Title
Development and Validation of a Deep Learning Model for Detecting Signs of Tuberculosis on Chest

Authors
Scott Lee¹, Shannon Fox¹, Raheem Smith¹, Kimberly A. Skrobarcek¹, Harold Keyserling², Christina R.

¹National Center for Emerging and Zoonotic Infectious Diseases, US Centers for Disease Control and - F2
1 C T r

Disclaimer

1
1
/

Radiographs among US-bound Immigrants and Refugees

Authors

Scott Lee^t, Shannon Fox¹, Raheem Smith¹, Kimberly A. Skrobarcek¹, Harold Keyserling², Christina R.

Phares⁵, Deborah Lee^t, Drew L. Posey^t

¹Nat **Authors**
Scott Lee¹, Shannon Fox¹, Raheem Smith¹, Kimberly A. Sk
Phares¹, Deborah Lee¹, Drew L. Posey¹
¹National Center for Emerging and Zoonotic Infectious Di
Prevention
²G2S Corporation
Disclaimer
The *A* < F 1 F2 【 T r *A* I ∈ ^ c c Scott Lee
Phares¹, D
¹National
Preventiol
²G2S Corp
Disclaime
The findin
represent
Abstract (Immigrant
exam, ove
≥15 years
often imp
does not c
We obtair , Shannon Fox

leborah Lee¹, D

Center for Eme

n

oration
 r

gs and conclus

the official pos
 **300 words manus

300 words manus

sand refugees

rseen by the U

old receive a cl

lement quality

urrently have a

ded** , Kaneem Smith
Irew L. Posey¹
Irging and Zoono
ions in this manu
ition of the Cent
x): (297)
Seeking admissi
S Centers for Dis
hest x-ray to look
control (QC) pro₈
method for con
est radiographs of , Kimberly A. Skrobarcek
tic Infectious Diseases, US
script are those of the au
ers for Disease Control ar
on to the United States m
ease Control and Prevent
of tor signs of tuberculosis.
grams to ensure radiograp
ducting , Harold Keyserling
S Centers for Diseas
thors and do not ne
ind Prevention (CDC)
ust first undergo ar
ion (CDC), during w
Although individua
phs are interpreted
vs at scale. , Emmanus

e Control and

e Control and

net all persol

screening si

correctly, the

medical exar Phares
¹Nation
Prevent
²G2S Co
Disclain
The find
represe
Abstrac
exam, o
215 yea
often in
does no
We obt:
We obt: , Deborah Lee
al Center for Elion
ion
rrporation
ner
dings and concl
nt the official _K
t (300 words n
ants and refuge
verseen by the
rs old receive a
plement quali
t currently hav
ained digitized
adiographs from , Drew L. Posey

merging and Zoc

merging and Zoc

sosition of the Constitution

Discreed Seeking adm

Discreed and directions of the series of the control of the contro Prevention

Sisclaimer

The findings and conclusions in this manuscript are those of the authors and do not necessarily

epresent the official position of the Centers for Disease Control and Prevention (CDC).

Abstract (30 Prevention

Prevention

Pisclaimer

The finding

represent the

Abstract (31

Immigrants

Exam, overs

Prevention

Dising radio_i

three tasks:

There tasks:

There tasks:

There tasks:

There tasks:

There tasks:

There Abstract (300 words max): (297)
Immigrants and refugees seeking admission to the United States must first undergo an overseas medical

Disclaimer
Disclaimer
The findings and c
epresent the office
Mastract (300 word
minigrants and reasures by:
15 years old recent
offer implement does not currently
We obtained digit
Jsing radiographs
hree tasks: identi
uber represent the official position of the Centers for Disease Control and Prevention (CDC).
 Abstract (300 words max): (297)

Immigrants and refugees seeking admission to the United States must first undergo an overseas

ex represent the official position of the United States must first undergo an
axam, overseen by the US Centers for Disease Control and Prevention (CDC), during wh
215 years old receive a chest x-ray to look for signs of tuber ノー・ミ く へ し せ ハ ドー・ノ せ く exam, overseen by the US Centers for Disease Control and Prevention (CDC), during which all persons
215 years old receive a chest x-ray to look for signs of tuberculosis. Although individual screening sites
often implement 215 years old receive a chest x-ray to look for signs of tuberculosis. Although individual screening sites
often implement quality control (QC) programs to ensure radiographs are interpreted correctly, the CI
does not curr often implement quality control (QC) programs to ensure radiographs are interpreted correctly, the CD
does not currently have a method for conducting similar QC reviews at scale.
We obtained digitized chest radiographs co

does not currently have a method for conducting similar QC reviews at scale.
We obtained digitized chest radiographs collected as part of the overseas immigration medical exam.
Using radiographs from applicants 15 years ol We obtained digitized chest radiographs collected as part of the overseas im
Using radiographs from applicants 15 years old and older, we trained deep lea
three tasks: identifying abnormal radiographs; identifying abnormal いしせい ドー・ノ せららく くらく Using radiographs from applicants 15 years old and older, we trained deep learning models to perform
three tasks: identifying abnormal radiographs; identifying abnormal radiographs suggestive of
tuberculosis; and identifyi three tasks: identifying abnormal radiographs; identifying abnormal radiographs suggestive of
tuberculosis; and identifying the specific findings (e.g., cavities or infiltrates) in abnormal radiographs.
We then evaluated t tuberculosis; and identifying the specific findings (e.g., cavities or infiltrates) in abnormal radio,
We then evaluated the models on both internal and external testing datasets, focusing on two
performance metrics: indiv We then evaluated the models on both internal and external testing datasets, focusing on two classes
performance metrics: individual-level metrics, like sensitivity and specificity, and sample-level metrics,
like accuracy performance metrics: individual-level metrics, like sensitivity and specificity, and sample-level metrics,
like accuracy in predicting the prevalence of abnormal radiographs.
A total of 152,012 images (one image per applic ike accuracy in predicting the prevalence of abnormal radiographs.

A total of 152,012 images (one image per applicant; mean applicant age 39 years) were used for model

training. On our internal test dataset, our models p A total of 152,012 images (one image per applicant; mean applicant
training. On our internal test dataset, our models performed well bo
suggestive of TB (area under the curve [AUC] of 0.97; 95% confidence
testimating sampl ノ t v v v v v v v v / l l l (training. On our internal test dataset, our models performed well both in identifying abnormalities suggestive of TB (area under the curve [AUC] of 0.97; 95% confidence interval [CI]: 0.95, 0.98) and in estimating sample-l suggestive of TB (area under the curve (AUC) of 0.97; 95% confidence interval [CI]: 0.95, 0.98) and internal internal internal test datasets, our models performed similarly well in identifying both generic abormalities abo suggestive of the same of Table and the same (-2% absolute percentage error; 95% CIC: -8%, 6%). On the external test datasets, our models performed similarly well in identifying both generic abnormalities (AUCs ranging fro external test datasets, our models performed similarly well in identifying both generic abnormalities (AUCs ranging from 0.89 to 0.92) and those suggestive of TB (AUCs from 0.94 to 0.99). This performance was consistent ac

Strong performance relative to high-quality radiological reference standards across a variety of datasets

Abbreviations

(AUCs ranging from 0.89 to 0.92) and those suggestive of TB (AUCs from 0.94 to 0.99). This performa
was consistent across metrics, including those based on thresholded class predictions, like sensitivity
specificity, and F was consistent across metrics, including those based on thresholded class predictions, like sensitivity, specificity, and F1 score.
Strong performance relative to high-quality radiological reference standards across a vari specificity, and F1 score.
Strong performance relative to high-quality radiological reference standards across a variety of datases
uggests our models may make reliable tools for supporting chest radiography QC activities Strong performance relat
Strong performance relat
suggests our models may
Abbreviations
Al, artificial intelligence;
Disease Control and Prev
Migration Health; HaMLE
Organization for Migratic
NOTE: This preprint reports no Suggests our models may make reliable tools for supporting chest radiography QC activities at CDC.
Abbreviations
Al, artificial intelligence; AP, anterior-posterior; AUC, area under the ROC curve; CDC, Centers for
Disease Supported and Preventions
Al, artificial intelligence; AP, anterior-posterior; AUC, area under the ROC curve; CDC, Centers for
Disease Control and Prevention; M; CNN, convolutional neural network; DGMH, Division of Global
 Abbreviations
Al, artificial inte
Disease Contro
Migration Heal
Organization fo
NOTE: This prep Disease Control and Prevention; M; CNN, convolutional neural network; DGMH, Division of Globa
Migration Health; HaMLET, Harnessing Machine Learning to Eliminate Tuberculosis; IOM, Internat
Organization for Migration; IRHB, Migration Health; HaMLET, Harnessing Machine Learning to Eliminate Tuberculosis; IOM, Internation
Organization for Migration; IRHB, Immigrant and Refugee Health Branch; MiMOSA, Migrant
NOTE: This preprint reports new resea Organization for Migration; IRHB, Immigrant and Refugee Health Branch; MiMOSA, Migrant
NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

Introduction

management Operational System Application, the pyriodical character recognition; PII, personally-
identifiable information; PPV, positive predictive value; QC, quality control; ROC, receiver operating
characteristic; TPP, characteristic; TPP, target product profile; WHO, World Health Organization
Introduction
Tuberculosis is an infectious disease caused by *Mycobacterium tuberculosis* (MTB) that typically affe
the lungs.¹¹Those who are in Introductions, TET, target product prome, whic), wond nearly organization
Tuberculosis is an infectious disease caused by *Mycobacterium tuberculosis* (the lungs.¹¹Those who are infected but do not show symptoms have lat Tuberculosis is an infectious disease caused by Mycobacterium tuberculosis (MTB) that typically affects Tuberculosis an infectious disease caused by mycobacterium text-culosis inferfection (LTBI)
and may never develop tuberculosis disease. LTBI is not infectious but still needs to be treated to
prevent the progression into t

Division of Global Migration Health (DGMH) within the Centers for Disease Control and Prevention (CDC)
has regulatory responsibility to oversee the medical examinations of these applicants. The examinations and may never develop tuberculosis disease. LTBI is not infectious but still needs to be treated to
prevent the progression into tuberculosis disease. Tuberculosis disease causes coughing, chest pain,
prevent the progressi prevent the progression into tuberculosis disease. Tuberculosis disease causes coughing, chest paratigue, weight loss, fever, and many other symptoms, and is contagious.²² It is the 13th leading care death in the world ratigue, weight loss, fever, and many other symptoms, and is contagious.²² It is the 13th leading cause
death in the world, and the second leading infectious killer after COVID-19.¹ In the United States,
tuberculosis death in the world, and the second leading infectious killer after COVID-19.¹ in the United States,
tuberculosis rates have been declining, and the tuberculosis incidence rate for 2021 was 2.4 cases per
100,000 persons, death in the world, and the second leading infectious killer after COVID-19.
tuberculosis rates have been declining, and the tuberculosis incidence rate f
100,000 persons, with the majority of reported cases occurring amon or 2021 was 2.4 cases
US-born persons (71.4
In the United States, The
Control and Prevention
pplicants. The examing
pplicants. The examing
ons for panel physicia
e US Department of S
Organization for Migration
igration of 10,000 persons, with the majority of reported cases occurring among non-US-born persons (71.4%).
Non-US born persons had an incidence rate 15.8 times higher (12.5 cases per 100,000) when compare
to US-born persons had an i Non-US born persons had an incidence rate 15.8 times higher (12.5 cases per 100,000) when compare
to US-born persons had an incidence rate 15.8 times higher (12.5 cases per 100,000) when compare
to US-born persons (0.8 cas to US-born persons (0.8 cases per 100,000).³
Every year, approximately 550,000 immigrants and refugees apply to enter the United States. The
Division of Global Migration Health (DGMH) within the Centers for Disease Contr to US-born persons (0.8 cases per 100,000).
Every year, approximately 550,000 immigran
Division of Global Migration Health (DGMH)
has regulatory responsibility to oversee the r
are conducted overseas in accordance with (
p 「ヽヽ (´ ヒ r r ̄ s l s ̄ ; ´ 〃 EIH a FF(villtratum Distribution of Global Migration Health (DGMH) within the Centers for Disease Control and Prevention
Division of Global Migration Health (DGMH) within the Centers for Disease Control and Prevention
has regulatory responsib In September 2018, DGMH began receiving digital copies of chest x-ray images from panel sites. This from the required verseas in accordance with CDC DGMH's Technical Instructions for panel physicians. All
panel physicians are conducted overseas in accordance with CDC DGMH's Technical Instructions for panel physicians. All
panel physicians are licensed local medical doctors on an agreement with the US Department of State to
perform these exa

panel physicians are licensed local medical doctors on an agreement with the US Department of State to
perform these examinations, and many are affiliated with the International Organization for Migration
(IOM), an intergo

perform these examinations, and many are affiliated with the International Organization for Migration (IOM), an intergovernmental agency under the United Nations system that supports migrants. IOM works closely with US Dep (IOM), an intergovernmental agency under the United Nations system that supports migrants. IOM
works closely with US Department of State and CDC to ensure the healthy migration of US-bound
immigrants and refugees.
DGMH's T works closely with US Department of State and CDC to ensure the healthy migration of US-bound
immigrants and refugees.
DGMH's Technical Instructions for tuberculosis seek to prevent disease importation by detecting an
trea immigrants and refugees.
DGMH's Technical Instructions for tuberculosis seek to prevent disease importation by detecting
treating infectious tuberculosis before arrival, and to reduce tuberculosis-related morbidity and
mor DGMH's Technical Instruct
treating infectious tubercu
mortality in these populat
applicants 15 years and ol-
tuberculosis, signs or symp
specimens collected for sn
In September 2018, DGMH
was due to the rollout of t
from t Itrats InfaIk Aa treating infectious tuberculosis before arrival, and to reduce tuberculosis-related morbidity and
mortality in these populations. Requirements include a medical history and physical examination. All
applicants 15 years and mortality in these populations. Requirements include a medical history and physical examination
applicants 15 years and older receive chest x-rays, and anyone with a chest x-ray suggestive of
tuberculosis, signs or symptom applicants 15 years and older receive chest x-rays, and anyone with a chest x-ray suggestive of
tuberculosis, signs or symptoms suggestive of tuberculosis, or known HIV, then has three sputum
specimens collected for smears tuberculosis, signs or symptoms suggestive of tuberculosis, or known HIV, then has three sputus specimens collected for smears and cultures ^{4,5}
In September 2018, DGMH began receiving digital copies of chest x-ray images specimens collected for smears and cultures ⁴⁵
In September 2018, DGMH began receiving digital copies of chest x-ray images from panel sites. T
was due to the rollout of the eMedical system, an electronic health processi In September 2018, DGMH began receiving digity
was due to the rollout of the eMedical system,
from the required overseas immigrant examina
applicants were collected, raising the possibility
DGMH's already effective oversig I ヽ f a I k k a c N t

was due to the rollout of the eMedical system, an electronic health processing system that collects day
from the required overseas immigrant examinations. In 2018 alone, 124,551 images for 521,270
applicants were collected from the required overseas immigrant examinations. In 2018 alone, 124,551 images for 521,270
applicants were collected, raising the possibility of using machine learning methods to complement
DGMH's already effective overs applicants were collected, raising the possibility of using machine learning methods to compleme
DGMH's already effective oversight for the radiologic components of tuberculosis screening for U
bound immigrants and refugee DGMH's already effective oversight for the radiologic components of tuberculosis screening for US-
bound immigrants and refugees.⁶
Brief overview of AI for Chest radiography
Artificial intelligence (AI), especially as bound immigrants and retugees.
Brief overview of AI for Chest radi
Artificial intelligence (AI), especia
applications in medical imaging. E
cancer histopathology⁹⁻¹¹, and im-
Models are typically designed for
the presenc c III,
a (II しんしょう しょうしょう しゅうしゃ Brief overview by Alfor Chest radiography
Artificial intelligence (AI), especially as ena
applications in medical imaging. Examples
cancer histopathology⁹⁻¹¹, and imaging for
Models are typically designed for diagnost
th applications in medical imaging. Examples include diabetic retinopathy⁷, cardiovascular risk prediction⁸, cancer histopathology⁹⁻¹¹, and imaging for musculoskeletal^{12,13}, cardiac¹⁴, and pulmonary¹⁵conditions.
 applications in medical imaging. Examples include diabetic retinopathy
cancer histopathology⁹⁻¹¹, and imaging for musculoskeletal^{12,13}, cardiac¹⁴
Models are typically designed for diagnostic tasks, like segmenting an , cardiovascular risk prediction⁻
⁴, and pulmonary¹⁵conditions.
omical structures or indicating
ic tasks, like predicting survival \mathcal{L} Models are typically designed for diagnostic tasks, like segmenting anatomical structures or indicating
the presence of disease, but they have also been designed for prognostic tasks, like predicting surviva
time for pati Models are typically designed for an expectation of the presence of disease, but they have also been designed for prognostic tasks, like predicting survival
time for patients from histopathology whole-slide images.¹⁶ time for patients from histopathology whole-slide images.¹⁶ time for patients from histopathology whole-slide images.¹⁶

