagnosed with an ICD code representing pneumonia and not referred for a CXR by the physician, we offered the patient a referral for a CXR if they wished to

confirm the diagnosis. However, we offered only patients a CXR after 28.03.23, as we only had sufficient permissions after that date. The study included the ICD codes J15, J18, and A24 for pneumonia, as these were the codes used by the physicians in their diagnoses.

The 95% Confidence Intervals (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-values < .05 to be significant. We implemented data analysis in Python (version 3.8) and used SciPy(36) (version 1.10.1) for statistics. To simplify the comparison, discussion, and description of results, we refer to risk groups 1-5 as the low-risk (LR) group and 6-10 as the high-risk group (HR). All tables and images represent the statistical analysis of participants from the filtered group.

Results

512 patients with 554 contacts were recruited from a single clinic at the PHCCA between October 25, 2022, and February 28, 2024. As of June 2024, the clinic had 15,393 registered patients. After filtering out patients based on presenting complaints, 197 patients with 204 contacts remained. The study flowchart is displayed in Figure 1. The population demographics, ICD-10 code distribution, and outcomes are displayed in Table 1 for the whole population and the filtered one.

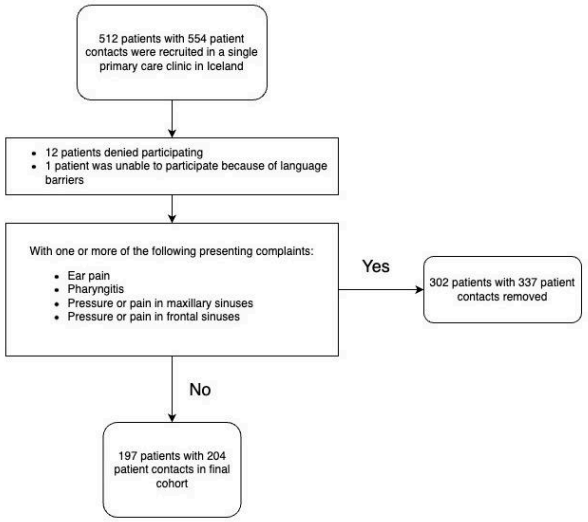


Figure 1: Study flowchart for patient selection

Demographics, ICD-10 Code Distributions, and Outcomes		
	Filtered population	**Whole population**
Patient contacts, count	204	541
Age, mean (range)	48.28 (18 - 88)	45.17 (18 - 88)
Sex, female, %	46.6	48.80
ICD-10 codes, %		
B34	25.69	20.80
R05	17.89	11.36
J45	8.72	5.07
J20	7.34	4.72
J18	5.05	2.44
J32	3.67	2.62
J22	3.67	2.45
J40	3.21	5.24
J02	2.75	6.12
R06	2.29	1.40
No ICD code	2.29	2.62
J01	1.83	4.37
J00	1.38	2.80
Patient outcomes		
CRP value, mean (95% CI), mg/dL	31.53 (0.7 - 220.56)	28.69 (1 - 201.5)
Re-evaluation in PC within 7 days, count (%)	18 (8.8)	31 (5.73)
Evaluation in ED within 7 days, count (%)	5 (2.45)	5 (0.92)
Antibiotics prescribed, count (%)	51 (25.00)	178 (32.90)
Pneumonia ICD code, count (%)	14 (6.86)	20 (3.70)
CXRs ordered, count (%)	19 (9.31)	24 (4.44)
CXRs pneumonia, count (%)	9 (4.41)	10 (1.85)
CXRs incidentaloma, count (%)	2 (0.98)	4 (0.73)
CT thorax ordered, count (%)	4 (1.96)	8 (1.48)
Lung cancer diagnoses, count (%)	1 (0.5)	2 (0.37)
Immediate ED referrals, count (%)	4 (1.98)	5 (0.92)

Table 1: Comparison of demographics and outcomes between the LR and HR groups The table compares the original dataset with the filtered dataset. The filtration process increases the number of patients with diagnoses that the model was originally trained and validated on (J00, J10, J11, J18, J20, J44, J45) while reducing the number of patients with other types of infections (J01, J02, J32). Of the more serious outcomes, one patient with lung cancer, one patient with a positive CXR, five patients with a pneumonia diagnosis and one patient with an immediate ED referral was filtered out.

