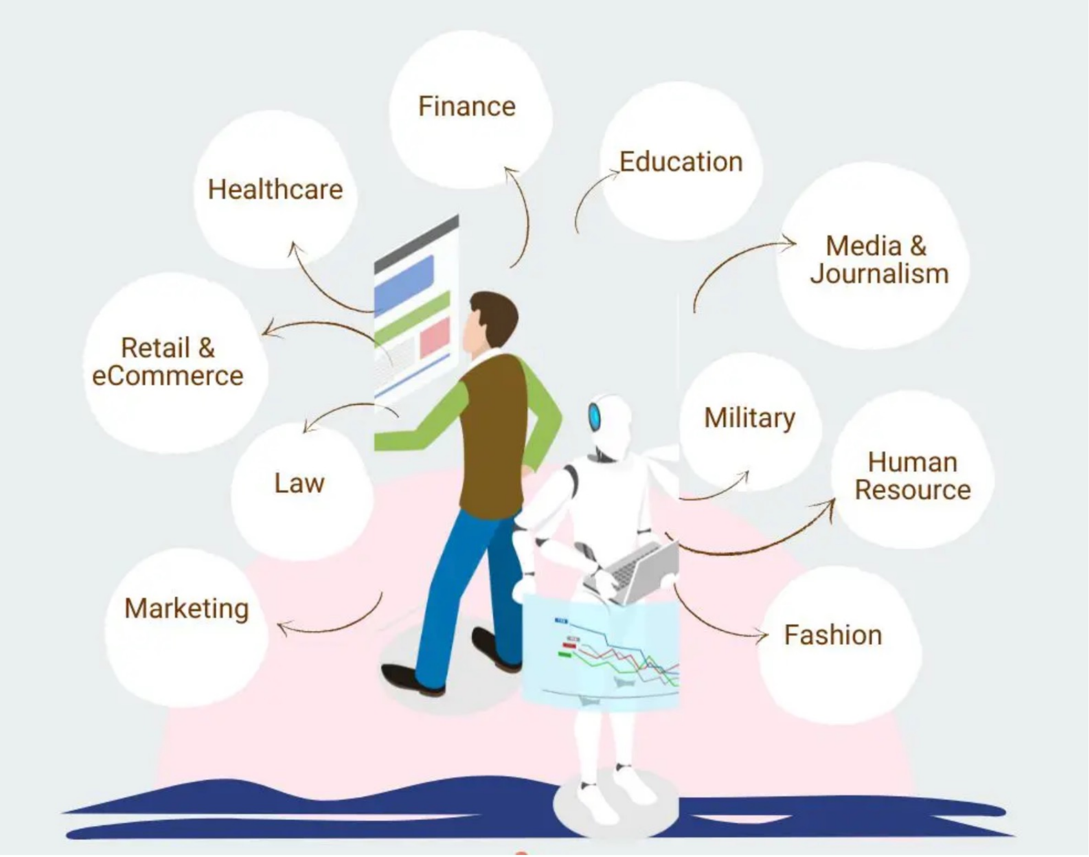
LLM for Health in Industry + A case study