| s ヽ ┟ t r c c r c r c s r

specific diseases, like pneumonia^{-17,18}, COVID-19⁻⁹, lung cancer^{26/21}, and tuberculosis.^{15,17,22} Recent
work^{23,24} has broadened the scope to include abnormalities in general. Studies focusing on tubercu
work^{23,} $\frac{1}{2}$ work^{23,24} has broadened the scope to include abnormalities in general. Studies focusing on tuberchave ranged from the narrow evaluation of specific models (typically commercial) on relatively sm
test sets^{25,2} have ranged from the narrow evaluation of specific models (typically commercial) on relatively small
test sets^{35,26} to the development of original algorithms from custom largescale training sets.^{23,27,28} The
references the net states and the development of original algorithms from custom largescale training sets.²³²⁷²⁸ The
references standards for these studies are often mixed, comprising radiological findings, clinical
diagnoses, micr references standards for these studies are often mixed, comprising radiological findings, clinical
diagnoses, microbiological testing, and nucleic acid amplification testing (NAAT).
Of special note, when laboratory tests a diagnoses, microbiological testing, and nucleic acid amplification testing (NAAT).
Of special note, when laboratory tests are used as reference standards, model performance tend
of dop relative to performance against a rad diagnoses are used as reference standards, model per drop relative to performance against a radiological standard; however, a small numet the World Health Organization's (WHO) Target Product Profile (TPP) for tube 90% sens

 drop relative to performance against a radiological standard; however, a small number of models have the World Health Organization's (WHO) Target Product Profile (TPP) for tuberculosis triage tests 90% sensitivity and at l met the World Health Organization's (WHO) Target Product Profile (TPP) for tuberculosis triage tests at 90% sensitivity and at least 70% specificity^{26,29} relative to NAAT or culture, even when testing does not rely on in 90% sensitivity and at least 70% specificity^{26,29} relative to NAAT or culture, even when testing does not
90% sensitivity and at least 70% specificity^{26,29} relative to NAAT or culture, even when testing does not
rely rely on initial radiographic interpretation to identify images with abnormalities (see Khan 2020 and Qin
2021, where all study participants received both a chest x-ray and either a GeneXpert MTB/RIF test or a
50% sputum cu 2021, where all study participants received both a chest x-ray and either a GeneXpert MTB/RIF test or a
sputum culture upon enrollment).^{26,30}
Project goal
The primary use-cases of models in the literature have mostly bee 2021, sputtum culture upon enrollment),^{26,30}

2021, The primary use-cases of models in the literature have mostly been clinical decision support and

2021, The primary use-cases of models in the literature have mostly be Project goal
The primary use-cases of models in the
workflow improvement, with special e
measured by AUC), interpretability, an
the potential benefit for models to bol
rank-ordering radiographs in batches b
project, we eva | ノート | イート | ドード ミーム(ノーニ ミード |ハーハ||オード||ドミミ((コミート Project goal
The primary
measured by
the potentia
rank-orderin
project, we e
performance
a given batcl
screening pro
(an unplanne
(QC) with the
To achieve o
abnormal (Ti
identifying tl
identifying tl
identify the construct workflow improvement, with special emphasis on individual-level classification performance (oft
measured by AUC), interpretability, and usability. With respect to TB, emphasis has also been pla
the potential benefit for mo measured by AUC), interpretability, and usability. With respect to TB, emphasis has also been placed of
the potential benefit for models to bolster TB screening and diagnosis in low-resource settings, e.g., by
ank-ordering

the potential benefit for models to bolster TB screening and diagnosis in low-resource settings, e.g., by
praised correcting radiographs in batches by their probability of disease to guide manual review. For this
project, mank-ordering radiographs in batches by their probability of disease to guide manual review. For this
project, we evaluated our models' ability to meet these goals, and we also sought to evaluate their
performance in estim project, we evaluated our models' ability to meet these goals, and we also sought to evaluate their
performance in estimating sample-level prevalence, i.e., in predicting the number of abnormal x-rays
a given batch. These proformance in estimating sample-level prevalence, i.e., in predicting the number of abnormal x-ray agiven batch. These measures mirror two important operational use-cases of the model in the overscreening program: support is a given batch. These measures mirror two important operational use-cases of the model in the overseas screening program: supporting panel physicians in providing high-quality initial reads during the exams (an unplanned screening program: supporting panel physicians in providing high-quality initial reads during the exams
(an unplanned but potentially impactful application), and enabling DGMH to conduct quality control
(QC) with the radio (an unplanned but potentially impactful application), and enabling DGMH to conduct quality control
(QC) with the radiographs once they have been collected (the primary focus of our current project).
To achieve our goals, w (QC) with the radiographs once they have been collected (the primary focus of our current project).
To achieve our goals, we trained and validated models for performing three tasks: classifying images
abnormal (Task 1), cl To achieve our goals, we trained and validated models for performing three tasks: classifying images
abnormal (Task 1), classifying images as abnormal and suggestive of tuberculosis (Task 2), and
identifying the specific a 丁 w j j j z c r z j l l l l l l l l l l l l l abnormal (Task 1), classifying images as abnormal and suggestive of tuberculosis (Task 2), and
identifying the specific abnormalities in the images (Task 3) (we use the same numbering scheme to
identify the corresponding m identifying the specific abnormalities in the images (Task 3) (we use the same numbering scheendientifying the specific abnormalities in the images (Task 3) (we use the same numbering scheendientify the corresponding model identify the corresponding models). To meet the two use-cases above, we tested our models on a
variety of data sets, both internal and external, and we measured their performance using two
operating points, one chosen to o variety of data sets, both internal and external, and we measured their performance using two operating points, one chosen to optimize individual-level classification performance, and one chosed pointize accuracy in predic operating points, one chosen to optimize individual-level classification performance, and one cloptimize accuracy in predicting prevalence. Although we did not formally test abnormality local methods, e.g., via object dete optimize accuracy in predicting prevalence. Although we did not formally test abnormality localization methods, e.g., via object detection models, we implemented a number of common saliency methods for visualizing suspecte

Methods
Internal dataset curation and description

methods, e.g., via object detection models, we implemented a number of common saliency methods for sinalizing suspected abnormalities on the input images to improve model interpretability and pilot interactive methods for independent of the methods for manual review.

Simulating suspected abnormalities on the input images to improve model interpretability and pilot

interactive methods for manual review.
 Methods

Internal dataset curatio interactive methods for manual review.

Wethods

Internal dataset curation and description

For our internal datasets (hereafter HaMLET, from our project title, Harnessing Machine Learning to

Eliminate Tuberculosis), we o Methods
Internal dataset curation and description
For our internal datasets (hereafter HaN
Eliminate Tuberculosis), we obtained an
from four sources: eMedical, the US Dep
application for recording and transmittir **『 /** F E
F E f a For our internal datasets (hereafter HaMLET, from our project title, Harnessing Machine Learning to Eliminate Tuberculosis), we obtained an initial convenience sample of 327,650 digitized radiographs
from four sources: eMedical, the US Department of State's immigrant health data system, a web-base
application for recordi From four sources: eMedical, the US Department of State's immigrant health data system, a web-based
proplication for recording and transmitting immigrant medical cases between the Panel Physicians, U:
application for recor application for recording and transmitting immigrant medical cases between the Panel Physicians, US

application for recording and transmitting immigrant medical cases between the Panel Physicians, US

dependence of the St application for recording and transmitting immigrant medical cases between the Panel Physicians, US-

(MiMOSA), the International Organization for Migration's (IOM) refugee health data system; IOI
Global Teleradiology and Quality Control Centre (GTQCC); and a small number of individual US
Global Teleradiology and Quality C

Global Teleradiology and Quality Control Centre (GTQCC); and a small number of individual US
immigrant panel sites that screen a relatively high number of applicants with tuberculosis each year
immigrant panel sites that immigrant panel sites that screen a relatively high number of applicants with tuberculosis each
names are provided in the Acknowledgments). Importantly, all these sites have experienced ra
and most conduct either double or names are provided in the Acknowledgments). Importantly, all these sites have experienced radiologists
and most conduct either double or triple readings on all chest x-ray images as a measure of quality
control. Regardless and most conduct either double or triple readings on all chest x-ray images as a measure of quality control. Regardless of source, all radiographs were stored as Digital Imaging and Communications in Medicine (DICOM) files control. Regardless of source, all radiographs were stored as Digital Imaging and Communications in
Medicine (DICOM) files, and all radiographic findings were extracted directly from the structured er
in the DS-3030 Tuberc Medicine (DICOM) files, and all radiographic findings were extracted directly from the structured ent
in the DS-3030 Tuberculosis Worksheet³² instead of from free-text radiology reports by way of natural
language process in the DS-3030 Tuberculosis Worksheet³² instead of from free-text radiology reports by way of natural
language processing (NLP).
The set assembled for this project was taken from screenings conducted during a ten-year pe language processing (NLP).

The set assembled for this project was taken from screenings conducted during a ten-year period from

October 2011 to October 2021 and not exclusively from the digitized radiographs routinely re The set assembled for this p
October 2011 to October 2(
DGMH since 2018 (Supplem
digitized radiographs began
from private immigrant pan
radiographs from applicant.
whose pixel arrays were milextraction (n=107,115) (Figithe 「く【くげいらせミ けくくくく October 2011 to October 2021 and not exclusively from the digitized radiographs routinely received by
DGMH since 2018 (Supplemental Table 2 shows the distribution of exams by region and year). The
digitized radiographs beg DGMH since 2018 (Supplemental Table 2 shows the distribution of exams by region and year). The
digitized radiographs began in 2018 due to the eMedical rollout, but we also received screenings direct
from private immigrant

digitized radiographs began in 2018 due to the eMedical rollout, but we also received screenings difform private immigrant panel sites and from IOM that predated the eMedical rollout. We excluded radiographs from applicant from private immigrant panel sites and from IOM that predated the eMedical rollout. We excluded
radiographs from applicants less than 15 years of age (n=52,523), as well as those stored in DICOM files
whose pixel arrays we inality and the methant sites and in Syemmathy and the syem and the resultion and the syem and the promatilities and the promatilities the set of the entrarction (n=107,115) (Figure 1 shows a flow diagram providing a detai whose pixel arrays were missing, corrupt, or otherwise unreadable by the software we used for
extraction (n=107,115) (Figure 1 shows a flow diagram providing a detailed numerical accounting of
extraction (n=107,115) (Figur extraction (n=107,115) (Figure 1 shows a flow diagram providing a detailed numerical accountin
these two exclusion steps). The remaining 168,012 radiographs constituted our final dataset, wh
split into training, validation these two exclusion steps). The remaining 168,012 radiographs constituted our final dataset, which split into training, validation, and testing portions following the procedure described below.

Radiologist annotations

Ch split into training, validation, and testing portions following the procedure described below.

Radiologist annotations

Chest radiograph abnormalities for the immigration exam fall into one of two groups: those suggestive Radiologist annotations

Radiologist annotations

Chest radiograph abnormalities for the immigration exam fall into one of two groups: those

och tuberculosis and those not. Abnormalities suggestive of tuberculosis include りくくくく r w c 【 \ (_ w c c c r w c _ r

Radiologist annotations
Chest radiograph abnorr
of tuberculosis and those
consolidations; reticular
defined margins; pleural
opacities; discrete nodul
reaction. Abnormalities i
abnormalities; smooth p
calcified lymph nodes of the tuberculosis and those not. Abnormalities suggestive of tuberculosis include: inflitrates or
consolidations; reticular markings suggestive of fibrosis; cavitary lesions; nodules or mass with poorly
defined margins; consolidations; reticular markings suggestive of fibrosis; cavitary lesions; nodules or mass wit
defined margins; pleural effusions; hilar or mediastinal adenopathy; miliary findings; discrete
opacities; discrete nodule(s) defined margins; pleural effusions; hilar or mediastinal adenopathy; miliary findings; discrete linear
opacities; discrete nodule(s) without calcification; volume loss or retraction; and irregular thick pleura
reaction. Ab opacities; discrete nodule(s) without calcification; volume loss or retraction; and irregular thick pleureaction. Abnormalities not suggestive of tuberculosis include cardiac, musculoskeletal, or other
abnormalities; smoot reaction. Abnormalities not suggestive of tuberculosis include cardiac, musculoskeletal, or other
abnormalities; smooth pleural thickening; diaphragmatic tenting; calcified pulmonary nodules; and
aclcified lymph nodes.
Mos is a standard.

Most abnormalities; smooth pleural thickening; diaphragmatic tenting; calcified pulmonary nodules; and

calcified lymph nodes.

Most abnormal images in our internal validation and test sets were suggestive calcified lymph nodes.

Most abnormal images in our internal validation and test sets were suggestive of tuberculosis. Althowe did benchmark our generic model against two open datasets with a wider range of abnormalities
 Most abnormal images
we did benchmark our
(described below), we
Importantly, however,
applicants with active t
had previously been sc
tuberculosis sputum te
microbiological or bact
standard.
External test sets
To supplemen 『ハ(=ミトサミ ノイトハ(==) we did benchmark our generic model against two open datasets with a wider range of abnormalities (described below), we focused primarily on the tuberculosis classification task for our analysis.
(mportantly, however, becau (described below), we focused primarily on the tuberculosis classification task for our analysis.
Importantly, however, because only a small number of the abnormal images (1,551) were from applicants with active tuberculo ²

Importantly, however, because only a small number of the abnormal images (1,551) were from

applicants with active tuberculosis at the time of screening—the vast majority were from applic

had previously been screened applicants with active tuberculosis at the time of screening—the vast majority were from applic
had previously been screened, diagnosed with tuberculosis, and received treatment, or whose
tuberculosis sputum testing result

¹ Had previously been screened, diagnosed with tuberculosis, and received treatment, or whose
tuberculosis sputum testing results were negative—we chose not to benchmark our models against a
microbiological or bacteriolo tuberculosis sputum testing results were negative—we chose not to benchmark our models age
microbiological or bacteriologic reference standard, focusing instead on a purely radiological re
standard.
External test sets
To s microbiological or bacteriologic reference standard, focusing instead on a purely radiological reference standard.