Abbreviations:,
ICD-10,"International Classification of Diseases, Tenth Revision",
CRP,"C-Reactive Protein",
CXR,"Chest X-Ray",
CT,"Computed Tomography",
ED,"Emergency Department",

A comparison of the patient outcomes in the HR and LR groups is displayed in Table 2 with p-values and Odds Ratios (OR). The distribution of patients receiving pneumonia ICD codes

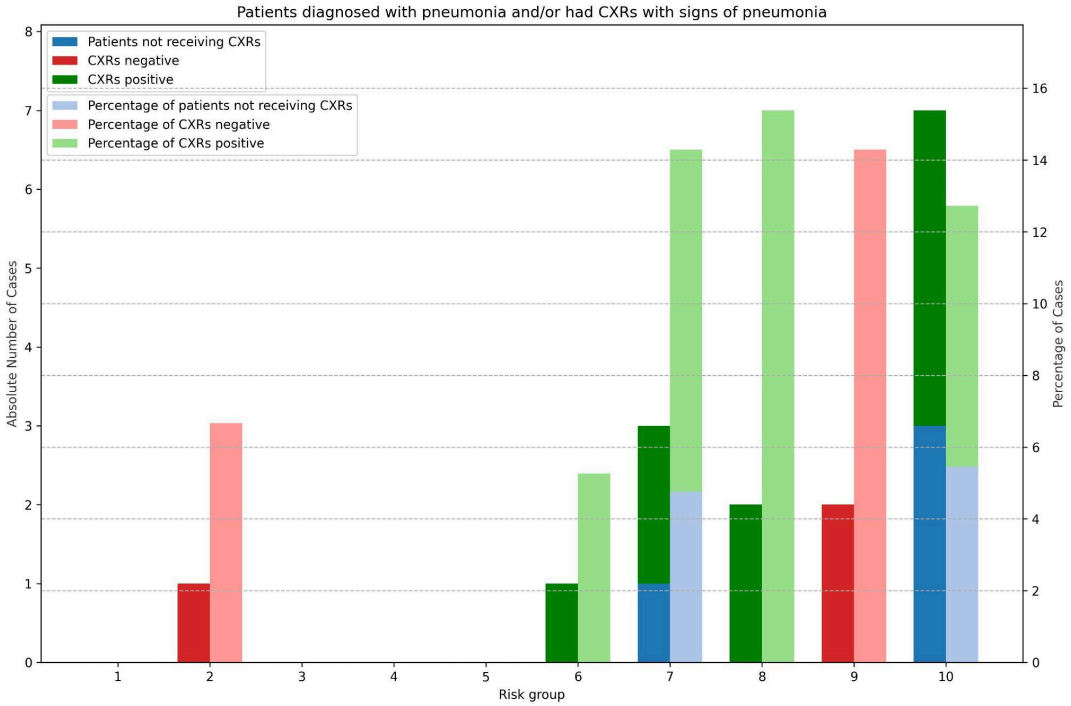


Figure 2: Distribution of pneumonia diagnoses and positive CXR results for pneumonia across risk groups. The left y-axis shows the absolute number of cases, while the right y-axis displays the percentage of cases. The x-axis corresponds to the different risk groups. In the LR group, one patient was diagnosed with pneumonia, but the subsequent CXR was negative. In the HR group, two patients received pneumonia diagnoses despite negative CXR results.

and CXRs with signs of pneumonia are shown in Figure 2, as well as CXR results if referred and their risk group placement. Three patients clinically diagnosed with pneumonia underwent a CXR as a part of the study; all three X-rays showed normal results. One of them was in the LR group. Two patients were diagnosed with lung cancer. One was removed after filtering presenting complaints, and both were in risk group number ten. Four patients were immediately referred to the ED, two in risk group 7 and two in risk group 10. Four patients were referred for a CT of thorax, one in risk group 1 and three in risk group 10. Incidentalomas were found in 2 of the 19 CXRs referred, both clinically non-significant and resulting in a CT image of the thorax as a follow-up. One patient was in risk group 1, and the other was in risk group 10.