BIODS 271 / CS 277

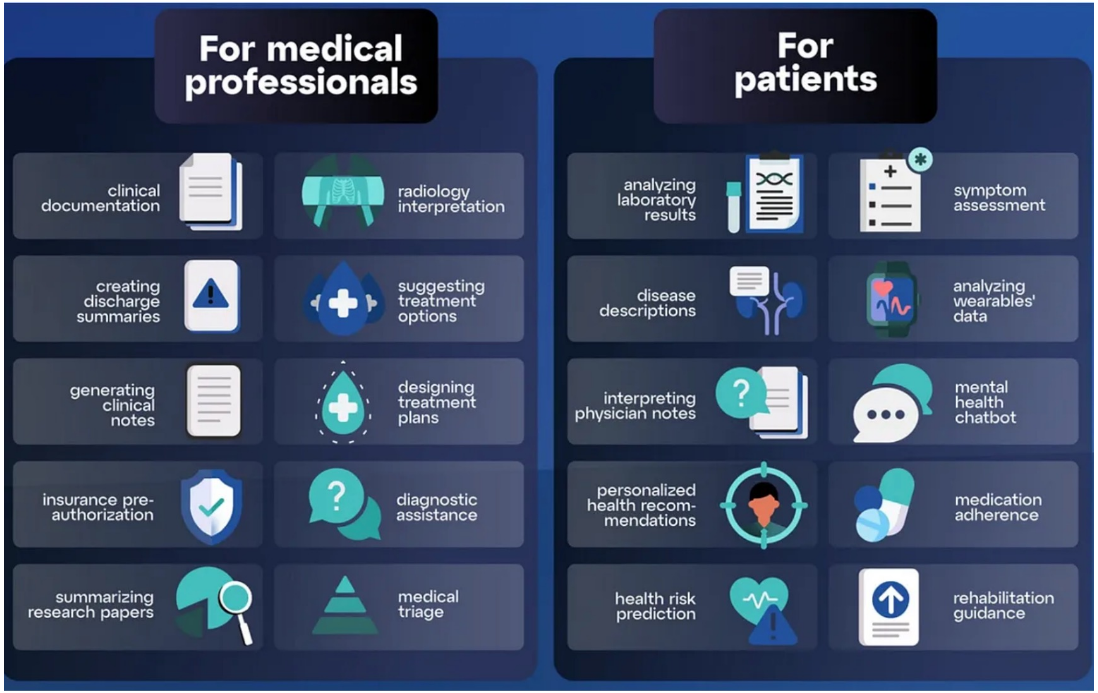
Tanveer Syeda-Mahmood

Industries where LLMs are useful



https://datasciencedojo.com/blog/llm-use-cases-top-10/

LLM for Healthcare – Patient/Provider



For medical professionals

- clinical documentation
- radiology interpretation
- creating discharge summaries
- suggesting treatment options
- generating clinical notes
- designing treatment plans
- insurance pre-authorization
- diagnostic assistance
- summarizing research papers
- medical triage

For patients

- analyzing laboratory results
- symptom assessment
- disease descriptions
- analyzing wearables' data
- interpreting physician notes
- mental health chatbot
- personalized health recommendations
- medication adherence
- health risk prediction
- rehabilitation guidance

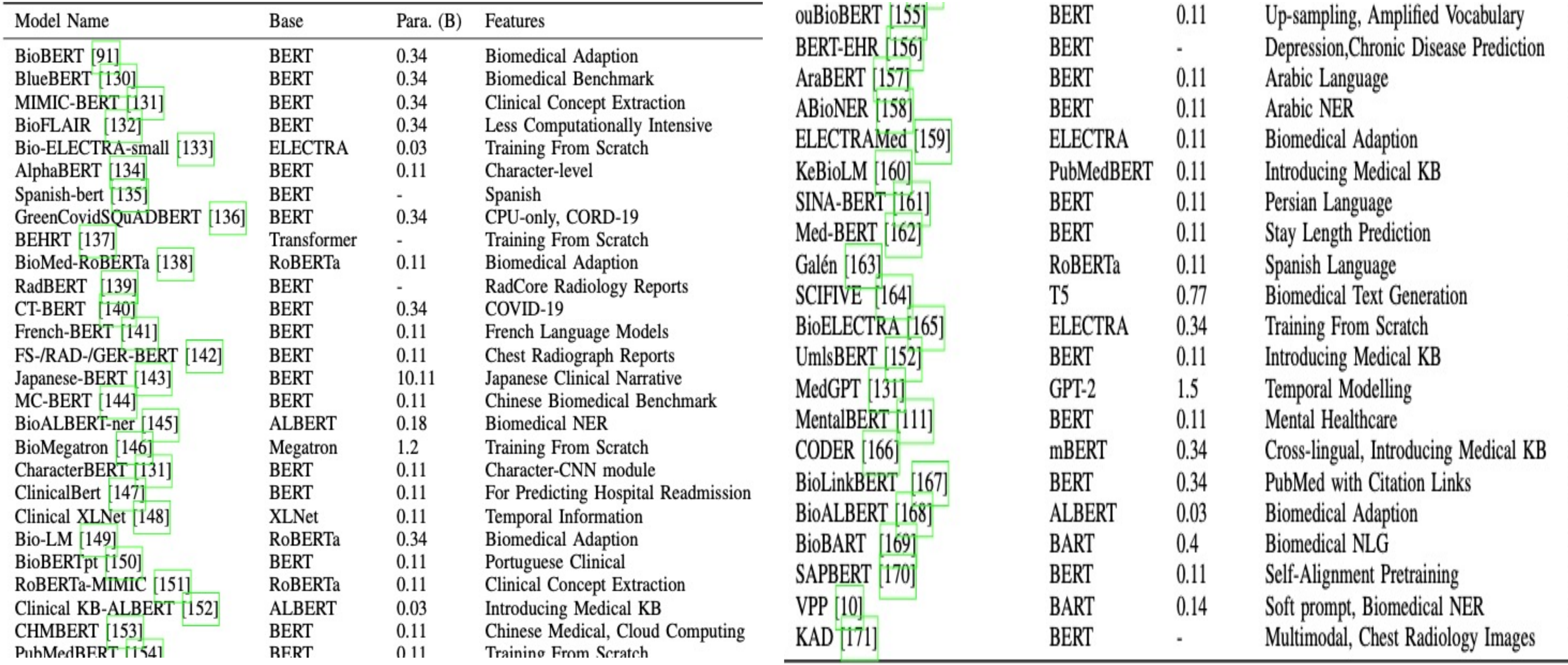
https://medium.com/@aiwizard/navigating-the-regulatory-landscape-for-large-language-models-in-healthcare-7af7f51836bd

Evolution of LLM in healthcare

- Pre-trained language models (PLM)
  - BioBERT, Clinical BERT
- LLM
  - ChatGPT
- Medical LLM
  - MedPALM, Clinical GPT, ChatDoctor, Visual Med-Alpaca
- Tasks:
  - Entity Recognition (NER), Relation Extraction (RE)
  - Text Classification (TC), Semantic Textual Similarity (STS), QA
  - Dialog systems (chatbots), report generation
  - De-identification