External test sets

To supplement our internal testing data, we benchmarked our binary models 1 and 2 on f External test sets
To supplement our internal testing data, we benchmarked our binary models 1 and 2 on four external
datasets: ChestX-ray8³³; the Montgomery County, USA (MCU) and the Shenzhen, China (SHN)
tuberculosis d External to
To supple
datasets: (
tuberculo:
which wer
(ChestX-ra 1 c t ヽ (External test sets
To supplement ou
datasets: ChestX-
tuberculosis datas
which were sugge
(ChestX-ray8). For datasets: ChestX-ray8³³; the Montgomery County, USA (MCU) and the Shenzhen, China (SHN)
tuberculosis datasets³⁴; and VinDr-CXR.³⁵ VinDr-CXR was the largest with 2,971 images in total, 161 of
which were suggestive of tuberculosis datasets³⁴; and VinDr-CXR.³⁵ VinDr-CXR was the largest with 2,971 images in tota
which were suggestive of tuberculosis; the others ranged in size from 138 images (MCU) to 81
(ChestX-ray8). For all datasets which were suggestive of tuberculosis; the others ranged in size from 138 images (MCU) to 810 images
(ChestX-ray8). For all datasets, we use the testing splits specified in their original publications, and the
let use of t (ChestX-ray8). For all datasets, we use the testing splits specified in their original publications, and the

interval of the original publications, and the

interval of the specified in their original publications, and th (ChestX-ray). For all datasets, we use the testing specified in the testing specified in the testing specified in the
References

Google^{23,27}.
MCU, SHN, and VindDr-CXR had labels indicating the suggested presence of tuberculosis (reference
standards varied by dataset and included radiographic, clinical, and laboratory evidence of disease), bunly Ch MCU, SHN, a
standards va
standards va
only ChestX
this imperfe
only dataset
(ChestX-ray)
Dataset spli
For our inte
achieve a 59
the Supplem
Testing and
QA/QC prog
For ChestX-l 『ミ c t c (゜ / F (´ F ´ T C ̄ F ヽ standards varied by dataset and included radiographic, clinical, and laboratory evidence of disease),
only ChestX-ray8 and VinDr-CXR also had labels indicating a variety of other abnormalities. Because
this imperfect overl only ChestX-ray8 and VinDr-CXR also had labels indicating a variety of other abnormalities. Because of
this imperfect overlap between our classification tasks and the labels in the datasets, VinDr-CXR is the
only dataset o ² this imperfect overlap between our classification tasks and the labels in the datasets, VinDr-CXR is the only dataset on which we test both binary models (1 and 2); for the other three, we test only Model 1 (ChestX-ray only dataset on which we test both binary models (1 and 2); for the other three, we test only Model 1
(ChestX-ray8) or Model 2 (MCU and SHN).
Dataset splitting
For our internal data, we began with 168,012 images in total, (ChestX-ray8) or Model 2 (MCU and SHN).

Dataset splitting

For our internal data, we began with 168,012 images in total, which we then randomly split into traini

(152,012; 15% abnormal), validation (8,000; 50% abnormal), Mataset splitting

For our internal data, we began with 168,0

(152,012; 15% abnormal), validation (8,000

following a sample size calculation we use

achieve a 5% margin of error in estimating

the Supplemental Methods). ムド(futficide) しんせいしょく Dataset splitting
For our internal data, we began with 168,012 images in total, which we then randomly split into training (152,012; 15% abnormal), validation (8,000; 50% abnormal), and testing (8,000; 50% abnormal) sets,
following a sample size calculation (8,000; 50% abnormal), and testing (8,000; 50% abnormal) sets,
following a sample size

(162))

(162) following a sample size calculation we used to determine the number of images we would need to

achieve a 5% margin of error in estimating sensitivity (technical details on the procedure are provided

the Sup achieve a 5% margin of error in estimating sensitivity (technical details on the procedure are provid
the Supplemental Methods). Training images were single-read images randomly drawn from all site
Testing and validation i the Supplemental Methods). Training images were single-read images randomly drawn from all sites.
Testing and validation images for Task 2 had either been double-read as part of the IOM Teleradiology
QA/QC program or sing Testing and validation images for Task 2 had either been double-read as part of the IOM Teleradiology
QA/QC program or single-read at a handful of panel sites in areas with high TB burden.

For ChestX-ray8, we reserved an QA/QC program or single-read at a handful of panel sites in areas with high TB burden.

For ChestX-ray8, we reserved an additional 8,000 images from the original training data to serve as

validation data for Task 1 (in ou ²

For ChestX-ray8, we reserved an additional 8,000 images from the original training data

validation data for Task 1 (in our internal validation dataset, abnormalities not suggestive

were underrepresented, as the abno トノノサ くりっく ゴード・くろく **Example 12**
For Chest Article and the Task 1 (in our internal validation dataset, abnormalities not suggestive of tubercu
were underrepresented, as the abnormal images were almost always abnormal and suggestive of
tubercu were underrepresented, as the abnormal images were almost always abnormal and suggestive of
tuberculosis).
Operating point selection
Operating point selection
When validation data was available, we used it to select two op tuberculosis).

Operating point selection

When validation data was available, we used it to select two operating points for thresholding the

Models' predictions on the corresponding test sets: one that maximized Youden's Operating poil
Operating poil
When validati
models' predict
models' predict
porating poir
images in our
latter being ge
period of time
minimized err
characteristics
Supplemental
datasets excer
on the test set
Mage preproc
 (\ r c cil F c c c c / / ´i *Deriating point selection*
When validation data was
models' predictions on thone
that minimized the re
operating points the "J" a
images in our internal test
latter being generally repr
period of time, the count-
minimize models' predictions on the corresponding test sets: one that maximized Youden's J index (all tasks)
one that minimized the relative error in predicted prevalence (Tasks 2 and 3 only). We named thes
operating points the "J" one that minimized the relative error in predicted prevalence (Tasks 2 and 3 only). We named these two operating points the "J" and "count" operating points, respectively. Because the proportion of abnormal latter being ge operating points the "J" and "count" operating points, respectively. Because the proportion of abnormal
images in our internal test set was different than the corresponding proportion in the training set, the
latter being images in our internal test set was different than the corresponding proportion in the training set, the
latter being generally representative of the screening program's data distribution over a multiyear
period of time, t latter being generally representative of the screening program's data distribution over a multiyear
period of time, the count-based operating points were selected using a reweighting scheme that
minimized error in predicti period of time, the count-based operating points were selected using a reweighting scheme that
minimized error in predicting the proportion from the training set using the model's performance
characteristics (sensitivity a minimized error in predicting the proportion from the training set using the model's performance
characteristics (sensitivity and specificity) on the validation set; this procedure is described in full
Supplemental Methods characteristics (sensitivity and specificity) on the validation set; this procedure is described in full is
upplemental Methods. Finally, when validation data was not available, as was the case for all ext
datasets except

Supplemental Methods. Finally, when validation data was not available, as was the case for all external datasets except for ChestX-ray8, we selected a single operating point that maximized Youden's J index on the test sets datasets except for ChestX-ray8, we selected a single operating point that maximized Youden's J index
on the test sets. We provide all operating points in Supplemental Table 2.
Image preprocessing
After discarding DICOM fi on the test sets. We provide all operating points in Supplemental Table 2.
Image preprocessing
After discarding DICOM files with corrupt pixel arrays, we extracted the pixel arrays and saved them as
1024x1024-pixel PNG fil Image preprocessing
After discarding DICOM files with corrupt pixel arrays, we extracted the pi
1024x1024-pixel PNG files. We then used optical character recognition (OC
images with evidence of burned-in patient metadata a Image preprocessing
After discarding DICOM files with corrupt pixel arrays, we extracted the pixel arrays and saved them as 1024x1024-pixel PNG files. We then used optical character recognition (OCR) software to identify images with evidence of burned-in patient metadata and removed them from the dataset. We describe both of these procedures mo

Model architecture and training procedures

images with evidence of burned-in patient metadata and removed them from the dataset. We des
both of these procedures more fully in the Supplemental Methods.
Model architecture and training procedures
To improve the model' both of these procedures more fully in the Supplemental Methods.
Model architecture and training procedures
To improve the model's ability to generalize to unseen data, we used a custom image augmentation
layer as the inpu Model architecture and training procedures
To improve the model's ability to generalize to unseen data, we use
layer as the input layer, randomly perturbing brightness, contrast, s
to the radiographs during training; value To improve the model's ability to generalize to unseen data, we used a custom image augmentation layer as the input layer, randomly perturbing brightness, contrast, saturation, and other characteristic to the radiographs during training; value ranges for these perturbations were taken from Majkowska
to the radiographs In the radiographs during training; value ranges for these perturbations were taken from Majkowska et
to the radiographs during training; value ranges for these perturbations were taken from Majkowska et
and the radiograph to the radiographs during training; value σ is the ranges for the set perturbations were taken from Majkowska et
al. () when σ and σ

Statistical inference

which was pretrained on ImageNet.³⁷ The final layers in our model were a dropout layer (probability
0.5, held fixed) and then a dense layer with a sigmoid activation and binary cross-entropy loss.
We trained all models i 0.5, held fixed) and then a dense layer with a sigmoid activation and binary cross-entropy loss.
We trained all models in minibatches of 12 images (4 per GPU) with the Adam³⁸ optimizer and a fixed
learning rate of 1e-4. We trained all models in minibatches of 12 images (4 per GPU) with the Adam³⁸ optimizer and idearning rate of 1e-4. For all tasks, we allowed training to continue until AUC began to decrease
validation data at which poin ハーハ へんこう りょうしょう elearning rate of 1e-4. For all tasks, we allowed training to continue until AUC began to decrease on the validation data at which point we saved the model weights and proceeded to testing.

Statistical inference

We calcu decrease of a hard many for the sate of the model weights and proceeded to testing.

Statistical inference

We calculated common classification performance metrics for all models and test sets, including AUC,

Sensitivity Statistical inference
We calculated common classification performance metrics for all models and test sets
sensitivity, specificity, and F1. For tuberculosis-specific datasets, we also calculated spensitivity and sensitiv We calculated common classification performance metrics for all models and test sets, including AUC, sensitivity, specificity, and F1. For tuberculosis-specific datasets, we also calculated specificity at 90%
sensitivity and sensitivity at 70% specificity, in line with the WHO's TPP for tuberculosis triage tests for
use i

sensitivity and sensitivity at 70% specificity, in line with the WHO's TPP for tuberculosis triage tests for
use in community settings. For the HaMLET test set, we calculated the model's relative error in
predicting preval see in community settings. For the HaMLET test set, we calculated the model's relative error in
predicting prevalence (i.e., the true number of abnormal-TB images), mirroring our primary operationa
use-case for the model predicting prevalence (i.e., the true number of abnormal-TB images), mirroring our primary ope
use-case for the model as a tool for internal QC activities.
For all metrics, we calculated bias-corrected and accelerated (BCA is use-case for the model as a tool for internal QC activities.
For all metrics, we calculated bias-corrected and accelerated (BCA) bootstrap confidence intervals^{39,39}
down-sampling abnormal images in the bootstrap repli The rate of the model and score for all metrics, we calculated bias-corrected and accelerat
down-sampling abnormal images in the bootstrap replicat
each was equal to the percentage in the training data (tar,
1; see the Sup For all metrics, we calculated and metrics, we calculated and according to the percentage of abnormal images i
each was equal to the percentage in the training data (target percentage of abnormal images i
each was equal to n
ab
deach was equal to the percentage in the training data (target percentage for each task provided in Tab
1; see the Supplemental Methods for more details). We did not adjust the intervals for multiplicity.
Abnormality locali 1; see the Supplemental Methods for more details). We did not adjust the intervals for multiplicity.
A*bnormality localization*
We used two saliency methods, Grad-CAM⁴⁰ and XRAI⁴¹, to generate abnormality heatmaps for This understand Methods, Grad-CAM⁴⁰ and XRA¹⁴¹, to generate abnormality heatmaps for the limages. We examined a small selection of the heatmaps for true-positive and false-positive images (abnormal and normal images, r メハ i (も ‐ く ミ / ミ し も せ Abnormally localization
We used two saliency me
images. We examined a s
(abnormal and normal in
explore their use as appr
truth bounding box anno
Software and Hardware
Our code is publicly avails
software and hardware u
Et images. We examined a small selection of the heatmaps for true-positive and false-positive images
(abnormal and normal images, respectively, with high model-based probabilities of abnormality) to
explore their use as appro

Software and Hardware
Our code is publicly available at https://github.com/cdcai/hamlet.git. Complete information on the

(abnormal and normal images, respectively, with high model-based probabilities of abnormality) to
explore their use as approximate abnormality localization methods. Because we did not have groun
truth bounding box annotati (explore their use as approximate abnormality localization methods. Because we did not have ground
truth bounding box annotation for the images, this step was primarily exploratory.

Software and Hardware

Our code is publ Ethical considerations
This project was proposed, reviewed, and approved in accordance with CDC institutional review policies

Results
Demographic characteristics of our study sample

the overall and training data, followed by applicants between 55 and 64 (12%) and those between 45 **Software and Hardware**

Our code is publicly available at $\underline{https://github.com/cdcai/hamlet_git}$. Complete infis

software and hardware used is available in the Supplemental Methods.
 Ethical considerations

This project was proposed, r こくら しょうしょうしょう しょうしょう software and hardware used is available in the Supplemental Methods.
 Ethical considerations

This project was proposed, reviewed, and approved in accordance with CDC institutional review pol

and procedures. Because it Finical considerations
This project was proposed, reviewed, and approved in accordance with
and procedures. Because it received a non-research determination, revi
board was not required. Neither trained model weights nor r d Takt Malakta d Ct and procedures. Because it received a non-research determination, review by an institutional review
board was not required. Neither trained model weights nor raw images will be made publicly available
to protect applicant board was not required. Neither trained model weights nor raw images will be made publicly available
to protect applicant privacy.
Results
Demographic characteristics of our study sample
Table 1 shows the demographic cha to protect applicant privacy.
 Results

Demographic characteristics of our study sample

Table 1 shows the demographic characteristics of the applicants in our study sample, and where their

screening exams were conducte Results
Demographic characteristics
Table 1 shows the demograp
screening exams were condu
between the ages of 15 and the
overall and training data,
and 54 (15%). In our validatic
underrepresented (19% in eac
overrepresente Bemographic entracteristics of our study sample
Table 1 shows the demographic characteristics of
screening exams were conducted. Overall, 55% of
between the ages of 15 and 44. Applicants 65 yea
the overall and training dat screening exams were conducted. Overall, 55% of the applicants were women, and most (64%) were
between the ages of 15 and 44. Applicants 65 years and older were the rarest, constituting 7% of both
the overall and training between the ages of 15 and 44. Applicants 65 years and older were the rarest, constituting 7% of bott
the overall and training data, followed by applicants between 55 and 64 (12%) and those between 45
and 54 (15%). In our the overall and training data, followed by applicants between 55 and 64 (12%) and those between 45 and 54 (15%). In our validation and test sets, applicants in the youngest age group (15 to 24) were underrepresented (19% i and 54 (15%). In our validation and test sets, applicants in the youngest age group (15 to 24) were underrepresented (19% in each), while those in the oldest age group (65 and over) were overrepresented (13% in each), rel underrepresented (19% in each), while those in the oldest age group (65 and over) were
overrepresented (13% in each), relative to the age distribution in both the overall sample and the
training data. overrepresented (13% in each), relative to the age distribution in both the overall sample
training data.
 $\frac{d}{dt}$ over presented (13% in each), relative to the age distribution in both the overall sample and the overall straining data. training data.

Corcastic Histly Southern Asia (13%), and Eastern Asia (7%) being the three primary contributors to the region. Bregion, most exams were conducted in Latin America and the Caribbean (23%), which was the meontributor to exam volume in the contributor to exam volume in the Americas region (26% of exams overall). By contrast, Oceania (1%)
and Europe (9%) had the smallest representation by volume. In our validation and test datasets, the
percentages changed su and Europe (9%) had the smallest representation by volume. In our validation and test datasets, these
percentages changed substantially, with Latin America and the Caribbean contributing 60% of the
images to both, and Sout percentages changed substantially, with Latin America and the Caribbean contributing 60% of the
images to both, and Southeastern and Southern Asia contributing 31%, with the remaining 9%
comprising images from Africa (6%) images to both, and Southeastern and Southern Asia contributing 31%, with the remaining 9%
comprising images from Africa (6%) and Europe (3%).
By data source, the majority of our images (88%) came from eMedical, the US Dep comprising images from Africa (6%) and Europe (3%).
By data source, the majority of our images (88%) came from eMedical, the US Department of Simmigrant health data system. Of the remaining images, 6% came from MiMOSA, IO By data source, the majority of our images (88%) cam
immigrant health data system. Of the remaining image
system, 3% from the IOM Teleradiology QC program, *i*
training dataset skewed heavily toward images from e
were even Filst / Fatt / Tass sa mmigrant health data system. Of the remaining images, 6% came from MiMOSA, IOM's health data
system, 3% from the IOM Teleradiology QC program, and 3% from our partner panel sites. Although t
training dataset skewed heavil

Distribution of abnormalities

system, 3% from the IOM Teleradiology QC program, and 3% from our partner panel sites. Although
training dataset skewed heavily toward images from eMedical (92%), the validation and test dataset
were evenly split between fraining dataset skewed heavily toward images from eMedical (92%), the validation and test datasets
were evenly split between eMedical (50%) and the IOM Teleradiology QC program (25%) and our
partner panels (25%). As menti were evenly split between eMedical (50%) and the IOM Teleradiology QC program (25%) and our
partner panels (25%). As mentioned above, the latter two primarily contributed images that were
abnormal and suggestive of TB, and matther panels (25%). As mentioned above, the latter two primarily contributed images that were
abnormal and suggestive of TB, and so we used eMedical as the source for normal images, noting
these latter images were also d abonormal and suggestive of TB, and so we used eMedicial as the source for normal images, noting
these latter images were also drawn from screenings performed by our partner panel sites and no
the system at random.
Distrib these latter images were also drawn from screenings performed by our partner panel sites and not from
the system at random.
Distribution of abnormalities
Table 2 shows the distribution of general and specific findings acr the system at random.