	Low-Risk Group	**High Risk-Group**	**P-value**	**Odds Ratio**
Count	82	122	N/A	N/A
Mean CRP values (95% CIs)	19.91 (0.5 - 94.95)	34.04 (1 - 228.60)	0.55	N/A
No. CRP measurements (%)	23 (29.58)	45 (36.85)	0.23	1.50
Antibiotics prescribed (%)	17 (20.73)	34 (27.86)	0.25	1.49
PC re-evaluation (%)	6 (7.32)	12 (9.84)	0.62	1.38
ED evaluation (%)	1 (1.21)	4 (3.28)	0.65	2.75*
ED referral (%)	0 (0)	4 (3.28)	0.15	6.26*
CXR referrals (%)	5 (6.10)	14 (11.48)	0.22	1.99
CXR incidentalomas (%)	1 (1.22)	1 (0.81)	0.47	0.30***
CXR pneumonia (%)	0 (0)	9 (7.38)	**0.03**	19.00* ***
Pneumonia ICD code (%)	1 (1.45)**	13 (9.63)	**0.009**	9.66
CT referrals (%)	1 (1.21)	3 (2.45)	0.65	2.04

Table 2: Comparison of clinical outcomes between low-risk and high-risk groups. Statistically significant values are marked in bold, with significant findings in pneumonia diagnoses via ICD codes (p=0.009, OR=9.66) and CXR-confirmed pneumonia (p=0.03, OR=19.00)

*Corrected with Haldane-Anscombe correction
**Subsequent CXR was normal
***Includes only patients who were referred for a CXR by a physician
****Includes only patients who were referred for a CT by a physician

Figure 3 displays the mean CRP values in each risk group with 95% CIs. The patient diagnosed with pneumonia, who had a negative CXR in risk group 2, had a CRP of 112, causing the spike in risk group 2. One patient diagnosed with J01 (Acute Sinusitis) in risk group 1 had a CRP of 81 causing the raised values in group 1. Four patients had a CRP value above 130 and were all diagnosed with pneumonia, and three of them had positive CXRs as well. They were all in the HR group. Outcome distributions stratified by risk group are shown in Figure 4 in absolute and proportional values. The outcome distributions generally show increasing proportional values as we approach higher risk groups.

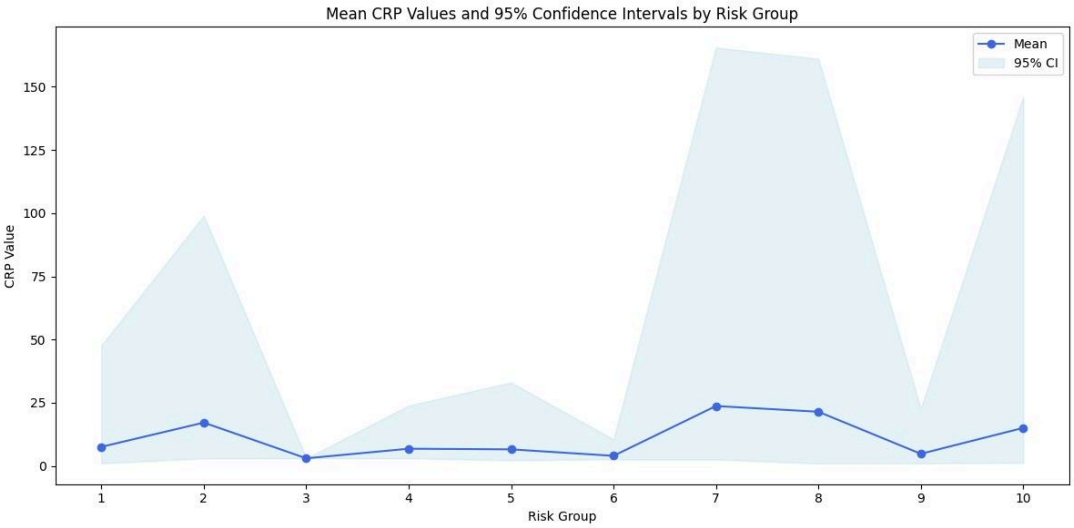


Figure 3: Mean CRP values stratified by risk group with 95% CIs.

Filtered patients

Among the filtered-out patients, six were clinically diagnosed with pneumonia, two of whom were in the LR group. One was referred to a CXR by the physician, which was normal. The second did not have a CXR taken. One pneumonia patient in the HR group had a negative CXR as part of the study, bringing the total number of pneumonia patients with negative CXRs to four, of which three had a CXR offered by the study, and one was referred by a physician. Two were in the LR group, and two in the HR group. Other pneumonia patients were in the HR groups, and two had positive CXRs. In the appendix, Figure A1 is identical to Figure 1 but includes all patients in the dataset. Two additional patients were immediately referred to the ED, both in the HR group.