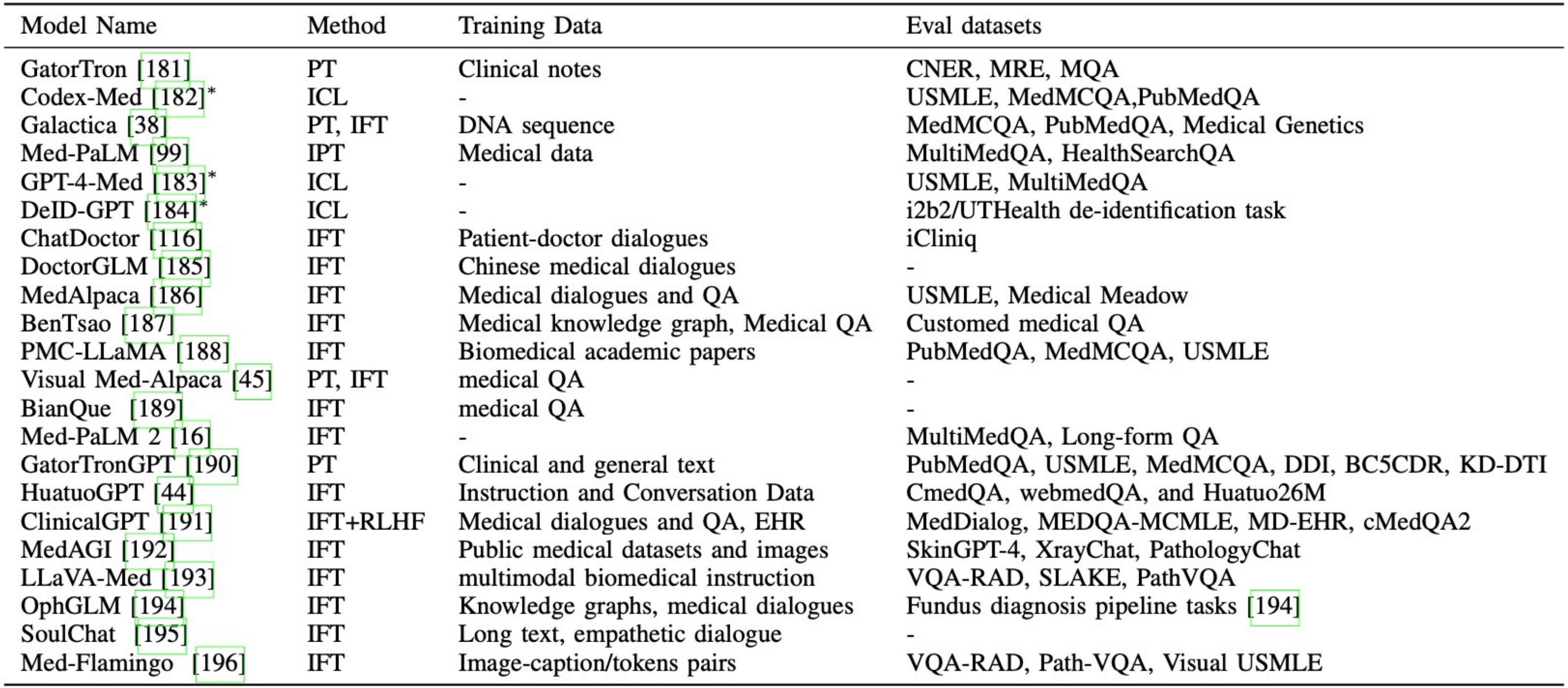
Summary of PLM models

https://arxiv.org/abs/2310.05694



Model Name	Base	Para. (B)	Features
BioBERT [91]	BERT	0.34	Biomedical Adaption
BlueBERT [130]	BERT	0.34	Biomedical Benchmark
MIMIC-BERT [131]	BERT	0.34	Clinical Concept Extraction
BioFLAIR [132]	BERT	0.34	Less Computationally Intensive
Bio-ELECTRA-small [133]	ELECTRA	0.03	Training From Scratch
AlphaBERT [134]	BERT	0.11	Character-level
Spanish-bert [135]	BERT	-	Spanish
GreenCovidSQuADBERT [136]	BERT	0.34	CPU-only, CORD-19
BEHRT [137]	Transformer	-	Training From Scratch
BioMed-RoBERTa [138]	RoBERTa	0.11	Biomedical Adaption
RadBERT [139]	BERT	-	RadCore Radiology Reports
CT-BERT [140]	BERT	0.34	COVID-19
French-BERT [141]	BERT	0.11	French Language Models
FS-/RAD-/GER-BERT [142]	BERT	0.11	Chest Radiograph Reports
Japanese-BERT [143]	BERT	10.11	Japanese Clinical Narrative
MC-BERT [144]	BERT	0.11	Chinese Biomedical Benchmark
BioALBERT-ner [145]	ALBERT	0.18	Biomedical NER
BioMegatron [146]	Megatron	1.2	Training From Scratch
CharacterBERT [131]	BERT	0.11	Character-CNN module
ClinicalBert [147]	BERT	0.11	For Predicting Hospital Readmission
Clinical XLNet [148]	XLNet	0.11	Temporal Information
Bio-LM [149]	RoBERTa	0.34	Biomedical Adaption
BioBERTpt [150]	BERT	0.11	Portuguese Clinical
RoBERTa-MIMIC [151]	RoBERTa	0.11	Clinical Concept Extraction
Clinical KB-ALBERT [152]	ALBERT	0.03	Introducing Medical KB
CHMBERT [153]	BERT	0.11	Chinese Medical, Cloud Computing
PubMedBERT [154]	BERT	0.11	Training From Scratch
ouBioBERT [155]	BERT	0.11	Up-sampling, Amplified Vocabulary
BERT-EHR [156]	BERT	-	Depression,Chronic Disease Prediction
AraBERT [157]	BERT	0.11	Arabic Language
ABioNER [158]	BERT	0.11	Arabic NER
ELECTRAMed [159]	ELECTRA	0.11	Biomedical Adaption
KeBioLM [160]	PubMedBERT	0.11	Introducing Medical KB
SINA-BERT [161]	BERT	0.11	Persian Language
Med-BERT [162]	BERT	0.11	Stay Length Prediction
Galén [163]	RoBERTa	0.11	Spanish Language
SCIFIVE [164]	T5	0.77	Biomedical Text Generation
BioELECTRA [165]	ELECTRA	0.34	Training From Scratch
UmlsBERT [152]	BERT	0.11	Introducing Medical KB
MedGPT [131]	GPT-2	1.5	Temporal Modelling
MentalBERT [111]	BERT	0.11	Mental Healthcare
CODER [166]	mBERT	0.34	Cross-lingual, Introducing Medical KB
BioLinkBERT [167]	BERT	0.34	PubMed with Citation Links
BioALBERT [168]	ALBERT	0.03	Biomedical Adaption
BioBART [169]	BART	0.4	Biomedical NLG
SAPBERT [170]	BERT	0.11	Self-Alignment Pretraining
VPP [10]	BART	0.14	Soft prompt, Biomedical NER
KAD [171]	BERT	-	Multimodal, Chest Radiology Images