Distribution of abnormalities

Table 2 shows the distribution of general and specific findings across our internal training, validation,

and test sets. In the training data, 12% of the images were *Distribution of abnorme*
Table 2 shows the distr
and test sets. In the tra
suggestive of tuberculo
smear- or culture-posit
sets, these percentages
abnormal images to inc
additional images from
and our partner panel s
By s Distribution of abnormantes
Table 2 shows the distribution
and test sets. In the training c
suggestive of tuberculosis wit
smear- or culture-positive for
sets, these percentages chang
abnormal images to increase
additional and test sets. In the training data, 12% of the images were abnormal and 5% were abnormal and
suggestive of tuberculosis with 0.1% of images in the latter category from applicants who were either
smear- or culture-positive suggestive of tuberculosis with 0.1% of images in the latter category from applicants who were ein
smear- or culture-positive for tuberculosis disease at the time of screening. In the validation and
sets, these percentages sumear- or culture-positive for tuberculosis disease at the time of screening. In the validation and test
sets, these percentages changed to 50%, 50%, and 9% respectively, both because we up-sampled
abnormal images to incr

sets, these percentages changed to 50%, 50%, and 9% respectively, both because we up-sampled
abnormal images to increase precision in estimating sensitivity, and because we had requested
additional images from smear- or cu abnormal images to increase precision in estimating sensitivity, and because we had requested
additional images from smear- or culture-positive applicants from the IOM Teleradiology QC prog
and our partner panel sites to u additional images from smear- or culture-positive applicants from the IOM Teleradiology QC pre
and our partner panel sites to use for testing.
By specific finding, the most common abnormalities were the discrete linear opa and our partner panel sites to use for testing.
By specific finding, the most common abnormalities were the discrete linear opacity (2.4% in training;
20% in validation; 20% in testing) and the infiltrate or consolidation By specific finding, the most common abnorm
20% in validation; 20% in testing) and the infil-
data, the rarest abnormalities were miliary fin
(0.1%), and hilar/mediastinal adenopathy (0.1
data, despite the up-sampling of a - H- \ C C C / / / / C v C w L L C _ N 20% in validation; 20% in testing) and the infiltrate or consolidation (1.4%; 30%; 30%). In the training
data, the rarest abnormalities were miliary findings (<1%), cavitary lesions (0.1%), pleural effusions
(0.1%), and h data, the rarest abnormalities were miliary findings (<1%), cavitary lesions (0.1%), pleural effusions (0.1%), and hilar/mediastinal adenopathy (0.1%), all of which remained rare in the validation and test data, despite th (0.1%), and hilar/mediastinal adenopathy (0.1%), all of which remained rare in the validation and te data, despite the up-sampling of abnormal images.
 Binary classification performance
 Binary classification performan data, despite the up-sampling of abnormal images.

Binary classification performance

Table 3 shows the performance metrics for our models on the two binary classification tasks. For Task 2,

Table 3 shows the performance Binary classification performance
Table 3 shows the performance metrics for our mod
AUCs were consistently high, ranging from 0.99 (95
CXR. Specificity at a sensitivity of 0.90 was similarly
0.98 (0.89, 1.0) on MCU, althou ノコノ くく ミドロ トくへん しょうしゃ しょうしゃ しょうしゃ しょうしゃ しょうしゃ しょうしゃ しょうしゃ しょうしゃ しょうしゃ しゅうしゃ しんしゃ しんしゃ しんしゃ しんしゃ しんしゃ Binary classification performance
Table 3 shows the performance metrics for our models on the two binary classification tasks. For Task 2, AUCS were consistently high, ranging from 0.99 (95% CI 0.97, 1.0) on MCU to 0.94 (0.93, 0.95) on VinDr-CXR. Specificity at a sensitivity of 0.90 was similarly high, ranging from 0.83 (0.78, 0.87) on VinDr-CXR. Specificity 0.98 (0.89, 1.0) on MCU, although for reasons we provide in the discussion, we do not suggest whether any of these would meet the WHO's TPP for tuberculosis triage tests. With the count-based operating point, the model als any of these would meet the WHO's TPP for tuberculosis triage tests. With the count-based operating
point, the model also fared well in predicting the number of abnormal images suggestive of tuberculosis
in our internal te

point, the model also fared well in predicting the number of abnormal images suggestive of tuberculos
in our internal test set, achieving a relative error of only -2% (-8%, 6%).
Performance was similar, although slightly l in our internal test set, achieving a relative error of only -2% (-8%, 6%).
Performance was similar, although slightly lower, on Task 1. On ChestX-ray8, the model achieved an AUC
of 0.92 (0.90, 0.93) and an optimal sensit in our internal test set, achieving a relative error of only set, $(2.8, 6.9)$.

Of 0.92 (0.90, 0.93) and an optimal sensitivity and specificity of 0.82 (0.7

VinDr-CXR, these numbers were 0.89 (0.88, 0.90), 0.89, (0.87, 0 |
|
| of 0.92 (0.90, 0.93) and an optimal sensitivity and specificity of 0.82 (0.78, 0.84) and 0.86 (0.81, 0.89); on
VinDr-CXR, these numbers were 0.89 (0.88, 0.90), 0.89, (0.87, 0.90), and 0.73 (0.72, 75), respectively.
NotinDr VinDr-CXR, these numbers were 0.89 (0.88, 0.90), 0.89, (0.87, 0.90), and 0.73 (0.72, 75), respectively.
VinDr-CXR, these numbers were 0.89 (0.88, 0.90), 0.89, (0.87, 0.90), and 0.73 (0.72, 75), respectively. $V_{\rm c}$ (1.89, 0.88, 0.90), 0.89, (0.88, 0.90), 0.89, (0.89, 75), $V_{\rm c}$ (1.89, 75), respectively.

Because we did not have internal testing data for this task, we did not test this model with the count-

Multilabel classification performance

Multilabel classification
Table 4 shows the perf(0.8) in predicting discre
operating point, sensitive coverating point, sensitive
cavities, pleural effusio
operating point was hig
specificity for discrete r
metrics that de Table 4 shows the performance metrics for our models on Task 3. AUCs ranged from 0.78 (95% CI 0.73, 0.8) in predicting discrete linear opacities to 0.96 (0.65, 0.99) in predicting pleural effusions. With the J
operating point, sensitivity and specificity jointly reached 0.8 for four of the abnormalities: infiltrates,
cav

operating point selection is crucial for accuracy on this task with relative errors rising by several orders
of magnitude when the J operating point was used instead of the count operating point (Supplemental operating point, sensitivity and specificity jointly reached 0.8 for four of the abnormalities: infiltrates, cavities, pleural effusions, and volume loss or retraction. For the other abnormalities: infiltrates, cavities pl cavities, pleural effusions, and volume loss or retraction. For the other abnormalities, sensitivity at the operating point was higher than specificity, sometimes by a large margin (e.g., 0.91 sensitivity and 0.5 specifici operating point was higher than specificity, sometimes by a large margin (e.g., 0.91 sensitivity and 0.53
specificity for discrete nodules without classification). Because most abnormalities were rare in our data
metrics t specificity for discrete nodules without classification). Because most abnormalities were rare in our datametrics that depend on prevalence, like PPV and F1, were consistently low, with F1 reaching a maximum of 0.13 (0.10, metrics that depend on prevalence, like PPV and F1, were consistently low, with F1 reaching a maximum
of 0.13 (0.10, 0.15) for infiltrates and dipping as low as 0.00 (0.00, 0.01) for hilar adenopathies.
With the count oper of 0.13 (0.10, 0.15) for infiltrates and dipping as low as 0.00 (0.00, 0.01) for hilar adenopathies.
With the count operating point, the model produced fairly accurate prevalence estimates for a number
of the abnormalities Of the aboromalities. With the aboute produced fairly accurate prevalence estimates for a

for the aboromalities, with three absolute relative errors under 10% (2% for infiltrates, -6% for d

Ilnear opacities, and -9% for N C II C C T A II G II C C II A of the abnormalities, with three absolute relative errors under 10% (2% for infiltrates, -6% for discrete
linear opacities, and -9% for both kinds of nodules), and three under 20% (12% for irregular thick pleura
reactions,

Ilnear opacities, and -9% for both kinds of nodules), and three under 20% (12% for irregular thick pleur
reactions, 17% for pleural effusions, and 19% for volume loss or retraction). We note here that
operating point selec reactions, 17% for pleural effusions, and 19% for volume loss or retraction). We note here that
coperating point selection is crucial for accuracy on this task with relative errors rising by several orders
of magnitude whe repreacting point selection is crucial for accuracy on this task with relative errors rising by severa
of magnitude when the J operating point was used instead of the count operating point (Supple
Table 2).
Approximate abn of magnitude when the J operating point was used instead of the count operating point (Supplemental
Table 2).
Approximate abnormality localization
Figure 2 shows Grad-CAM (second and third columns) and XRAI (fourth and fif Table 2).

Approximate abnormality localization

Figure 2 shows Grad-CAM (second and third columns) and XRAI (fourth and fifth columns) for five

radiographs correctly identified as abnormal by Model 1; the radiographs we

Discussion
Applications of current models

Approxim
Figure 2 s
Figure 2 s
radiograp
abhorma
but for find
Discussio
Applicatio
Applicatio
Cur binar
points. Altasks, e.g
top-perfo
largely to
the exter Alfrakca II Ackthitti *Approximate abnormally hocalization*
Figure 2 shows Grad-CAM (second and
radiographs correctly identified as abn
abnormal examples in ChestX-ray8, SH
but for five images incorrectly identified
datasets. In general, the tw radiographs correctly identified as abnormal by Model 1; the radiographs were drawn from the
abnormal examples in ChestX-ray8, SHN, and VinDR-CXR. Figure 3 shows the same panel of heatm
but for five images incorrectly iden informal examples in ChestX-ray8, SHN, and VinDR-CXR. Figure 3 shows the same panel of head
but for five images incorrectly identified as abnormal, also by Model 1 and drawn from the same
datasets. In general, the two meth but for five images incorrectly identified as abnormal, also by Model 1 and drawn from the same
datasets. In general, the two methods identify similar regions of the radiographs as being abnormal,
although they do occasion datasets. In general, the two methods identify similar regions of the radiographs as being abnorm
although they do occasionally diverge in both the extent and the severity of the highlighted regio
Discussion
Applications o although they do occasionally diverge in both the extent and the severity of the highlighted regions.
 Discussion
 Applications of current models

Our binary models showed strong results on our internal datasets at bot Discussion
Discussion
Applications of current models
Our binary models showed strong results on our internal datasets at both of our chosen operating
Our binary models showed strong results on our internal datasets at bot 【 / C ドセーセイ セ ヽ w ff r / v f Our binary models showed strong results on our internal datasets at both of our chosen operating points. Although we do not currently intend for either model to be used for individual-level classifiests, e.g., as part of the overseas immigration exam clinical workflow, their performance is on-partop-performing models tasks, e.g., as part of the overseas immigration exam clinical workflow, their performance is on-par with top-performing models published in the literature, including several commercial products designed largely to detect top-performing models published in the literature, including several commercial products designed
largely to detect radiographic signs of tuberculosis (see Codlin 2021 and Kik 2022).^{42,43} Performance on
the external data largely to detect radiographic signs of tuberculosis (see Codlin 2021 and Kik 2022).^{42,43} Performance
the external datasets also suggests that the models may generalize well to unseen data, even on tas
for which we did n the external datasets also suggests that the models may generalize well to unseen data, even on tasks
for which we did not have clean validation or testing data, like Task 1, on which our models come withit
two percentage for which we did not have clean validation or testing data, like Task 1, on which our models come with
two percentage points of what we believe to be the current state-of-the-art on AUC (0.92 from Model
sv. 0.94 from Nabul two percentage points of what we believe to be the current state-of-the-art on AUC (0.92 from Model 1
vs. 0.94 from Nabulsi et al. 2021).²³ The main exception to this trend is Model 3, which despite achieving
good AUCs i vs. 0.94 from Nabulsi et al. 2021).²³ The main exception to this trend is Model 3, which despite achieving good AUCs in identifying many of the specific findings, is not likely to be clinically useful, at least in the po

good AUCs in identifying many of the specific findings, is not likely to be clinically useful, at least in the
population represented by our internal data, owing to the rarity of most of the findings and the model's
result population represented by our internal data, owing to the rarity of most of the findings and the model'
resulting poor PPV.
A relatively clear use-case for our tuberculosis-specific models (2 and 3), however, is in estimat population represents and the resulting poor PPV.
A relatively clear use-case for our tuberculosis-specific models (2 and 3), however, is in estimating
sample-level counts of abnormal images and, depending on the abnormali A relatively clear use
sample-level counts
findings. As a tool fc
exam, we can imagi ノ
sf
(Sample-level counts of abnormal images and, depending on the abnormality in question, specific findings. As a tool for conducting internal QC on radiographs read during the overseas immigration exam, we can imagine running sample-level conducting internal QC on radiographs read during the overseas immigration
exam, we can imagine running the models on batches of incoming images and comparing their nu
exam, we can imagine running the models finding we can imagine running the models on batches of incoming images and comparing their num
exam, we can imagine running the models on batches of incoming images and comparing their num-
exam, we can imagine running th exam, we can imagine running the models on batches of incoming images and comparing their number

counts exceeds a predefined threshold and triggering a model-guided manual review for further
investigation. In the case of Model 2, for example, which underestimated the number of abnormal
images suggestive of tuberculosi investigation. In the case of Model 2, for example, which underestimated the number of abnorm
images suggestive of tuberculosis in our internal test set by only 2%, a reasonable threshold mig
10%, just beyond the bounds of

images suggestive of tuberculosis in our internal test set by only 2%, a reasonable threshold might
10%, just beyond the bounds of the 95% CI (-8% to 6%), and the corresponding manual review might
begin with the images wit 10%, just beyond the bounds of the 95% CI (-8% to 6%), and the corresponding manual review might begin with the images with the highest model-based probability of abnormality among those initially reported normal. This ki begin with the images with the highest model-based probability of abnormality among those initially
reported normal. This kind of process may be stratified by key operational variables, like exam site or
country, and may a reported normal. This kind of process may be stratified by key operational variables, like exam site or
country, and may also be informed by existing epidemiologic information, like the expected backgrou
country, and may a country, and may also be informed by existing epidemiologic information, like the expected backgrount and ographs confirmed by earlier QC efforts.
The expected backgrount and ographs confirmed by earlier QC efforts.
Direc rates of tuberculosis disease in the screening areas or site-specific historical rates of abnormal
radiographs confirmed by earlier QC efforts.
Directions for future research
In our case, a natural first step for future re radiographs confirmed by earlier QC efforts.