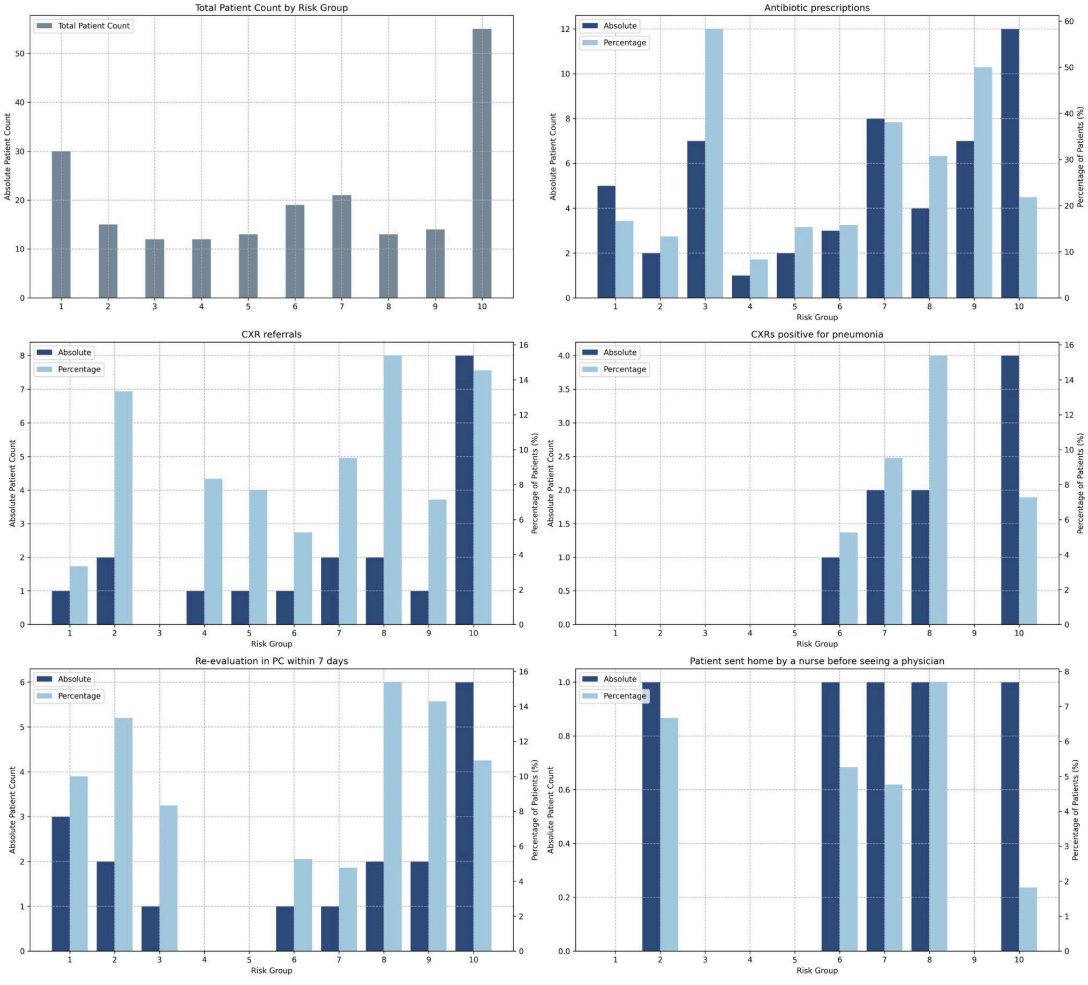


Figure 4: Outcome distributions stratified by risk group. CXR = chest x-ray; PC = primary care.

Discussion

Summary

In this prospective study, we examined the performance of the RSTM, which had shown promise in earlier work. All the outcome distributions show a similar pattern, with increased proportional values with higher risk groups, especially for the more objective outcomes such as CRP and the rates of CXRs positive for pneumonia. Importantly, there are no lung cancer diagnoses, ED referrals, or positive CXRs for pneumonia in the LR group. Interestingly, all patients who underwent a CXR as a part of the study after being clinically diagnosed with pneumonia had normal CXRs, highlighting the complexity of pneumonia diagnosis. Three patients in the LR group were diagnosed with pneumonia. Two of these patients had normal CXRs, while the third patient, unfortunately, did not undergo a CXR. Our results indicate that the model is safe in this small population, but a larger prospective study is needed to confirm our findings.