Summary of LLM models



Model Name	Method	Training Data	Eval datasets
GatorTron [181]	PT	Clinical notes	CNER, MRE, MQA
Codex-Med [182]*	ICL	-	USMLE, MedMCQA,PubMedQA
Galactica [38]	PT, IFT	DNA sequence	MedMCQA, PubMedQA, Medical Genetics
Med-PaLM [99]	IPT	Medical data	MultiMedQA, HealthSearchQA
GPT-4-Med [183]*	ICL	-	USMLE, MultiMedQA
DeID-GPT [184]*	ICL	-	i2b2/UTHealth de-identification task
ChatDoctor [116]	IFT	Patient-doctor dialogues	iCliniq
DoctorGLM [185]	IFT	Chinese medical dialogues	-
MedAlpaca [186]	IFT	Medical dialogues and QA	USMLE, Medical Meadow
BenTsao [187]	IFT	Medical knowledge graph, Medical QA	Customed medical QA
PMC-LLaMA [188]	IFT	Biomedical academic papers	PubMedQA, MedMCQA, USMLE
Visual Med-Alpaca [45]	PT, IFT	medical QA	-
BianQue [189]	IFT	medical QA	-
Med-PaLM 2 [16]	IFT	-	MultiMedQA, Long-form QA
GatorTronGPT [190]	PT	Clinical and general text	PubMedQA, USMLE, MedMCQA, DDI, BC5CDR, KD-DTI
HuatuoGPT [44]	IFT	Instruction and Conversation Data	CmedQA, webmedQA, and Huatuo26M
ClinicalGPT [191]	IFT+RLHF	Medical dialogues and QA, EHR	MedDialog, MEDQA-MCMLE, MD-EHR, cMedQA2
MedAGI [192]	IFT	Public medical datasets and images	SkinGPT-4, XrayChat, PathologyChat
LLaVA-Med [193]	IFT	multimodal biomedical instruction	VQA-RAD, SLAKE, PathVQA
OphGLM [194]	IFT	Knowledge graphs, medical dialogues	Fundus diagnosis pipeline tasks [194]
SoulChat [195]	IFT	Long text, empathetic dialogue	-
Med-Flamingo [196]	IFT	Image-caption/tokens pairs	VQA-RAD, Path-VQA, Visual USMLE

Methods include: pretraining (PT), Instructional pre-training (IPT), Prompt tuning (PT), Instructional fine-tuning (IFT), Reinforcement learning human feedback (RLHF), In-context learning (ICL)

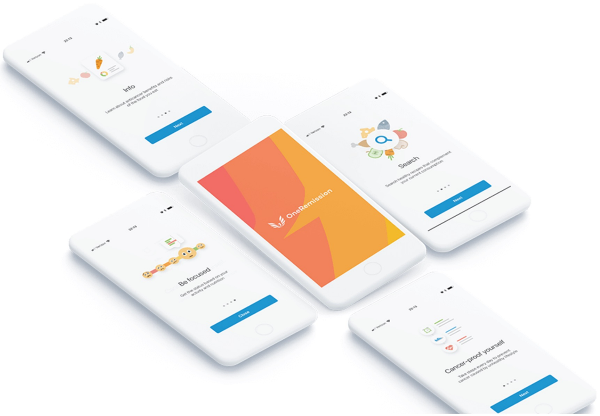
Issues with LLM approaches

- Hallucinations
  - Misinformation
- Ethical implications:
  - Biases
  - Toxicity
  - Stereotypes
- Privacy
  - Data leakage
- Performance
  - Lack of transparency
  - Accuracy and reliability
  - Data quality