Directions for future research

In our case, a natural first step for future research would be a follow-up validation study with n

review to explore our models' utility tools Directions for future research

In our case, a natural first step for future rese

review to explore our models' utility tools fo

examined the performance of already-traine

detecting abnormalities⁴³ and, in certain cas 1 | r c c | a c f c c c c c r | k t r Directions for future research
In our case, a natural first step for future research would be a follow-up validation study with manual

review to explore our models' utility tools for supporting internal QC efforts. A number of studies have
examined the performance of already-trained models, mostly in the form of commercial software, in
detecting abnormali examined the performance of already-trained models, mostly in the form of commercial software, in
detecting abnormalities⁴³ and, in certain cases, tuberculosis disease (Khan 2020; Qin 2021).^{25,30} To our
knowledge, ours detecting abormalities⁴³ and, in certain cases, tuberculosis disease (Khan 2020; Qin 2021).^{26,30} To our knowledge, ours is the first study to propose evaluating models' ability to estimate sample-level coun
and more ev knowledge, ours is the first study to propose evaluating models' ability to estimate sample-level count
and more evaluation would be needed before integrating them with existing QC workflows. A designe
QC study would also and more evaluation would be needed before integrating them with existing QC workflows. A designed QC study would also allow for the evaluation of attribution methods, like the saliency heatmaps we produced for Figures 2 a QC study would also allow for the evaluation of attribution methods, like the saliency heatmaps we
produced for Figures 2 and 3, as tools for abnormality localization to assist with manual review, which
only a small numbe produced for Figures 2 and 3, as tools for abnormality localization to assist with manual review, which, yand number of prior studies have rigorously addressed.^{44,45}
Similarly, an operational analysis to decide when, wh only a small number of prior studies have rigorously addressed.⁴⁴⁴⁵
Similarly, an operational analysis to decide when, where, and how to use the models to improve
screening programs would fill a gap in the literature, w only a small number of prior studies have rigorously addressed.

Similarly, an operational analysis to decide when, where, and how t

screening programs would fill a gap in the literature, which to date

model performance c's rikt Kesa Kardit ra screening programs would fill a gap in the literature, which to date has focused primarily on exam
model performance in clinical contexts rather than the downstream effects of incorporating the
harger workflows. Remaining model performance in clinical contexts rather than the downstream effects of incorporating them into
larger workflows. Remaining problems include establishing best practices for selecting operating points
by country or fac

larger workflows. Remaining problems include establishing best practices for selecting operating points
by country or facility to optimize detection given local constraints on resources and background rates of
tuberculosis by country or facility to optimize detection given local constraints on resources and background rates of
tuberculosis; estimating minimum diagnostic performance needed to achieve cost-effectiveness and
programmatic effici therculosis; estimating minimum diagnostic performance needed to achieve cost-effectiveness and
programmatic efficiency under different operating scenarios; and evaluating the epidemiologic and
economic impact of allowing programmatic efficiency under different operating scenarios; and evaluating the epidemiologic and
economic impact of allowing radiologists to use the models for decision support during the overseas
screening exams. There i economic impact of allowing radiologists to use the models for decision support during the overseas
screening exams. There is evidence that similar models can improve turnaround time²³ or lower cos
associated with diagno screening exams. There is evidence that similar models can improve turnaround time²³ or lower cost
associated with diagnostic workflows, and because of the scope of the CDC's overseas screening
program, these seem like p associated with diagnostic workflows, and because of the scope of the CDC's overseas screening
program, these seem like potentially fruitful avenues of investigation.
A final direction for future research is in developing program, these seem like potentially fruitful avenues of investigation.
A final direction for future research is in developing and evaluating models for predicting active
tuberculosis disease from chest radiographs in comb A final direction for future research is in developing and evaluating mortube
crubics disease from chest radiographs in combination with relevimmunologic information. To our knowledge, no model in the literatur
to predict Atitrifient and and area of the controller and a securious controller and a controller and controller and controll tuberculosis disease from chest radiographs in combination with relevant clinical, demographic,
immunologic information. To our knowledge, no model in the literature has been trained from to
pradiographic, clinical, and mi immunologic information. To our knowledge, no model in the literature has been trained from the sto predict tuberculosis disease from the radiograph directly, although training datasets may use a mixalographic, clinical, a to predict tuberculosis disease from the radiograph directly, although training datasets may use a mix of
radiographic, clinical, and microbiological results as reference standards; and no model has been
developed that acc radiographic, clinical, and microbiological results as reference standards; and no model has been
developed that accepts multimodal (i.e., non-radiographic) inputs. Given these limitations, performance
in predicting tuberc developed that accepts multimodal (i.e., non-radiographic) inputs. Given these limitations, perform in predicting tuberculosis disease is determined primarily by two pieces of information: a model's performance in identify in predicting tuberculosis disease is determined primarily by two pieces of information: a model's
performance in identifying images with abnormalities, and the correlation between the presence of
those abnormalities and t performance in identifying images with abnormalities, and the correlation between the presence of those abnormalities and tuberculosis disease in the target population. Occasionally the correlation high enough for models t in those abnormalities and tuberculosis disease in the target population. Occasionally the correlation is high enough for models to meet the TPP, but often it is not, with observed specificities at 90% sensitively strangi high enough for models to meet the TPP, but often it is not, with observed specificities at 90% sensitive ranging from well over 70%²⁶ to 60% and below²⁵ in several well-known commercial algorithms. Using other availa ranging from well over 70%²⁶ to 60% and below²⁵ in several well-known commercial algorithms. Using
other available information for prediction, whether by stacking an additional model on top of the image-
processing mo other available information for prediction, whether by stacking an additional model on top of the imag
processing module or by altering the module itself to accept multimodal inputs, may yield diagnostic
gains, not only be process in the module or by alternative or by alternative increases to accept model in the model and the model in
the model inputs of the model in pulle inputs of the model of the model of the
same state and the model in t gains, not only because the approach mirrors the process by which clinical diagnoses are made and thus β

seems a prior sensible, out also uceause patient transferences these my status and unstorped
tuberculosis are known to affect model performance based on the radiographs alone.²⁶⁴⁶ Alt
some technical innovation may be req some technical innovation may be required to make this approach feasible, it may well improve our
ability to predict tuberculosis disease, especially in low-resourced settings where access to trained
ability to predict tub ability to predict tuberculosis disease, especially in low-resourced settings where access to trained radiologists is limited, and thus seems well worth pursuing.
 Conclusion
 Conclusion
 Conclusion

Using data colle radiologists is limited, and thus seems well worth pursuing.

Conclusion

Using data collected from immigrants and refugees during overseas immigration exams prior to ent

into the US, we trained and tested three deep lear **Conclusion**

Using data collected from immigrants and refugees during ointo the US, we trained and tested three deep learning mod

radiographs. The models performed well, achieving high scc

state-of-the-art on several ex Using data c
Using data c
into the US,
radiographs
state-of-the
Acknowled;
The authors
Medica Inte
de la Fronte
as well as M
Ortega, and
References
1. World H
https://www
Date: 2023
Date: 2023 into the US, we trained and tested three deep learning models for identifying abnormalities on chest
radiographs. The models performed well, achieving high scores on our internal test dataset, and neari
state-of-the-art on radiographs. The models performed well, achieving high scores on our internal test dataset, and near
state-of-the-art on several external test datasets.
Acknowledgements
The authors thank the following panel sites and orga state-of-the-art on several external test datasets.
 Acknowledgements

The authors thank the following panel sites and organizations for their contribution to this work: Clinica

Medica Internacional (Mexico), Servicios **Acknowledgements**

The authors thank the following panel sites and on

Medica Internacional (Mexico), Servicios Medicos

de la Frontera (Mexico), and St. Luke's Medical Ce

QC, and the International Panel Physician Associ The authors thank th
Medica Internaciona
de la Frontera (Mexic
QC, and the Internati
as well as Mary Naug
Ortega, and Zachary
References
1. World Health Or_i
https://www.who.in
Date: 2023
Date accessed: April
2. Centers f Medica Internacional (Mexico), Servicios Medicos Consulares (Dominican Republic), Servicios Medicos
de la Frontera (Mexico), and St. Luke's Medical Center Extension Clinic (the Philippines), IOM Telerad
QC, and the Interna de la Frontera (Mexico), and St. Luke's Medical Center Extension Clinic (the Philippines), IOM Telerad QC, and the International Panel Physician Association. We also thank colleagues for review and supportax evell as Mary de, and the International Panel Physician Association. We also thank colleagues for review and suppose well as Mary Naughton, Jessica Webster, Nina Marano, Sifrash Gelaw, Bhaskar Amatya, Alexandra Ortega, and Zachary White

References

1. World Health Organization. Tuberculosis. https://www.who.int/news-room/fact-sheets/detail/tuberculosis

and as Well as Mary Naughton, Jessica Webster, Nina Marano, Sifrash Gelaw, Bhaskar Amatya, Alexandra
Ortega, and Zachary White.
Ortega, and Zachary White.
References
1. World Health Organization. Tuberculosis.
Date: 2023
D Ortega, and Zachary White.
 References

1. World Health Organization. Tuberculosis.

https://www.who.int/news-room/fact-sheets/detail/tuberculosis

Date: 2023

2. Centers for Disease Control and Prevention. Latent **TB In** References

1. World Health Organizati

https://www.who.int/news-

Date accessed: April 21, 202

2. Centers for Disease Contr

https://www.cdc.gov/tb/top

Date accessed: April 21, 202

3. Centers for Disease Contr

https:/ World Health Organization: Tuberculosis:

S://www.who.int/news-room/fact-sheets,

S: 2023

Paccessed: April 21, 2023

enters for Disease Control and Prevention.

S://www.cdc.gov/tb/topic/basics/tbinfect

S: 2020

Paccessed Date: 2023

Date: 2023

Date: 2023

2. Centers for Disease Control and Prevention. Latent TB Infection

https://www.cdc.gov/tb/topic/basics/tbinfectiondisease.htm

Date: 2020

Date accessed: April 21, 2023

3. Centers for Particulary
Date access
2. Centers f
https://ww
Date: 2021
Date access
4. Centers f
https://ww
Date: 2021
Date access
5. Centers f 2. Centers for Disease Control
https://www.cdc.gov/tb/topic
Date: 2020
Date accessed: April 21, 2023
3. Centers for Disease Control
https://www.cdc.gov/tb/sta
Date: 2021
4. Centers for Disease Control
https://www.cdc.gov/i - "I I I I I I I I I I I 2. Centers for Disease Control and Prevention. Latent TB Infection and TB Disease.
2. Centers for Disease Control and Prevention. Reported Tuberculosis in the United
Date accessed: April 21, 2023
3. Centers for Disease Con https://www.cdc.gov/tb/statistics/reports/2022/national
Date: 2020
3. Centers for Disease Control and Prevention. Reported Tube
https://www.cdc.gov/tb/statistics/reports/2022/nationa
Date: 2021
4. Centers for Disease Contr Factorial
Date access
3. Centers f
https://ww
Date: 2021
Date access
4. Centers f
bttps://ww
Date: 2021
Date access
5. Centers f
https://ww
Date: 2022
Date access Base Controlling

2. Centers for Disease Control

https://www.cdc.gov/tb/sta

Date accessed: April 21,2023

4. Centers for Disease Control

https://www.cdc.gov/immigra

Date 2021

Date accessed: December 19,

5. Centers fo

3. Centers for Disease Control and Tevention. The ported Tuberculosis in the United States, 2022.

Attps://www.cdc.gov/tb/statistics/reports/2022/national_data.htm

Date: 2021

2021

4. Centers for Disease Control and Prev

https://www.cdc.gov/tb/statistics/reports/2022/national_data.htm
Date: 2021
4. Centers for Disease Control and Prevention. **Technical Instructions for P**
https://www.cdc.gov/immigrantrefugeehealth/panel-physicians.html
Dat Date access
4. Centers f
https://ww
Date: 2021
Date access
5. Centers f
https://ww
Date: 2022
Date access Face accessed: pm-11,2011
4. Centers for Disease Contro
https://www.cdc.gov/immigr
Date: 2021
Date accessed: December 19,
5. Centers for Disease Contro
https://www.cdc.gov/immigr
Date: 2022
Date accessed: December 19, 4. Centers for Disease Control and Tevention. Technical Instructions for Panel Physicians.
https://www.cdc.gov/immigrantrefugeehealth/panel-physicians.html
Date accessed: December 19, 2022
5. Centers for Disease Control an Date: 2021
Date: 2021
S. Centers for Disease Control and Prevention. Immigrant and Refugee
https://www.cdc.gov/immigrantrefugeehealth/index.html
Date: 2022
Date accessed: December 19, 2022 Date access
5. Centers f
https://ww
Date: 2022
Date access B. Centers for Disease Control and P
https://www.cdc.gov/immigrantref
Date: 2022
Date accessed: December 19, 2022 ים
|}
| 5. Centers for Disease Control and Prevention. Immigrant and Refugee Health.
https://www.cdc.gov/immigrantrefugeehealth/index.html
Date: 2022
Date accessed: December 19, 2022 Date: 2022
Date accessed: December 19, 2022
Ate accessed: December 19, 2022 Date: 2022 Date accessed: December 19, 2022

incidence in immigrants and refugees bound for the United States: a population-based cross-sectional
study. *Ann Intern Med.* 2015;162(6):420-428. doi:10.7326/M14-2082
7. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Naraya istudy. Ann Intern Med. 2015;162(6):420-428. doi:10.7326/M14-2082

7. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K,

Madams T, Cuadros J, Kim R. Development and validation of a deep Study. Ann Intern Med. 2015,162(6):420-428. doi:10:1526/M14-2062

7. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A

Madams T, Cuadros J, Kim R. Development and validation of a deep le

diabetic retinopathy - T C W C L C C C C T C C C C Madams T, Cuadros J, Kim R. Development and validation of a deep learning algorithm for detection diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016 Dec 13;316(22):2402-10.

8. Poplin R, Varadarajan AV, Blume diabetic retinopathy in retinal fundus photographs. JAMA. 2016 Dec 13;316(22):2402-10.

8. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, Peng L, Webster DR. Prediction

of cardiovascular risk factors diabetic retinopathy in retinal fundus photographs. JAMA: 2010 Dec 13,516(22):2402-10.

8. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, Peng L, Webster

of cardiovascular risk factors from retinal f そくし くうしょう こくしょう

of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical*
Engineering. 2018 Mar;2(3):158-64.
9. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, of cardiovascular risk factors from Religions and deep learning. Nurtie Biometical
Engineering. 2018 Mar;2(3):158-64.
9. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, Brogi E, Reu
VE, Kl Engineering. 2018 Mar,2(3):158-04.
9. Campanella G, Hanna MG, Genesl
VE, Klimstra DS, Fuchs TJ. Clinical-gran
on whole slide images. Nature Media
10. Jaroensri R, Wulczyn E, Hegde N,
Dabbs DJ, Olson N. Deep learning medise civic since since since it since

VE, Klimstra DS, Fuchs TJ. Clinical-grade computational pathology using weakly supervised deep learning
on whole slide images. *Nature Medicine*. 2019 Aug;25(8):1301-9.
10. Jaroensri R, Wulczyn E, Hegde N, Brown T, Flament on whole slide images. *Nature Medicine*. 2019 Aug;25(8):1301-9.
10. Jaroensri R, Wulczyn E, Hegde N, Brown T, Flament-Auvigne I, Tan F, Cai Y, Nagpal K, Rakha EA,
Dabbs DJ, Olson N. Deep learning models for histologic gra on whole slide images. *Nature Medicine*. 2019 Aug;25(8):1301-9.
10. Jaroensri R, Wulczyn E, Hegde N, Brown T, Flament-Auvigne I,
Dabbs DJ, Olson N. Deep learning models for histologic grading of
disease prognosis. *NPJ Br* - ^ [C   / C   / U   / f   / 『 『 『

Dabbs DJ, Olson N. Deep learning models for histologic grading of breast cancer and association with
Dabbs DJ, Olson N. Deep learning models for histologic grading of breast cancer and association with
disease prognosis. disease prognosis. *NPJ Breast Cancer.* 2022 Oct 4;8(1):1-2.
11. Tolkach Y, Dohmgörgen T, Toma M, Kristiansen G. High-accuracy prostate cancer pathology using
deep learning. *Noture Mochine Intelligence.* 2020 Jul;2(7):411 disease prognosis. *NP S Breast cancer. 2022 Oct 4*;6(1):1-2.
11. Tolkach Y, Dohmgörgen T, Toma M, Kristiansen G. High
deep learning. *Nature Machine Intelligence*. 2020 Jul;2(7):4
12. Pierson E, Cutler DM, Leskovec J, Mul - ^ C - ^ L - ^ f - ^ 『 l - ^ L - r - r deep learning. *Nature Machine Intelligence*. 2020 Jul;2(7):411-8.
12. Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducionexplained pain disparities in underserved populations deep learning. Mature Machine Intelligence. 2020 Jul₁2(17):411-0.
12. Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer
unexplained pain disparities in underserved populations. *Nature*
13. Tiulpin A, Thevenot unexplained pain disparities in underserved populations. *Nature Medicine*. 2021 Jan;27(1):136-40.
13. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic knee osteoarthritis diagnosis
from plain radiograp unexplained pain disparites in underserved populations. Mutare Medicine. 2021 Jan;27(1):136-40.
13. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic knee osteoarthritis diagnos
from plain radiographs: a (一斤) (斤) (斤) () () () () from plain radiographs: a deep learning-based approach. Scientific reports. 2018 Jan 29;8(1):1-0.
14. Ueda D, Ehara S, Yamamoto A, Iwata S, Abo K, Walston SL, Matsumoto T, Shimazaki A, Yoshiyama
Miki Y. Development and Val From planning-epiter a deep learning-sected preports. The station of T, Shimazaki A, Yoshiy,
Miki Y. Development and Validation of Artificial Intelligence-based Method for Diagnosis of Mitra
Regurgitation from Chest Radiog (一)【) () 【) 【) () () () () / , ()

シーII ドー・ペーパー ペーパー・パー ロー・ロー・ロー

Regurgitation from Chest Radiographs. Radiology: Artificial Intelligence. 2022 Mar 2;4(2):e210221.
15. Nam JG, Park S, Hwang EJ, Lee JH, Jin KN, Lim KY, Vu TH, Sohn JH, Hwang S, Goo JM, Park CM.
Development and validation Example, Park S, Hwang EJ, Lee JH, Jin KN, Lim KY, Vu TH, Sohn JH, Hwang S, Goo JM, Park CM.
Development and validation of deep learning-based automatic detection algorithm for mailgnant
pulmonary nodules on chest radiogra Development and validation of deep learning-based automatic detection algorithm for malignant
pulmonary nodules on chest radiographs. *Radiology*. 2019 Jan;290(1):218-28.
16. Courtiol P, Maussion C, Moarii M, Pronier E, Pi pulmonary nodules on chest radiographs. *Radiology*. 2019 Jan;290(1):218-28.
16. Courtiol P, Maussion C, Moarii M, Pronier E, Pilcer S, Sefta M, Manceron P, Toldo S, Zaslavskiy I
Stang N, Girard N. Deep learning–based clas pulmonary nodules on chest radiographs. Radiology. 2019 Jan;250(1):210-20.
16. Courtiol P, Maussion C, Moarii M, Pronier E, Pilcer S, Sefta M, Manceron P,
Stang N, Girard N. Deep learning-based classification of mesothelio こくてん こくり こじょう

Miki Y. Development and Validation of Artificial Intelligence--based Method for Diagnosis of Mitral
Regurgitation from Chest Radiographs. Radiology: Artificial Intelligence. 2022 Mar 2;4(2):e210221.
15. Nam JG, Park S, Hwa Stang N, Girard N. Deep learning-based classification of mesothelioma improves prediction of patient
outcome. Nature medicine. 2019 Oct;25(10):1519-25.
17. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklun outcome. Nature medicine. 2019 Oct;25(10):1519-25.
17. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskays
Seekins J. Chexpert: A large chest radiograph dataset with uncerta 17. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chut
Seekins J. Chexpert: A large chest radiograph dataset v
Proceedings of the AAAI Conference on Artificial Intellig
18. Jaiswal AK, Tiwari P, Kumar S, Gupta D, Kha $\frac{1}{2}$ defined by $\frac{1}{2}$ for $\frac{$ Seekins J. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI Conference on Artificial Intelligence. 2019 Jul 17 (Vol. 33, No. 01, pp. 590-597).
18. Jaiswal

Proceedings of the AAAI Conference on Artificial Intelligence. 2019 Jul 17 (Vol. 33, No. 01, pp. 590-597).
18. Jaiswal AK, Tiwari P, Kumar S, Gupta D, Khanna A, Rodrigues JJ. Identifying pneumonia in chest X-rays: A deep l Proceedings of the AAAI Conference on Artificial Intelligence. 2019 Jul 17 (vol. 55, No. 01, pp. 550-557).
18. Jaiswal AK, Tiwari P, Kumar S, Gupta D, Khanna A, Rodrigues JJ. Identifying pneumonia in chest X-
rays: A deep $\frac{1}{2}$ rays: A deep learning approach. Measurement. 2019 Oct 1;145:511-8.

A Al Emadi N, Reaz MB. Can Al help in screening viral and COVID-19 pneumonia?. IEEE Access. 2020 Jul
20;8:132665-76.
20. Lee JH, Sun HY, Park S, Kim H, Hwang EJ, Goo JM, Park CM. Performance of a deep learning algorithm
co 20,8:132665-76.

20,8:132665-76.

20. Lee JH, Sun HY, Park S, Kim H, Hwang EJ, Goo JM, Park CM. Performance of a deep learning algorithm

compared with radiologic interpretation for lung cancer detection on chest radiograp

20. Lee JH, Sun H
20. Lee JH, Sun H
screening popula
21. Yoo H, Kim Kł
detection of mali
1;3(9):e2017135-
22. Hwang EJ, Pa
Development and
diseases on chest
23. Nabulsi Z, Sel
Yu J. Deep learnii
two unseen disea - < c v = < < c - - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c - < c screening population. Radiology. 2020 Dec;297(3):687-96.
21. Yoo H, Kim KH, Singh R, Digumarthy SR, Kalra MK. Validation of a deep learning algorithm for the
detection of malignant pulmonary nodules in chest radiographs. J 21. Yoo H, Kim KH, Singh R, Digumarthy SR, Kalra MK. Valid
detection of malignant pulmonary nodules in chest radiogr
1;3(9):e2017135-.
22. Hwang EJ, Park S, Jin KN, Im Kim J, Choi SY, Lee JH, Goo
Development and validation こくりょう こくしょう こんしょう

compared with radiologic interpretation for lung cancer detection on chest radiographs in a health
screening population. Radiology. 2020 Dec;297(3):687-96.
21. Yoo H, Kim KH, Singh R, Digumarthy SR, Kalra MK. Validation of

detection of malignant pulmonary nodules in chest radiographs. JAMA network open. 2020 Sep
1;3(9):e2017135-.
22. Hwang EJ, Park S, Jin KN, Im Kim J, Choi SY, Lee JH, Goo JM, Aum J, Yim JJ, Cohen JG, Ferretti GR.
Developmen 1,3(9):e2017135-.
22. Hwang EJ, Park S, Jin KN, Im Kim J, Choi SY, Lee JH, Goo JM, Aum J, Yim JJ, Cohen JG, Ferretti
Development and validation of a deep learning-based automated detection algorithm for major
diseases on c 22. Hwang EJ, Park
Development and
diseases on chest r
23. Nabulsi Z, Selle
Yu J. Deep learning
two unseen diseas
24. Tang YX, Tang
Automated abnorr
NPJ digital medicir
25. Tavaziva G, Ma
commercially avail
confirmed pulmo こしい こうせい こんじ こくい Development and validation of a deep learning-based automated detection algorithm for major tho
diseases on chest radiographs. JAMA Network Open. 2019 Mar 1;2(3):e191095-.
23. Nabulsi Z, Sellergren A, Jamshy S, Lau C, Sant diseases on chest radiographs. JAMA Network Open. 2019 Mar 1;2(3):e191095-.
23. Nabulsi Z, Sellergren A, Jamshy S, Lau C, Santos E, Kiraly AP, Ye W, Yang J, Pilgrim R, Kazemzadeh S,
Yu J. Deep learning for distinguishing n 23. Nabulsi Z, Sellergren A, Jamshy S, Lau C, Santos E, Kiraly AP, Ye W, Yang J, Pilg
Yu J. Deep learning for distinguishing normal versus abnormal chest radiographs
two unseen diseases tuberculosis and COVID-19. Scientifi こうし こんし こくりょう 23. Tavaziva G, Majidulla A, Nazish A, Saeel S, Benedeti A, Khan AJ, Khan AJ, Khan AJ, Khan AJ, Summers RM
24. Tang YX, Tang YB, Peng Y, Yan K, Bagheri M, Redd BA, Brandon CJ, Lu Z, Han M, Xiao J, Summers RM
24. Tang YX, T

two unseen diseases tuberculosis and COVID-19. *Scientific Reports*. 2021 Sep 1;11(1):1-5.
24. Tang YX, Tang YB, Peng Y, Yan K, Bagheri M, Redd BA, Brandon CJ, Lu Z, Han M, Xiao J, Summers RN
Automated abnormality classifi two unseen diseases tuberculosis and COVID-19. Scientific Reports. 2021 Sep 1;11(1;1-3.)
24. Tang YX, Tang YB, Peng Y, Yan K, Bagheri M, Redd BA, Brandon CJ, Lu Z, Han M, Xiao J,
Automated abnormality classification of che こんし こくり こくりょう こくりょう Automated abnormality classification of chest radiographs using deep convolutional neural networks.

NPJ digital medicine. 2020 May 14;3(1):70.

25. Tavaziva G, Majidulla A, Nazish A, Saeed S, Benedetti A, Khan AJ, Khan FA NPJ digital medicine. 2020 May 14;3(1):70.

25. Tavaziva G, Majidulla A, Nazish A, Saeed S, Benedetti A, Khan AJ, Khan FA. Diagnostic accuracy of a

commercially available, deep learning-based chest X-ray interpretation so NAT angula mediatry acts and y any any any and the commercially available, deep learning-base
confirmed pulmonary tuberculosis. Internat
26. Qin ZZ, Ahmed S, Sarker MS, Paul K, Add
detection from chest x-rays for triaging こくく こくさ こうトー こうじょう commercially available, deep learning-based chest X-ray interpretation software for detecting culture-
confirmed pulmonary tuberculosis. International Journal of Infectious Diseases. 2022 Sep 1;122:15-20.
26. Qin ZZ, Ahmed 26. Qin ZZ, Ahmed S, Sarker MS, Paul K, Adel AS, Naheyan T, Barrett R, Banu S, Creswell J. Tuberculosis
detection from chest x-rays for triaging in a high tuberculosis-burden setting: an evaluation of five
artificial intel 14 C te (14 F te F C te F C t detection from chest x-rays for triaging in a high tuberculosis-burden setting: an evaluation of five
artificial intelligence algorithms. *The Lancet Digital Health.* 2021 Sep 1;3(9):e543-54.
27. Majkowska A, Mittal S, Ste artificial intelligence algorithms. *The Lancet Digital Health.* 2021 Sep 1;3(9):e543-54.
27. Majkowska A, Mittal S, Steiner DF, Reicher JJ, McKinney SM, Duggan GE, Eswaran K, Cameron C
PH, Liu Y, Kalidindi SR, Ding A. Che artificial intelligence algorithms. The Editect Digital Health. 2021 Sep 1;3(9):e343-54.
27. Majkowska A, Mittal S, Steiner DF, Reicher JJ, McKinney SM, Duggan GE, Eswaran
PH, Liu Y, Kalidindi SR, Ding A. Chest radiograph 14月八月 14月(14月) 10月 PH, Liu Y, Kalidindi SR, Ding A. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*. 2020
Feb;294(2):421-3

with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*. 2020
Feb;294(2):421-31.
28. Putha P, Tadepalli M, Reddy B, Raj T, Chiramal JA, Govil S, Sinha N, KS M, Reddivari S, Jagirdar Wirraamologist rugali M, Reddy B, Raj T, Chiramal JA, Govil S, Sinha N, KS M, Reddivari S, Jagirdar A, Radepolli M, Reddy B, Raj T, Chiramal JA, Govil S, Sinha N, KS M, Reddivari S, Jagirdar A, Radio. The D.Can artificial reb,254(2):421-51.
28. Putha P, Tadepa
P. Can artificial intel
on 2.3 million x-rays
29. Kazemzadeh S, \
Kiraly AP. Deep Lear
Clinical Performanc
30. Khan FA, Majidu
Saeed S. Chest x-ray

Kiraly AP. Deep Learning Detection of Active Pulmonary Tuberculosis at Chest Radiography Matched the
Clinical Performance of Radiologists. Radiology. 2022 Sep 6:212213. P. Can artificial intelligence reliably report chest x-rays?: Radiologist validation of an algorithm trained
on 2.3 million x-rays. arXiv preprint arXiv:1807.07455. 2018 Jul 19.
29. Kazemzadeh S, Yu J, Jamshy S, Pilgrim R, on 2.3 million x-rays. arXiv preprint arXiv:1807.07455. 2018 Jul 19.
29. Kazemzadeh S, Yu J, Jamshy S, Pilgrim R, Nabulsi Z, Chen C, Beladia N, Lau C, McKinney SM, Hughes
Kiraly AP. Deep Learning Detection of Active Pulmon 29. Kazemzadeh S, Yu J, Jamshy S, Pilgrim R, Nabulsi Z, Chen C, Bela
Kiraly AP. Deep Learning Detection of Active Pulmonary Tuberculos
Clinical Performance of Radiologists. *Radiology*. 2022 Sep 6:212213
30. Khan FA, Majid これは くうこく Kiraly AP. Deep Learning Detection of Active Pulmonary Tuberculosis at Chest Radiography Matched the
Clinical Performance of Radiologists. *Radiology*. 2022 Sep 6:212213.
30. Khan FA, Majidulla A, Tavaziva G, Nazish A, Abi

Clinical Performance of Radiologists. *Radiology*. 2022 Sep 6:212213.
SO. Khan FA, Majidulla A, Tavaziva G, Nazish A, Abidi SK, Benedetti A, Menzies D, Johnston JC, Khan AJ,
Saeed S. Chest x-ray analysis with deep learning Clinical Performance of Radiologists. Rudiology. 2022 Sep 0.212213.
30. Khan FA, Majidulla A, Tavaziva G, Nazish A, Abidi SK, Benedetti A,
Saeed S. Chest x-ray analysis with deep learning-based software as a $\frac{1}{2}$ Saeed S. Chest x-ray analysis with deep learning-based software as a triage test for pulmonary

Saeed S. Chest x-ray analysis with deep learning-based software as a triage test for pulmonary

The A, Menzie of the A, Menzie Saeed S. Chest x-ray analysis with deep learning-based software as a triage test for pulmonary

Approvals: Electronic Medical Examination for Visa or Refugee Applicant.
https://www.requlations.gov/document/DOS FRDOC 0001-5900

Date: 2020. Date accessed: July 3, 2023

Health. 2020 Nov 1;2(11):e573-81.

31. US Department of State. Agency Information Collection Activities; Proposals, Submissions, and

Approvals: Electronic Medical Examination for Visa or Refugee Applicant.

Approvals: Ele Health. 2020 Nov 1;2(11):e573-81. ミノト II ミトII ミ CC ミ C 31. O Separation of State Record, Technic Medical Examination for Visa or Refugee Applicant.
Approvals: Electronic Medical Examination for Visa or Refugee Applicant.

2022

Date accessed: June 30, 2023

32. OMB. DS-3030 Tu https://www.regulations.gov/document/DOS_FRDOC_0001-5900
Date: 2022
Date accessed: June 30, 2023
32. OMB. **DS-3030 Tuberculosis Worksheet: Medical Exan**
https://omb.report/icr/202010-1405-004/doc/105591400https://c
Date: 2 Date access

32. OMB. **D**

https://omb.

Date: 2020.

Date accesse

33. Wang X,

database ar

diseases. In

2097-2106)

34. Jaeger S

computer-a

Dec;4(6):47

35. Nguyen,

annotations Example 2022

22. OMB. **DS-3030 Tuberculos**

https://omb.report/icr/202010-14

Date accessed: July 3, 2023

33. Wang X, Peng Y, Lu L, Lu Z,

database and benchmarks on

diseases. In *Proceedings of the*

2097-2106).