Strengths and limitations

The prospective data collection strengthens the study, mimicking how the RSTM would perform in real clinical settings. Offering patients diagnosed with pneumonia a CXR to confirm the diagnosis strengthens the results, considering the superior diagnostic capability of pneumonia by CXRs compared to physicians(37). There is selection bias in the CRP values since it is up to the consulting physician if it is measured. Patients with more severe symptoms are more likely to be recruited by clinical staff since those patients meet with a nurse and a doctor, whereas patients with milder symptoms sometimes get sent home by a nurse. A larger prospective study

is needed to explore the performance of the RSTM for rare and more broad outcomes. A selection bias is introduced by the recruitment process as the healthcare personnel have knowledge of which patients are to be recruited. A broader group of patients would be in the initial population in real settings. Filtering the patients by presenting complaints limits the study since it introduces a selection bias to the results. The filtering method has not been validated and represents a novel approach to categorizing patients suitable for an ML model. This might make it hard to apply to a new population, but a larger study is needed to explore this further.

Comparison with existing literature

Some studies have explored the effectiveness of ML models in patient triage, recognizing their potential in this area(38,39). However, most of these studies are retrospective, and the prospective studies typically involve small populations. Multiple studies have attempted to derive a diagnostic rule for pneumonia where all but one(40) include features that make them incomparable to the RSTM. There is a slight linear correlation between that diagnostic rule and the output of the RSTM; see image A2 in the appendix. Furthermore, studies examining the performance of online symptom checkers indicate that their diagnostic performance is poor compared to humans(41).

Implications for research and/or practice

If the model had been applied to the population, 26.3% of CXR referrals and 25% of CT referrals (33% if the follow-up CTs for the incidentalomas are taken into account) could have been prevented, and all unnecessary diagnostics, including one CXR which caused the patient harm with an incidentaloma needing a thorax CT to follow-up. Studies have shown that antibiotic prescriptions for patients with respiratory symptoms are overused and often inappropriate, as most of these infections have viral causes(30,42). Patients in the LR group

received 33% of antibiotic prescriptions. Antibiotic prescribing could be significantly reduced if the RSTM truly identifies patients with milder symptoms. A randomized clinical trial is needed to explore that possibility further.

Citations

1. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017 Feb;542(7639):115–8.
2. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med. 2019 Jun;25(6):954–61.
3. Cheng JZ, Ni D, Chou YH, Qin J, Tiu CM, Chang YC, et al. Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. Sci Rep. 2016 Apr 15;6:24454.
4. Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, Canazart JA, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun. 2020 Apr 9;11(1):1760.
5. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016 Dec 13;316(22):2402–10.
6. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med. 2019 Jan;25(1):65–9.
7. Serra-Burriel M, Locher L, Vokinger KN. Development Pipeline and Geographic Representation of Trials for Artificial Intelligence/Machine Learning–Enabled Medical Devices (2010 to 2023). NEJM AI. 2023 Dec 11;1(1):AIp2300038.
8. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med. 2021 Apr;27(4):582–4.
9. Potnis KC, Ross JS, Aneja S, Gross CP, Richman IB. Artificial Intelligence in Breast Cancer Screening: Evaluation of FDA Device Regulation and Future Recommendations. JAMA Intern Med. 2022 Dec 1;182(12):1306–12.
10. Gottliebsen K, Petersson G. Limited evidence of benefits of patient operated intelligent primary care triage tools: findings of a literature review. BMJ Health Amp Care Inform. 2020 May 1;27(1):e100114.
11. Morahan-Martin JM. How internet users find, evaluate, and use online health information: a cross-cultural review. Cyberpsychology Behav Impact Internet Multimed Virtual Real Behav Soc. 2004 Oct;7(5):497–510.
12. Wyatt JC. Fifty million people use computerised self triage. BMJ. 2015 Jul 8;351:h3727.
13. Hill MG, Sim M, Mills B. The quality of diagnosis and triage advice provided by free online symptom checkers and apps in Australia. Med J Aust. 2020 Jun;212(11):514–9.
14. Hammoud M, Douglas S, Darmach M, Alawneh S, Sanyal S, Kanbour Y. Evaluating the Diagnostic Performance of Symptom Checkers: Clinical Vignette Study. JMIR AI. 2024 Apr 29;3:e46875.
15. Marshall M. Redefining quality: valuing the role of the GP in managing uncertainty. Br J Gen