Commercial LLMs for Healthcare

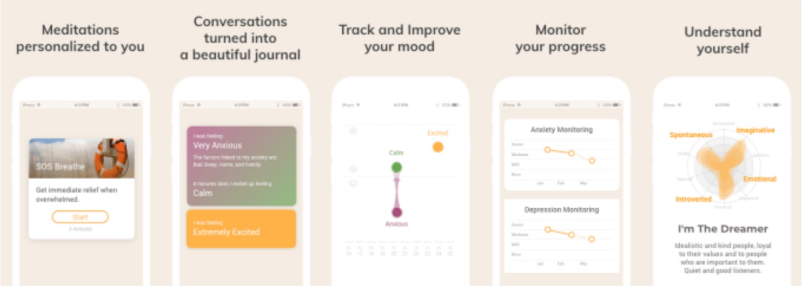
- Not many, still under development
  - Large companies
    - MedPalm2 (Google)
    - OpenAI (GPT4)
  - Mostly developed in-house by informatics teams within hospitals
  - Startups emerging in this space
    - OneRemission,

Chatbots in healthcare



OneRemission: Information for cancer patients

Youper


Meditations personalized to you

Conversations turned into a beautiful journal

Track and Improve your mood

Monitor your progress

Understand yourself

Source: https://techcrunch.com/

Information for psychological health

Woebot

Infermedica

Buoy Health

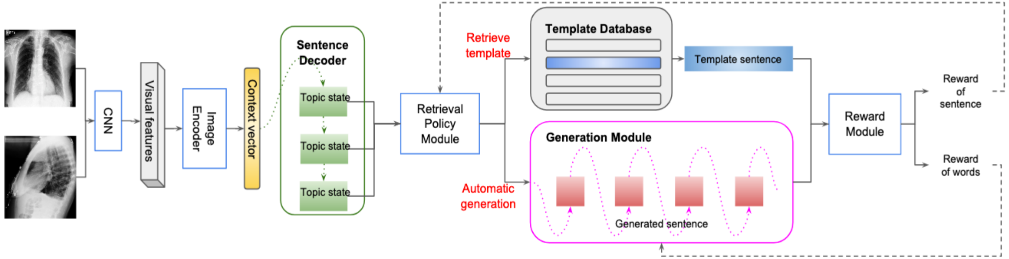
Report Generation Methods

- Recognition of findings as simple reports
- Findings as seed for language generation
- Direct input of the image in a vision language model
- Visual instructional tuning-based report generation

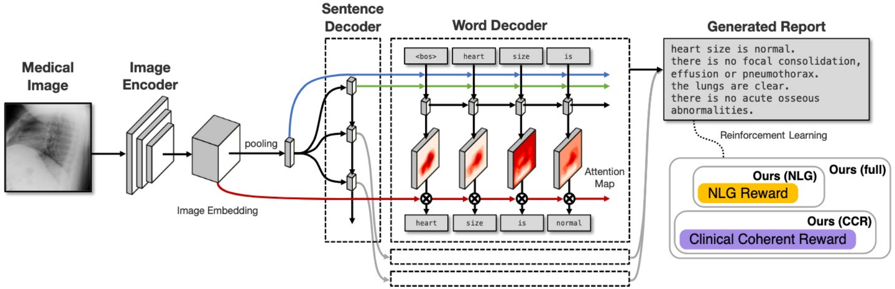
Report generation approaches

- Joint learning of images and text
  - Encoder-decoder architecture for semantic topics
  - hierarchical LSTM or RNN to generate the sentence
  - Result is not clinically accurate
- Pure template-based retrieval
  - Language is repeatable, coarse, and cookie cutter
- Hybrid approaches
  - Template sentences or a topic-based sentence generation module
- Clinically meaningful
  - Enforce constraints of +v,-ve, no mention, uncertain and reflect that in the sentence generated
  - Uses a reinforcement learning approach
- None of these approaches can ensure clinically meaningful and fine-grained description of findings in an unattended fashion

From Li,Y., Liang,X., Hu,Z., Xing,E.P.: "Hybrid retrieval-generation reinforced agent for medical image report generation," in Advances in Neural Information Processing Systems. pp. 1530– 1540 (2018)



Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, Marzyeh Ghassemi, "Clinically Accurate Chest X-Ray Report Generation," Proceedings of the 4th Machine Learning for Healthcare Conference, PMLR 106:249-269, 2019.



```
heart size is normal.
there is no focal consolidation,
effusion or pneumothorax.
the lungs are clear.
there is no acute osseous
abnormalities.
```

Limited report generation based on findings – Companies and data providers

NANOX

ARTERYS

- FDA Clearance for over 500 AI software
- Limited finding coverage
- Coarse-grained reports

Riverain TECHNOLOGIES

RADLogics™

aidoc

qure.ai

Behold.ai

Lunit

OXIPIT

DEEPTEK

annalise

TEMPUS

Large Labeled Dataset Providers

MIT Massachusetts Institute of Technology

Carnegie Mellon University

1850 MELIORA UNIVERSITY of ROCHESTER

BIMCV Medical Imaging Databank of the Valencia Region

NIH National Institutes of Health *Clinical Center*

CXR-RePAIR

- Used the reports/seeded sentences & images in encoders → focused on max similarity between img & text embeddings