34. J ミトII ミ(くる)ミ(I) ミ ミ 22. OMD: D23-3030 Tuberculosis Worksheet: "Medical Examination for immigrant or netroge expineding

Date: 2020.

33. Wang X, Peng Y, Lu L, Lu Z, https://omb.report/icr/202010-1405-004/doc/105591400https://omb.report/icr/202010-1405-004/doc/105591400
Date accessed: July 3, 2023
Date accessed: July 3, 2023
33. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx ミ(くくこう) くしょう ミラリー ミニュ database and benchmarks on weakly-supervised classification and localization of common thoradiseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 20
2097-2106).
34. Jaeger S, Candemi

diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017
2097-2106).
34. Jaeger S, Candemir S, Antani S, Wáng YX, Lu PX, Thoma G. Two public chest X-ray datasets for
computer-aided diseases. In Trocecanings of the IEEE Conference on Computer Vision and Pattern necognition. 2017 (pp.
2097-2106).
34. Jaeger S, Candemir S, Antani S, Wáng YX, Lu PX, Thoma G. Two public chest X-ray datasets for
2007-2106) ⁷
2097 S.
2007-2106:475
2007-2106:475
2097-2106:58
2009 IEEE co
2009 IEEE co
2009 IEEE co
2009 IEEE co
2009 IEEE co
2009 IEEE co ミ(I) ミ(ミ) ミノ ミニ ミー ミア computer-aided screening of pulmonary diseases. *Quantitative Imaging in Medicine and Surgery*.
Dec;4(6):475.
35. Nguyen, H.Q., Lam, K., Le, L.T. et al. VinDr-CXR: An open dataset of chest X-rays with radiologis
annotation

35. Nguyen, H
35. Nguyen, H
annotations. S
36. Tan M, Le
Machine Leari
37. Deng J, Do
2009 IEEE con
38. Kingma DF
Dec 22.
39. Efron B. Be
Mar 1;82(397)
40. Selvaraju F

computer-aided screening or pulmonary useases. Quantitative integrity in Medicine and Surgery. 2014
25. Nguyen, H.Q., Lam, K., Le, L.T. et al. VinDr-CXR: An open dataset of chest X-rays with radiologist's
annotations. Sci

ミミ ミノ ミニ ミー ミー くく annotations. Sci Data 9, 429 (2022). https://doi.org/10.1038/s41597-022-01498-w.
36. Tan M, Le Q. EfficientnetV2: Smaller models and faster training. In International Conference on
Machine Learning. 2021 Jul 1 (pp. 10096-1 annon and the state of the state 9, 120 (112). Smaller models and faster training. In International
Machine Learning. 2021 Jul 1 (pp. 10096-10106). PMLR.
37. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A la ミノ ミニ ミー ミーム くくし 36. Tan W, Economic Learning. 2021 Jul 1 (pp. 10096-10106). PMLR.
37. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image databa
2009 IEEE conference on computer vision and pattern Machine Ecarning. 2021 Jul 1 (pp. 10050-10100). PMLR.
37. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imager
2009 IEEE conference on computer vision and pattern re
38. Kingma DP, Ba J. Adam: A method for stochastic o ミニュ ミー・ミー くくく 2009 IEEE conference on computer vision and pattern recognition 2009 Jun 20 (pp. 248-255). IEEE.
28. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014
Dec 22.
39. Efron B. Be 2008 IEEE conference on confidence optimization. arxiv preprint arxiv:1412.6980. 2
20022.
2008 IEFron B. Better bootstrap confidence intervals. Journal of the American Statistical Association.
1
2008 IEFron B. Better boots ミー・ミー くくりょう くりょう Dec 22.
39. Efron B. Better bootstrap confidence intervals. Journal of the American Statistical Association. 1987
Mar 1;82(397):171-85.
40. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual e 39. Efro
39. Efro
Mar 1;8.
40. Selv:
deep ne
comput
41. Kapi
Proceed
42. Codl
Sidney-/ ミ『 』 () / [/ <] (/ [/ <] (39. Enomin. Better bootstrap connuetnee intervals. Journal of the American Sudratical Association. 1997
And 1,82(397):171-85.
40. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanation 40. Selvaraju RR, Cogsw
deep networks via grad
computer vision 2017 (
41. Kapishnikov A, Bolu
Proceedings of the IEEE
42. Codlin AJ, Dao TP, V
Sidney-Annerstedt K, So
detection of tuberculos z c c z f z c c z

deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision 2017 (pp. 618-626).
41. Kapishnikov A, Bolukbasi T, Viégas F, Terry M. Xrai: Better attributions through computer vision 2017 (pp. 618-626).
41. Kapishnikov A, Bolukbasi T, Viégas F, Terry M. Xrai: Better attributions through regions. In
Proceedings of the IEEE/CVF International Conference on Computer Vision 2019 (pp. 4948-49 2011 Kapishnikov A, Bolukbasi T, Viéga
Proceedings of the IEEE/CVF Internat
42. Codlin AJ, Dao TP, Vo LN, Forse R.
Sidney-Annerstedt K, Squire B. Indep
detection of tuberculosis. Scientific re ムト くらく Proceedings of the IEEE/CVF International Conference on Computer Vision 2019 (pp. 4948-49!
42. Codlin AJ, Dao TP, Vo LN, Forse RJ, Van Truong V, Dang HM, Nguyen LH, Nguyen HB, Nguye
Sidney-Annerstedt K, Squire B. Independe Proceedings of the IEEE, of the International Conference on Computer View Level, (pp. 4948-4957).
A2. Codlin AJ, Dao TP, Vo LN, Forse RJ, Van Truong V, Dang HM, Nguyen LH, Nguyen HB, Nguyen N
Sidney-Annerstedt K, Squire B. $\frac{2}{3}$ $62.$ Sidney-Annerstedt K, Squire B. Independent evaluation of 12 artificial intelligence solutions for the detection of tuberculosis. Scientific reports. 2021 Dec 13;11(1):23895. detection of tuberculosis. Scientific reports. 2021 Dec 13;11(1):23895.

The Connection of tuberculosis. Scientific reports. 2021 Dec 13;11(1):23895. detection of tuberculosis. Scientific reports. 2021 Dec 13;11(1):23895.

ノ(こを ノー ノー ノミヽ Garfin AM, Marks GB. Diagnostic accuracy of chest X-ray interpretation for tuberculosis by three
artificial intelligence-based software in a screening use-case: an individual patient meta-analysis of
global data. medRxiv.

artificial intelligence-based software in a screening use-case: an individual patient meta-analysis global data. medRxiv. 2022 Jan 27:2022-01.
44. Arun N, Gaw N, Singh P, Chang K, Aggarwal M, Chen B, Hoebel K, Gupta S, Pat global data. medRxiv. 2022 Jan 27:2022-01.
44. Arun N, Gaw N, Singh P, Chang K, Aggarwal M, Chen B, Hoebel K, Gupta S, Patel J, Gidwani M,
Adebayo J. Assessing the trustworthiness of saliency maps for localizing abnormalit global data. Media and Presentation 2022
44. Arun N, Gaw N, Singh P, Chang K, Aggary
Adebayo J. Assessing the trustworthiness of
imaging. Radiology: Artificial Intelligence. 20
45. Saporta A, Gui X, Agrawal A, Pareek A, T Adebayo J. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical
imaging. Radiology: Artificial Intelligence. 2021 Oct 6;3(6):e200267.
45. Saporta A, Gui X, Agrawal A, Pareek A, Truong SQ, imaging. Radiology: Artificial Intelligence. 2021 Oct 6;3(6):e200267.
45. Saporta A, Gui X, Agrawal A, Pareek A, Truong SQ, Nguyen CD, Ngo VD, Seekins J, Blankenberg
AY, Lungren MP. Benchmarking saliency methods for chest 45. Saporta A, Gui X, Agrawal A, Pareek A, Truong SQ, Nguyen CD, N
AY, Lungren MP. Benchmarking saliency methods for chest X-ray intelligence. 2022 Oct;4(10):867-78.
46. Kagujje M, Kerkhoff AD, Nteeni M, Dunn I, Mateyo K, ノノー くらい しょうしょう

AY, Lungren MP. Benchmarking saliency methods for chest X-ray interpretation. Nature Machine
Intelligence. 2022 Oct;4(10):867-78.
46. Kagujje M, Kerkhoff AD, Nteeni M, Dunn I, Mateyo K, Muyoyeta M. The performance of compu Intelligence. 2022 Oct;4(10):867-78.
A6. Kagujje M, Kerkhoff AD, Nteeni M, Dunn I, Mateyo K, Muyoyeta M. The performance of compaided detection digital chest X-ray reading technologies for triage of active tuberculosis amo 16. Kagujje M, Kerkhoff AD, Nteeni Naided detection digital chest X-ray re
with a history of previous tuberculos
with a history of previous tuberculos ム
こい
 and detection detection detection detection digital chest X-ray reading technologies for trial infectious Diseases. 2023 Feb 1;76(3):e894-901. with a history of previous tuberculosis. Clinical Infectious Diseases. 2023
The following Diseases. 2023
Feb 1;76(3):e894-901.

*Exam Region:
Africa: Algeria, Angola, Benin, Burkina Faso, Burundi, Cameroon, Cape Verde, Chad, Democratic Republic of the Congo, Republic of the Congo, Djibout
Gambia, Ghana, Guinea, Ivory Coast, Kenya, Liberia, Malawi, Africa: Algeria,
Gambia, Ghana
Togo, Tunisia, L
Americas: Arge
Jamaica, Mexic
Asia: Afghanist
Kuwait, Kyrgyzs
Thailand, Turke
Europe: Albani,
Hungary, Italy,
Hungary, Italy,

Western Asia 9196 5 9088 6 62 1 46 1 M/A 385 0 380 0 1 0 4 0
gy and Quality Control Center (GTQCC) and the private panel sites all have their own internal quality control processes.
Burkina Faso, Burundi, Cameroon, Cape Verde, Chad, Democratic Republic of the Control Center (GTQCC) and the private panel sites all have their own internal quality control processes.

Irundi, Cameroon, Cape Verde, Chad, Democratic Republic of the Congo, Republic of the Congo, Djibouti, Egypt,

peri Africa: Angola, Ghana, Guinea, Ivorkina 1995, Parlami, Cameroon, Cape Verde, Incornect Nepation, Incornect Nepation, Incorect Nepation, Morocco, Mozambique, Namibia, Niger, Nigeria, Rwanda, Senegal, South Africa, Sudan, Ta Togo, Tunisia, Uganda, Zambia, and Zimbabwe
Americas: Argentina, Barbados, Belize, Bolivia, Brazil, Canada, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, Guatemala, Guyana, Haiti, Honduras,
Amarica, Mexico, Nic **Americas:** Argentina, Barbados, Belize, Bolivia,
Jamaica, Mexico, Nicaragua, Panama, Paraguay
Asia: Afghanistan, Azerbaijan, Bahrain, Banglad
Kuwait, Kyrgyzstan, Laos, Lebanon, Malaysia, M
Thailand, Turkey, Turkmenistan Americas: Argentina, businca, businca, businca, businca, Unididal and Tobago, and Uruguay
Asia: Afghanistan, Azerbaijan, Bahrain, Bangladesh, Cambodia, China, Hong Kong, India, Indonesia, Iraq, Israel, Japan, Jordan, Kazak Asia: Afghanistan, Azerbaijan, Bahrain, Bangladesh, Cambodia, China, Hong Kong, India, I
Asia: Afghanistan, Azerbaijan, Bahrain, Bangladesh, Cambodia, China, Hong Kong, India, I
Kuwait, Kyrgyzstan, Laos, Lebanon, Malaysia, Asia: Anghanistan, Azerbaijan, Danian, Danigalesh, Cambodia, Unia, Thomg Kong, India, Ind Thailand, Turkey, Turkmenistan, United Arab Emirates, Uzbekistan, Vietnam, and West Bank
Europe: Albania, Austria, Belarus, Belgium, Bosnia and Hercegovina, Bulgaria, Croatia, Cyprus, Denmark, Estonia, Finland, France, Geo Europe: Albania, Austria, Belarus, Belgium, Bosnia and Hercegovina, Bulgaria, Croatia, Cyprus
Hungary, Italy, Kosovo, Latvia, Lithuania, Malta, Moldova, Netherlands, North Macedonia, No
Slovenia, Spain, Sweden, Switzerland Europe: Austria, Belarus, Austria, Belarus, Doshin and The Customa, Danian, School, Cyprus, Denmark, Estonia, Thania, Hercegovina, Denmark, Croatia, Croatia, Croatia, Serbia, Slovak Republic Slovak Republic Slovak Network Slovenia, Spain, Sweden, Switzerland, Ukraine, and United Kingdom
Oceania: Australia, Fiji, New Zealand, Papua New Guinea, and Tonga
Decania: Australia, Fiji, New Zealand, Papua New Guinea, and Tonga

Slovenia, Spain, Slovenia, Switzerland, Papua New Guinea, and Tonga
Oceania: Australia, Fiji, New Zealand, Papua New Guinea, and Tonga
 Oceania: Australia, Fiji, New Zealand, Papua New Guinea, and Tonga

SM/CX+ at Current Exam 1551 0.9 150 0.1 689 8.6 712 8.9
sitive TB Status (+) is defined as having positive sputum smear (SM) and/or culture (CX) results at either the latest
orded exam in the screening process ("Current") iatus (+) is defined as having positive sputum smear (SM) and/or culture (CX) results at either the latest in in the screening process ("Current") or at one of the previous exams in the same ("Prior").
In the screening pro recorded exam in the screening process ("Current") or at one of the previous exams in the same ("Prior").
Table 2. Counts and proportions of radiographic findings suggestive of tuberculosis (TB) in our in
training, validat recorded manufacturing process ("Current") or at our process factor in the process ("Current").

Table 2. Counts and proportions of radiographic findings suggestive of tuberculosis (TB

training, validation, and testing da Table 2. Counts and proportions of radiographic findings suggestive of tuberculosis (TB) in our internal Table 2. Counts and proportions of radiographic findings suggestive of tuberculosis (TB) in our internal
training, validation, and testing datasets. training, validation, and testing datasets.

*Datasets included: HaMLET; VinDr-CXR, an open dataset of chest x-rays with radiologist annotations from 2 major hospitals in Vietnam; Shenzhen, collected from outpatient clinics and collected in collaboration with Shenzhen No.3 People's Hospital, Guangdong Medical College, Shenzhen, China; Montgomery, chest x-ray dataset collected in Montgomery County, Maryland, in collaboration with the Department of Health and Human Services; ChestX-ray8, hospital-scale chest x-ray database with disease image labels, provided by the National Institutes for Health (NIH).

Table 3. Classification results for our abnormal-TB and abnormal models on our internal dataset and on four external datasets.

1.44)

Table 4. Classification metrics for our findings-specific model on our internal dataset.

* Data sources were eMedical, the US Department of State's immigrant health information system; the International `Data sources were eMedical, the US Department of State's immigrant health information system; the International
Organization for Migration's Global Teleradiology and Quality Control Centre (IOM GTQCC); individual private and MiMOSA, IOM's migrant health information system

*All ages ranged from 0 – 102 years old

Figure 1. Flow diagram detailing the number of radiographs by data source and abnormality status from before exclusion for age (All Ages), after exclusion for age under 15 years, and after exclusion for whether the DICOM pixel arrays were Python-readable (Valid Images). Radiographs were collected between the years of 2011 and 2021 and constitute a convenience sample of the total applicant population for that time period. convenience sample of the total applicant population for that time period.

Figure 2. XRAI (left) and GradCAM (right) heatmaps for true positive images. The original radiographs are
on the left, the Grad-CAM activations and heatmaps are in the middle, and the XRAI activations and
overlays are on t overlays are on the right. For the XRAI overlays, only regions reaching the 70th percentile of activation
strength are shown. strength are shown.

Figure 3. XRAI (left) and GradCAM (right) heatmaps for false positive images. The original radiographs are on the left, the Grad-CAM activations and heatmaps are in the middle, and the XRAI activations and overlays are on the right. For the XRAI overlays, only regions reaching the 70th percentile of activation strength are shown. activation strength are shown.