Pract. 2016 Feb 1;66(643):e146–8.
16. Gardner NP, Gormley GJ, Kearney GP. Learning to navigate uncertainty in primary care: a scoping literature review. BJGP Open [Internet]. 2024 Jul 1 [cited 2024 Aug 14];8(2). Available from: https://bjgpopen.org/content/8/2/BJGPO.2023.0191
17. Svedahl ER, Pape K, Toch-Marquardt M, Skarshaug LJ, Kaspersen SL, Bjørngaard JH, et al. Increasing workload in Norwegian general practice - a qualitative study. BMC Fam Pract. 2019 May 21;20(1):68.
18. Hobbs FDR, Bankhead C, Mukhtar T, Stevens S, Perera-Salazar R, Holt T, et al. Clinical workload in UK primary care: a retrospective analysis of 100 million consultations in England, 2007–14. The Lancet. 2016 Jun 4;387(10035):2323–30.
19. Institute for Government [Internet]. 2023 [cited 2024 May 12]. Performance Tracker 2023: General practice. Available from: https://www.instituteforgovernment.org.uk/publication/performance-tracker-2023/general-practice
20. Kontopantelis E, Roland M, Reeves D. Patient experience of access to primary care: identification of predictors in a national patient survey. BMC Fam Pract. 2010 Aug 28;11:61.
21. Medscape. Medscape. 2023 [cited 2023 Aug 21]. “I Cry but No One Cares”: Physician Burnout & Depression Report 2023. Available from: https://www.medscape.com/slideshow/2023-lifestyle-burnout-6016058
22. Medscape M. Medscape National Physician Burnout & Suicide Report 2020: The Generational Divide [Internet]. [cited 2020 Aug 7]. Available from: https://www.medscape.com/slideshow/2020-lifestyle-burnout-6012460
23. Orton P, Orton C, Pereira Gray D. Depersonalised doctors: a cross-sectional study of 564 doctors, 760 consultations and 1876 patient reports in UK general practice. BMJ Open. 2012 Feb 2;2(1):e000274.
24. Havers FP, Hicks LA, Chung JR, Gaglani M, Murthy K, Zimmerman RK, et al. Outpatient Antibiotic Prescribing for Acute Respiratory Infections During Influenza Seasons. JAMA Netw Open. 2018 Jun 8;1(2):e180243.
25. Office-Related Antibiotic Prescribing for Persons Aged ≤14 Years --- United States, 1993--1994 to 2007--2008 [Internet]. [cited 2024 Apr 10]. Available from: https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6034a1.htm
26. Pichichero ME. Dynamics of Antibiotic Prescribing for Children. JAMA. 2002 Jun 19;287(23):3133–5.
27. Antimicrobial resistance [Internet]. [cited 2022 Oct 21]. Available from: https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance
28. GOV.UK [Internet]. [cited 2023 Jan 18]. Health matters: antimicrobial resistance. Available from: https://www.gov.uk/government/publications/health-matters-antimicrobial-resistance/health-matters-antimicrobial-resistance
29. Murray CJL, Ikuta KS, Sharara F, Swetschinski L, Aguilar GR, Gray A, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. The Lancet. 2022 Feb 12;399(10325):629–55.
30. Ebell MH, Radke T. Antibiotic use for viral acute respiratory tract infections remains common. Am J Manag Care. 2015 Oct 1;21(10):e567-575.
31. Gulliford MC, Dregan A, Moore MV, Ashworth M, Staa T van, McCann G, et al. Continued high rates of antibiotic prescribing to adults with respiratory tract infection: survey of 568 UK general practices. BMJ Open. 2014 Oct 27;4(10):e006245.
32. Shobeirian F, Ghomi Z, Soleimani R, Mirshahi R, Sanei Taheri M. Overuse of brain CT scan for evaluating mild head trauma in adults. Emerg Radiol. 2021 Apr;28(2):251–7.
33. Armao D, Semelka RC, Elias J. Radiology’s ethical responsibility for healthcare reform: tempering the overutilization of medical imaging and trimming down a heavyweight. J Magn