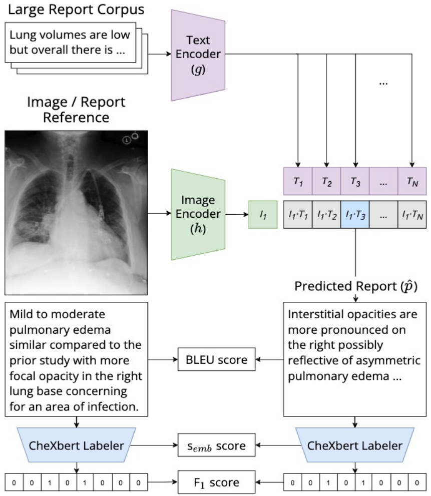
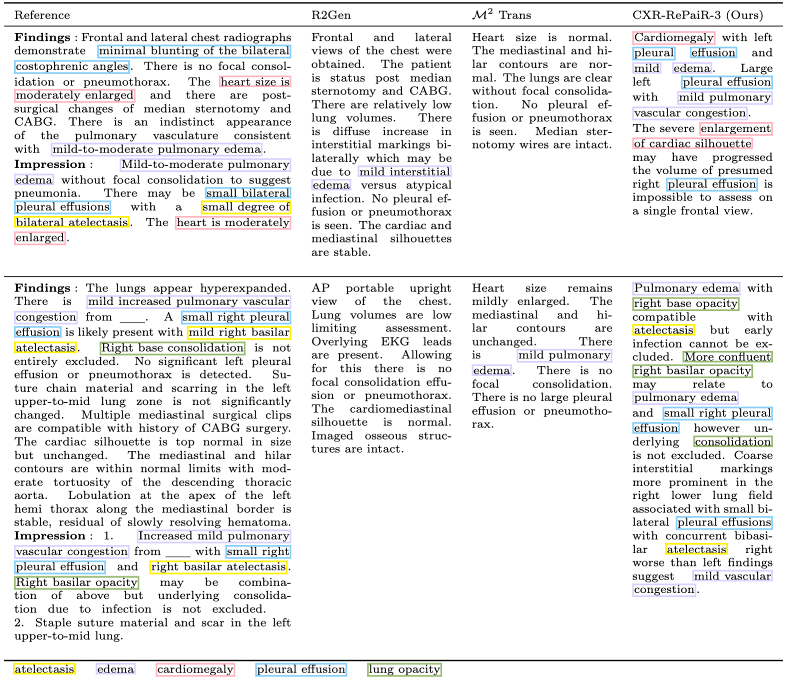

Large Report Corpus
Lung volumes are low but overall there is ...
Text Encoder (g)
...
Image / Report Reference
Image Encoder (h)
I_1
T_1
T_2
T_3
...
T_N
$I_1 \cdot T_1$
$I_1 \cdot T_2$
$I_1 \cdot T_3$
...
$I_1 \cdot T_N$
Predicted Report ($\hat{p}$)
Mild to moderate pulmonary edema similar compared to the prior study with more focal opacity in the right lung base concerning for an area of infection.
Interstitial opacities are more pronounced on the right possibly reflective of asymmetric pulmonary edema ...
BLEU score
CheXbert Labeler
s_{emb} score
CheXbert Labeler
0 0 1 0 1 0 0 0
F_1 score
0 0 1 0 1 0 0 0

Figure 1: **CXR-RePaiR approach.** Reports or report sentences from a large corpus are passed through a pre-trained text encoder, and the input chest X-ray is similarly passed through a pre-trained image encoder. We generate a prediction by selecting the report that maximizes the similarity between the text and image embeddings. The predicted and ground truth reports are then passed through a labeler and performance scores are computed.

Table 2: Examples of different methods' generated reports compared to the reference report. Correct positive predictions are color coded by finding type to improve readability. On these examples, our method does particularly well in providing clinically useful information that is accurate, clear, and actionable.