SUPPLEMENT
Additional Information on the Overseas Screening Procedures

X-ray images collected during the overseas health screenings are interpreted by a radiologist at the
panel physician's office (panel site) and reviewed by the panel physician. The process is to first determine if the chest x-ray image is normal or abnormal, if abnormal then the radiologist will parel physician's office (panel of an extending the panel site) and the panel in the padiologist will determine if the image suggests that the applicant has tuberculosis (TB). If the image is suggest
determine if the image determine is the chest x-ray image is normal or abnormal, is absorbed in the radiologist will
determine if the image suggests that the applicant has tuberculosis (TB). If the image is suggest
then the applicant needs sputu then the applicant needs sputum smears and culture testing done. For the images that suggest the
applicant has TB, the radiologist identifies and records the specific radiographic features associated with then the applicant needs spatial culture and culture testing active testing condets spaces in a supplicant has
applicant has TB, the radiologist identifies and records the specific radiographic features associated
this det this determination, and the panel physician then determines each applicant's TB classification and
records that information on the applicant's US Department of State TB Worksheet (DS-3030 form). records that information on the applicant's US Department of State TB Worksheet (DS-3030 form).

Any applicant diagnosed with pulmonary TB disease must receive a classification of Class A TB and is not cleared for travel until successfully finishing directly observed treatment under the supervision of a
panel physician. Those who receive a classification of Class B0 TB (applicants, previously classified as cleared for travel until successfully minimally in surgering the supervision and supervision of panel physician.
The supervision of a supervision of a supervision of also separate treatment under the supervision of a
Class panel physician. These who receive a classification of Class B1 TB (applicants with signs or symptoms or chass
Class A, who successfully completed treatment), Class B1 TB (applicants with signs or symptoms or ch
x-ray find x-ray findings suggestive of TB disease, or known HIV infection, but negative sputum smears and
cultures and are not diagnosed with TB disease), Class B2 TB (applicants with a positive IGRA, but yations and are not diagnosed with TB disease), Class B2 TB (applicants with a positive IGRA, but
otherwise have a negative evaluation for TB; i.e., latent TB infection), or Class B3 TB (applicants w cultures and are not diagnosed with TB disease), Cheese CR (applicants with a positive IGRA, and
otherwise have a negative evaluation for TB; i.e., latent TB infection), or Class B3 TB (applicants w
a recent contact of a k a recent contact of a known TB disease case, regardless of IGRA or TST results) are cleared and permitted to travel. a recent contact of a known TB disease case, regardless of TST results) are cleared and
permitted to travel.

Supplemental Methods $\ddot{\cdot}$

**Supplemental Methods
Sample size calculation
We used a simulation pr** validation and test sets to accurately estimate sensitivity, specificity, and other key performance metrics
for our models. As a starting point, we assumed the models would achieve a sensitivity and specificity of for our models. As a starting point, we assumed the models would achieve a sensitivity and specificity of
80% for the binary classification tasks, which is roughly in line with similar models in the literature (see 80% for the binary classification tasks, which is roughly in line with similar models in the literature (see
main manuscript for examples). The estimation procedure itself followed these steps: main manuscript for examples). The estimation procedure itself followed these steps:

- 1. Choose a sample size N_s . In our simulation, these ranged from 100 to 15,000.
- 2. Generate a single random variate k_{abn} from a binomial distribution with $n = N_s$ and $p = p_{abn}$, where p_{abs} is the expected prevalence of abnormal images suggestive of TB (in our case, this was 7%). where p_{abn is} the expected prevalence of abnormal images suggestive of TB (in our case, this was
7%).
Generate random variates for true positives (TP), false positives (FP), false negatives (FN), and
- ر
Gene
true true negatives (TN) with a single draw from a multinomial distribution, where $n = N_s$ and p_k is the probability of success for each category k as determined by the prespecified levels p_{se} and p_{sp} of sensitivity and specificity, respectively. Explicitly, these are: sensitivity and specificity, respectively. Explicitly, $P_{TP} = p_{\text{se}} * p_{\text{abn}}$

$$
P_{TP} = p_{se} * p_{abn}
$$

\n
$$
p_{TN} = p_{sp} * (1 - p_{abn})
$$

\n
$$
p_{FP} = 1 - p_{TP}
$$

\n
$$
p_{FN} = 1 - p_{TN}
$$

- 4. Repeat steps 1 and 2 a large number of times (in our case 1,000) and record the full spectrum of performance metrics based on the resulting confusion matrices.
4. Record the 2.5th and 97.5th percentiles for each metric, and take the half-width of the resulting
- performance and the resulting conditions and take
Record the 2.5th and 97.5th percentiles for each metric, and take
interval as the statistical precision in estimating the metric at the 5. Record the 2.5"" and 97.5"" percentiles for each metric, and take the half-width of the resulting
interval as the statistical precision in estimating the metric at the given sample size $N_{\rm s}$. interval as the statistical precision in estimating the metric at the given sample size Ns.

for the state sample size by 100 and repeat steps 2 through 6 until the maximum sample size has

been exceeded.
We aimed for a minimum sample size that would allow us to calculate sensitivity and specificity to $\frac{1}{2}$ within 5%. Both our validation and test sets exceeded this minimum by a large margin with an approximate precision for each metric of 1.25%.

approximate precision for each metric of 1.25%.

We began by excluding from our datasets radiographs with corrupt DICOM files or corrupt pixel arrays.
For DICOM files without such corruptions, we then extracted their raw pixel arrays, converted them to For DICOM files without such corruptions, we then extracted their raw pixel arrays, converted them to
grayscale, and exported them to 1,024x1,024-pixel PNG files. Once the image files were saved, we ran a For Dicommed mine is then then extracted the mail of the mine correct man pixel arrays, converted them to the
grayscale, and exported them to 1,024x1,024-pixel PNG files. Once the image files were saved, we ran a
separate separate script that used optical character recognition (OCR) software (detailed in Software and
hardware below) to identify radiographs with burned-in annotations and discarded any identified as separate script that used optical character recognition (OCR) software (detailed in Software and
hardware below) to identify radiographs with burned-in annotations and discarded any identified
containing more than three wo hardware below) to identify radiographs with burned-in annotations and discarded any identified as
containing more than three words of text. This exclusion step served mainly to protect the identities
the entrants by remov the entrants by removing personally identifiable information (PII) from the images, but it also prevented
the text information from affecting the model's performance. Finally, we discarded images with mean the text information from affecting, hermally information (PI) from the images, but it also prevented.
The text information from affecting the model's performance. Finally, we discarded images with mean
pixel values (out o pixel values (out of 255) below 50 or above 215 which were either too dark or too light, respectively, for radiographic features to be visible.

$\overline{}$ and $\overline{}$ of the visible to be visible.

To improve the model's ability to generalize to unseen data, we used a custom image augmentation layer as the input layer with the following transformations and deltas:

Horizontal flips (0.5 probability); Resizing (either 0.7x, 0.85x, 1.15x, or 1.3x the original dimensions);
Changes in brightness (delta range -0.525 to 0.525); Changes in brightness (delta range -0.525 to 0.525);
Changes in contrast (delta range 0.349 to 1.346); Uppes in contrast (delta range 0.349 to 1.346);
Changes in saturation (delta range 0.382 to 1.403); a c
Changes in saturation (delta range 0.382 to 1.403
Changes in hue (delta range -0.127 to 0.127). Changes in saturation (delta range of the transporter of the type of the state of the Changes in hue (delta range -0.127 to 0.127).

The delta ranges were taken from Majkowska 2020 and did not vary during training, and the ة
r transformations were turned off during validation and testing. The (sometimes perturbed) images were
then passed to the EfficientNetV2M feature extractor for processing, after which they passed through a then passed to the EfficientNetV2M feature extractor for processing, after which they passed through a
dropout layer (p=0.5) and then a randomly initialized final dense layer with either binary or categorical then passed to the Efficient Efficient Particle in the Efficient of processing, after them they passed through
dropout layer (p=0.5) and then a randomly initialized final dense layer with either binary or categorical
cross cross-entropy loss, depending on the task, for final classification.

cross-entropy losses in the task of the task.
Model training and validation procedures

Model training and vandation procedures
We trained all models with a minibatch siz
the feature extractor at once (in larger mi We trained all models with a minibature of 2 per GPU, with all the controlled under all blocks may need to be unfrozen sequentially to prevent out-of-memory errors). We trained each model
blocks may need to be unfrozen seq the feature extractor at once (in larger minibations) of the sectral minibative minibative concerns), the block
blocks may need to be unfrozen sequentially to prevent out-of-memory errors). We trained each mod
until its AU until its AUC on the validation data began to decrease; in our case, this yielded one epoch of training for
the abnormal/normal model (Model 1), two for the abnormal-TB model (Model 2), and two for the the abnormal/normal model (Model 1), two for the abnormal-TB model (Model 2), and two for the
multilabel model (Model 3). In all cases, we used the Adam optimizer (Kingma 2014) with a fixed multilabel model (Model 3). In all cases, we used the Adam optimizer (Kingma 2014) with a fixed
learning rate of 0.001. multilabel model (Model 3). In all cases, we used the Adam optimizer (Kingma 2014) with a fixed
learning rate of 0.001. learning rate of 0.001.
Operating point selection

$\frac{1}{1}$

Operating point selection
When validation data was
predictions on the test da predictions on the test data: one that maximized Youden's J index and one that minimized the model's predictions for the model of t predictions on the test data: one that maximized Youden's J index and one that minimized the model's

error in predicting the positive samples. We calculated the first directly from the model's predicted
probabilities on the validation data, but we calculated the second by reweighting the positive samples so
that test data weighted in the validation and test sets to improve precision in estimating sensitivity). Our reweighting that we informed by the following observations:
were updated in the value of the

- 1. Sensitivity (true positive rate, or TPR) and specificity (1 minus the false positive rate, or FPR) are conditional on the true label (e.g., abnormality or the presence of a specific abnormality) and thus independent of prevalence. conditional on the true label (e.g., about the true label in the single positive rate (FPR) and fall the prevalence and predicted prevalence are equal when the false positive rate (FPR) and fall
- True prevalence and predicted pr
negative rate (FNR) are equal. 2. True prevalence and predicted prevalence are equal when the value positive rate (FPR) and false
negative rate (FNR) are equal.
3. When FPR and FNR are equal, the number of counts in the off-diagonal cells of the confusi
- negative rate (FINI) are equal.
When FPR and FNR are equal,
matrix are equal (this is the ma matrix are equal (this is the mathematical insight behind McNemar's test for the difference in
paired proportions).
- 4. The off-diagonal cells are equal to the expressions FPR $*(1-p)$ and FNR $* p$, respectively, where p is equal to true prevalence.
- 5. FNR is equal to 1 sensitivity, and FPR is equal to 1 specificity, p is equal to true prevalence.
FNR is equal to 1 – sensitivity,
The ROC curve for a given clas
- 5. FRR is equal to 1 sensitivity, and FR is equal to 1 specificity.
6. The ROC curve for a given classifier on a test set contains all poss
specificity. specificity.
6. The relative difference in prevalence for any given tuple of TPR, FPR, and p is given by the
- .
The relative
expression expression $\left|\text{FPR}*(1-p)\right| - \left|\text{FNR} * p\right| / p$, where FPR = 1 - specificity and FNR = 1 - sensitivity.

To select the second operating point, then, we simply used the FPR, TPR pairs from a model's ROC curve ا
م To select the second operating point, then, we simply accurating the pairs from a model's Processity on the validation data to calculate the difference between true prevalence and predicted prevalence
given a target true on the validation data to calculate the difference between the prevalence and predicted prevalence
given a target true prevalence p_t for each possible operating point (using the formula from observationable providents) above) and selected the one with the smallest difference.

Software and hardware $\frac{3}{2}$

All models were built with Keras using the TensorFlow 2 backend. DICOM work was done with the pydicom Python package; image preprocessing with the pytesseract, scikit-image, and NumPy packages;
data visualization with the seaborn package; and statistical analysis with custom functions for bootstrap python and python package; individual paralysis with custom functions for bootstrap
That a visualization with the seaborn package; and statistical analysis with custom functions for bootstrap
That image, all code is public resampling. All code is publicly available at <u>https://github.com/cdcai/hamlet.git</u>, and top-level scripts are
designed to be rerun on different datasets to aid replication of our results and future research efforts. designed to be rerun on different datasets to aid replication of our results and future research efforts.

Models were trained on a scientific workstation with 32 logical processors, 128GB of RAM, and 3 NVIDIA |
|
| MODELS WERE TRAINED ON A SCIENCE WORKSTATION WITH 32 LOGICAL processors, 128GB of RAM, and 3 NVIDIA proce

RTX A6000 GPUs.
Bootstrap resampling procedures $\frac{1}{1}$

To create confidence intervals for the performance metrics, we used the nonparametric bootstrap with bias correction and acceleration (BCA), following Efron 1987. Because the test set contained an even
number (4,000) of normal and abnormal images each (we up-sampled abnormal images to increase number (4,000) of normal and abnormal images each (we up-sampled abnormal images to increase
precision in estimating AUROC and sensitivity), naively constructing the intervals would yield inflated precision in estimating AUROC and sensitivity), naively constructing the intervals would yield inflated
estimates for metrics like positive predictive value (PPV), F1-score, and the relative difference in estimates for metrics like positive predictive value (PPV), F1-score, and the relative difference in
prevalence that are affected by the underlying prevalence of the abnormal images. To account for this estimates for meaning predictive predictive value (PPV), F1-score, and the relative difference in
prevalence that are affected by the underlying prevalence of the abnormal images. To account for
inflation, we constructed e prevalence that are affected that an anti-trying prevalence of the abnormal images. To account for the test set,
inflation, we constructed each bootstrap replicate by sampling n_{ab} abnormal images from the test set,
whe where n_{ab} is a random draw from a binomial distribution $B(n, p)$, n is the total number of images in the test set (8,000), and p is the proportion of abnormal images in the training data. Each replicate was then test set (8,000), and p is the proportion of abnormal images in the training data. Each replicate was the
filled out with 1 – n_{ab} normal images, also sampled from the test data with replacement, and then the filled out with $1 - n_{ab}$ normal images, also sampled from the test data with replacement, and then the filled out with $1 - \frac{1}{2}$, also sampled from the test data with replacement, and then the

replicate was used to calculate the full range of performance metrics, including sensitivity and specificity
(although these are not affected by prevalence). We used this sampling distribution to generate the bias correction parameters z0 and acceleration parameters for the metrics and construct the resulting 95% confidence intervals. As noted in the main manuscript, we did not adjust the intervals for multiplicity.

${\sf Supplemental}$ Results. As noted in the main manuscript, we did not adjust the intervals for multiplicity. As not multiplicity. As not adjust the intervals for multiplicity. As not adjust the intervals for multiplicity. As not a $\ddot{}$ Supplemental Results

Supplemental Table 1. Operating points for our internal validation set that optimized three criteria: Youden's J index (j); the absolute relative error in predicted counts (count); and the same relative but
reweighted to take the difference in prevalence for each finding between the validation set and the reweighted to take the difference in prevalence for each finding between the validation set and the
total available data into account (count_adj). We used only the first and third operating points in our total available data into account (count_adj). We used only the first and third operating points in oui
analysis. to a variable data into account (count_adj). We used only the first and the operating points in our
analysis. analysis.

֦ Supplemental Table 2. Distribution of health screenings by year and region. Now and column
totals are on the margins. totals are on the margins.

$\ddot{}$ Supplemental References

Efron B. Better bootstrap confidence intervals. Journal of the American Statistical Association. 1987 Mar
1;82(397):171-85.

Majkowska A, Mittal S, Steiner DF, Reicher JJ, McKinney SM, Duggan GE, Eswaran K, Cameron Chen PH, Liu Y, Kalidindi SR, Ding A. Chest radiograph interpretation with deep learning models: assessment with Liu Y, Amamura, Ding A. Chest radiograph interpretation with deep learning models: assessment with
radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*. 2020
Feb;294(2):421-31. radiologist-adjudicated reference standards and population-adjusted evaluation. Radiology. 2020 Feb;294(2):421-31.