Reson Imaging JMRI. 2012 Mar;35(3):512–7.
34. Smith-Bindman R, Miglioretti DL, Larson EB. Rising Use Of Diagnostic Medical Imaging In A Large Integrated Health System. Health Aff Proj Hope. 2008;27(6):1491–502.
35. Ellertsson S, Hlynsson HD, Loftsson H, Sigurðsson EL. Triaging Patients With Artificial Intelligence for Respiratory Symptoms in Primary Care to Improve Patient Outcomes: A Retrospective Diagnostic Accuracy Study. Ann Fam Med. 2023 May 1;21(3):240–8.
36. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020 Mar;17(3):261–72.
37. Speets AM, Hoes AW, Graaf Y van der, Kalmijn S, Sachs APE, Mali WPTM. Chest radiography and pneumonia in primary care: diagnostic yield and consequences for patient management. Eur Respir J. 2006 Nov 1;28(5):933–8.
38. Miles J, Turner J, Jacques R, Williams J, Mason S. Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: a systematic review. Diagn Progn Res. 2020 Oct 2;4(1):16.
39. ScienceDaily [Internet]. [cited 2024 Jun 3]. AI outperforms clinicians' judgment in triaging postoperative patients for intensive care. Available from: https://www.sciencedaily.com/releases/2019/10/191029182456.htm
40. 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 Jan;37(3):215–25.
41. Chambers D, Cantrell AJ, Johnson M, Preston L, Baxter SK, Booth A, et al. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. BMJ Open. 2019 Aug;9(8):e027743.
42. Creer DD, Dilworth JP, Gillespie SH, Johnston AR, Johnston SL, Ling C, et al. Aetiological role of viral and bacterial infections in acute adult lower respiratory tract infection (LRTI) in primary care. Thorax. 2006 Jan 1;61(1):75–9.

Appendix

Filtering patients by presenting complaints

As the first step in the study, all participants had to select a presenting complaint before proceeding to the questionnaire on the web interface. They could either click on one of the main complaints listed in the interface or search for an appropriate option in a dropdown menu. Participants were instructed to inform the study presenter if a suitable complaint was unavailable. However, this situation did not occur, as all participants were able to find a fitting main complaint. After selecting their presenting complaints, participants were asked to rank

them by severity or importance. They then answered the questions necessary for the RSTM to calculate their risk score. We removed patients who reported any of the following complaints:

- ear pain
- sore throat
- pain or pressure over the maxillary sinuses
- pain or pressure over the frontal sinuses

All other patients were included in the final study population.

Questionnaire procedure and a full list of questions

Patients were instructed to answer the following questions based on their current state unless explicitly directed otherwise (not considering the course of their illness). For the following questions, they could respond with "yes," "no," or "I don't know":

- Do you have symptoms of the common cold?
- Have you been diagnosed with asthma by a physician?
- Is there mucus / expectorate coming up from your lungs?
- Have you been coughing up blood, or has there been blood in the mucus from your lungs?
- Have you experienced a fever in the course of your symptoms?
- Have you experienced shortness of breath?
- Have you been diagnosed by a physician with chronic obstructive pulmonary disease (COPD)?
- Do you have any allergies?
- Have you vomited?

- Do you have a fever?

If the question “Do you have a fever?” was answered with a yes, the users were asked to input their temperature value in degrees Celsius. The input allowed for floating point numbers between 35 and 42. If they had not measured the temperature, they could choose “I have not measured my temperature”. In such cases, a default value of 37°C was used as input to the model.

The following items were presented as a list of checkboxes associated with the question, “Do you have any of the following symptoms?”:

- Cough
- Sore throat
- Runny Nose (Rhinorrhea)
- Nasal congestion / Stuffy nose

Additionally, the users answered the question “Do you have a smoking history?” by selecting one of the four choices:

- active smoker
- used to smoke but quit
- social smoker
- no

Age and sex are also part of the predictors of the RSTM and were collected automatically when each participant signed the consent electronically.