Reference	R2Gen	$\mathcal{M}^2$ Trans	CXR-RePaiR-3 (Ours)
Findings : Frontal and lateral chest radiographs demonstrate minimal blunting of the bilateral costophrenic angles. There is no focal consolidation or pneumothorax. The heart size is moderately enlarged and there are post-surgical changes of median sternotomy and CABG. There is an indistinct appearance of the pulmonary vasculature consistent with mild-to-moderate pulmonary edema. **Impression** : Mild-to-moderate pulmonary edema without focal consolidation to suggest pneumonia. There may be small bilateral pleural effusions with a small degree of bilateral atelectasis. The heart is moderately enlarged.	Frontal and lateral views of the chest were obtained. The patient is status post median sternotomy and CABG. There are relatively low lung volumes. There is diffuse increase in interstitial markings bilaterally which may be due to mild interstitial edema versus atypical infection. No pleural effusion or pneumothorax is seen. The cardiac and mediastinal silhouettes are stable.	Heart size is normal. The mediastinal and hilar contours are normal. The lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. Median sternotomy wires are intact.	Cardiomegaly with left pleural effusion and mild edema. Large left pleural effusion with mild pulmonary vascular congestion. The severe enlargement of cardiac silhouette may have progressed the volume of presumed right pleural effusion is impossible to assess on a single frontal view.
Findings : The lungs appear hyperexpanded. There is mild increased pulmonary vascular congestion from ____. A small right pleural effusion is likely present with mild right basilar atelectasis. Right base consolidation is not entirely excluded. No significant left pleural effusion or pneumothorax is detected. Suture chain material and scarring in the left upper-to-mid lung zone is not significantly changed. Multiple mediastinal surgical clips are compatible with history of CABG surgery. The cardiac silhouette is top normal in size but unchanged. The mediastinal and hilar contours are within normal limits with moderate tortuosity of the descending thoracic aorta. Lobulation at the apex of the left hemi thorax along the mediastinal border is stable, residual of slowly resolving hematoma. **Impression** : 1. Increased mild pulmonary vascular congestion from ____ with small right pleural effusion and right basilar atelectasis. Right basilar opacity may be combination of above but underlying consolidation due to infection is not excluded. 2. Staple suture material and scar in the left upper-to-mid lung.	AP portable upright view of the chest. Lung volumes are low limiting assessment. Overlying EKG leads are present. Allowing for this there is no focal consolidation effusion or pneumothorax. The cardiomediastinal silhouette is normal. Imaged osseous structures are intact.	Heart size remains mildly enlarged. The mediastinal and hilar contours are unchanged. There is mild pulmonary edema. There is no focal consolidation. There is no large pleural effusion or pneumothorax.	Pulmonary edema with right base opacity compatible with atelectasis but early infection cannot be excluded. More confluent right basilar opacity may relate to pulmonary edema and small right pleural effusion however underlying consolidation is not excluded. Coarse interstitial markings more prominent in the right lower lung field associated with small bilateral pleural effusions with concurrent bibasilar atelectasis right worse than left findings suggest mild vascular congestion.

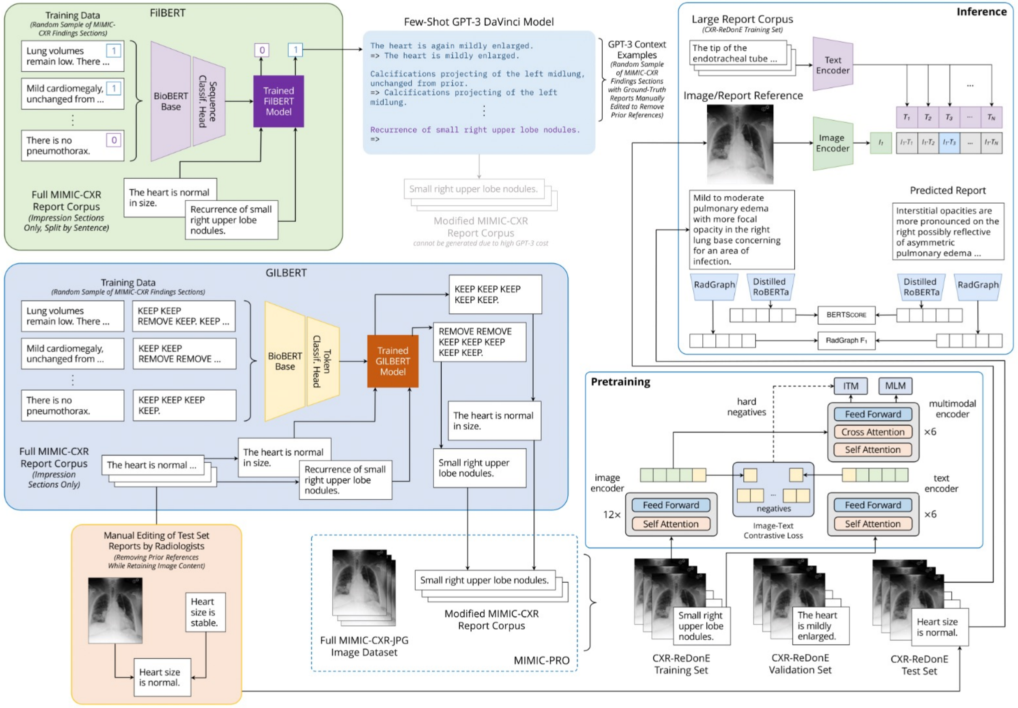
atelectasis   edema   cardiomegaly   pleural effusion   lung opacity

Endo, M., Krishnan, R., Krishna, V., Ng, A. Y., & Rajpurkar, P. (2021, November 28). *Retrieval-based chest X-ray report generation using a pre-trained contrastive language-image model*. PMLR. https://proceedings.mlr.press/v158/endo21a.html

CXR-ReDonE

- Used few-shot approach to rewrite reports (using GILBERT, GPT-3)
  - BioBERT token classification to remove words in priors
  - modified MIMIC-CXR files to MIMIC-PRO
- trained ReDonE to use contrastive learning & output similar reports

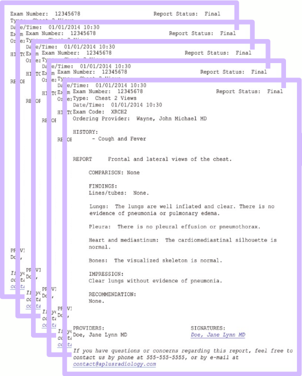
Ramesh, V., Chi, N. A., & Rajpurkar, P. (2022, October 13). *Improving radiology report generation systems by removing hallucinated references to non-existent priors*. arXiv.org. https://arxiv.org/abs/2210.06340