Figure A1

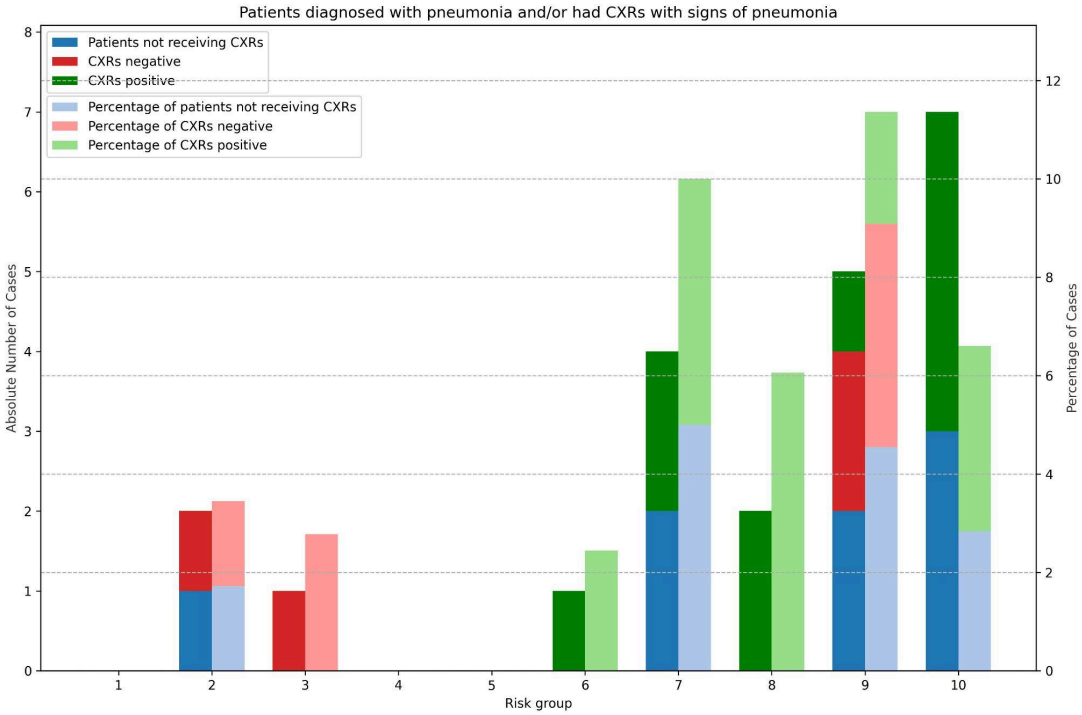
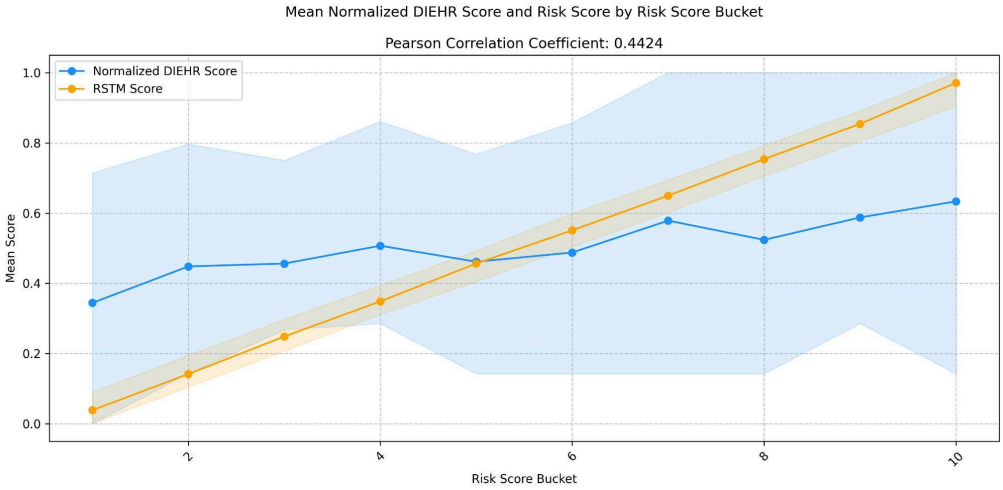


Figure A2


Mean Normalized DIEHR Score and Risk Score by Risk Score Bucket
Pearson Correlation Coefficient: 0.4424
Normalized DIEHR Score
RSTM Score
Mean Score
Risk Score Bucket
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10