FilBERT
Training Data
(Random Sample of MIMIC-CXR Findings Sections)
Lung volumes remain low. There ...
1
Mild cardiomegaly, unchanged from ...
1
There is no pneumothorax.
0
BioBERT Base
Sequence Classif. Head
Trained FilBERT Model
0
1
Full MIMIC-CXR Report Corpus
(Impression Sections Only, Split by Sentence)
The heart is normal in size.
Recurrence of small right upper lobe nodules.
Few-Shot GPT-3 DaVinci Model
The heart is again mildly enlarged.
=> The heart is mildly enlarged.
Calcifications projecting of the left midlung, unchanged from prior.
=> Calcifications projecting of the left midlung.
Recurrence of small right upper lobe nodules.
=>
GPT-3 Context Examples
(Random Sample of MIMIC-CXR Findings Sections with Ground-Truth Reports Manually Edited to Remove Prior References)
Small right upper lobe nodules.
Modified MIMIC-CXR Report Corpus
cannot be generated due to high GPT-3 cost
GILBERT
Training Data
(Random Sample of MIMIC-CXR Findings Sections)
Lung volumes remain low. There ...
KEEP KEEP REMOVE KEEP. KEEP ...
Mild cardiomegaly, unchanged from ...
KEEP KEEP REMOVE REMOVE ...
There is no pneumothorax.
KEEP KEEP KEEP KEEP.
BioBERT Base
Token Classif. Head
Trained GILBERT Model
KEEP KEEP KEEP KEEP KEEP.
REMOVE REMOVE KEEP KEEP KEEP KEEP KEEP.
Full MIMIC-CXR Report Corpus
(Impression Sections Only)
The heart is normal ...
The heart is normal in size.
Recurrence of small right upper lobe nodules.
The heart is normal in size.
Small right upper lobe nodules.
Manual Editing of Test Set Reports by Radiologists
(Removing Prior References While Retaining Image Content)
Heart size is stable.
Heart size is normal.
Full MIMIC-CXR-JPG Image Dataset
Small right upper lobe nodules.
Modified MIMIC-CXR Report Corpus
MIMIC-PRO
Inference
Large Report Corpus
(CXR-ReDonE Training Set)
The tip of the endotracheal tube ...
Text Encoder
Image/Report Reference
Image Encoder
T_1 T_2 T_3 ... T_N
I_1
$I_1{\cdot}T_1$ $I_1{\cdot}T_2$ $I_1{\cdot}T_3$ ... $I_1{\cdot}T_N$
Mild to moderate pulmonary edema with more focal opacity in the right lung base concerning for an area of infection.
Predicted Report
Interstitial opacities are more pronounced on the right possibly reflective of asymmetric pulmonary edema ...
RadGraph
Distilled RoBERTa
Distilled RoBERTa
RadGraph
BERTSCORE
RadGraph F_1
Pretraining
ITM
MLM
hard negatives
multimodal encoder
Feed Forward
Cross Attention
Self Attention
×6
image encoder
12×
Feed Forward
Self Attention
negatives
Image-Text Contrastive Loss
text encoder
Feed Forward
Self Attention
×6
Small right upper lobe nodules.
CXR-ReDonE Training Set
The heart is mildly enlarged.
CXR-ReDonE Validation Set
Heart size is normal.
CXR-ReDonE Test Set

Figure 1: **CXR-ReDonE pipeline.** We first generate MIMIC-PRO by passing reports from MIMIC-CXR through GILBERT. It should be noted that we also investigate a secondary pathway to synthesize MIMIC-PRO—the two-step pipeline FilBERT+GPT-3—but do not employ it due to its decreased accuracy and higher cost. We then train CXR-ReDonE by passing reports and chest X-rays from MIMIC-PRO through a text encoder and image encoder, respectively. Finally, CXR-ReDonE outputs the generated report with the highest dot-product similarity between the text and image embeddings, and performance metrics are calculated by comparing the ground truth to the predicted reports.

Automatic Report Generation – RAG approach



Exam Number: 12345678 Report Status: Final
Type: Chest 2 Views
Date/Time: 01/01/2014 10:30
Exam Code: XRCH2
Ordering Provider: Wayne, John Michael MD

HISTORY:
- Cough and Fever

REPORT Frontal and lateral views of the chest.

COMPARISON: None

FINDINGS:
Lines/tubes: None.

Lungs: The lungs are well inflated and clear. There is no evidence of pneumonia or pulmonary edema.

Pleura: There is no pleural effusion or pneumothorax.

Heart and mediastinum: The cardiomediastinal silhouette is normal.

Bones: The visualized skeleton is normal.

IMPRESSION:
Clear lungs without evidence of pneumonia.

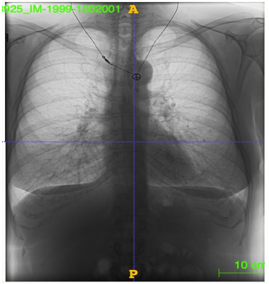
RECOMMENDATION:
None.

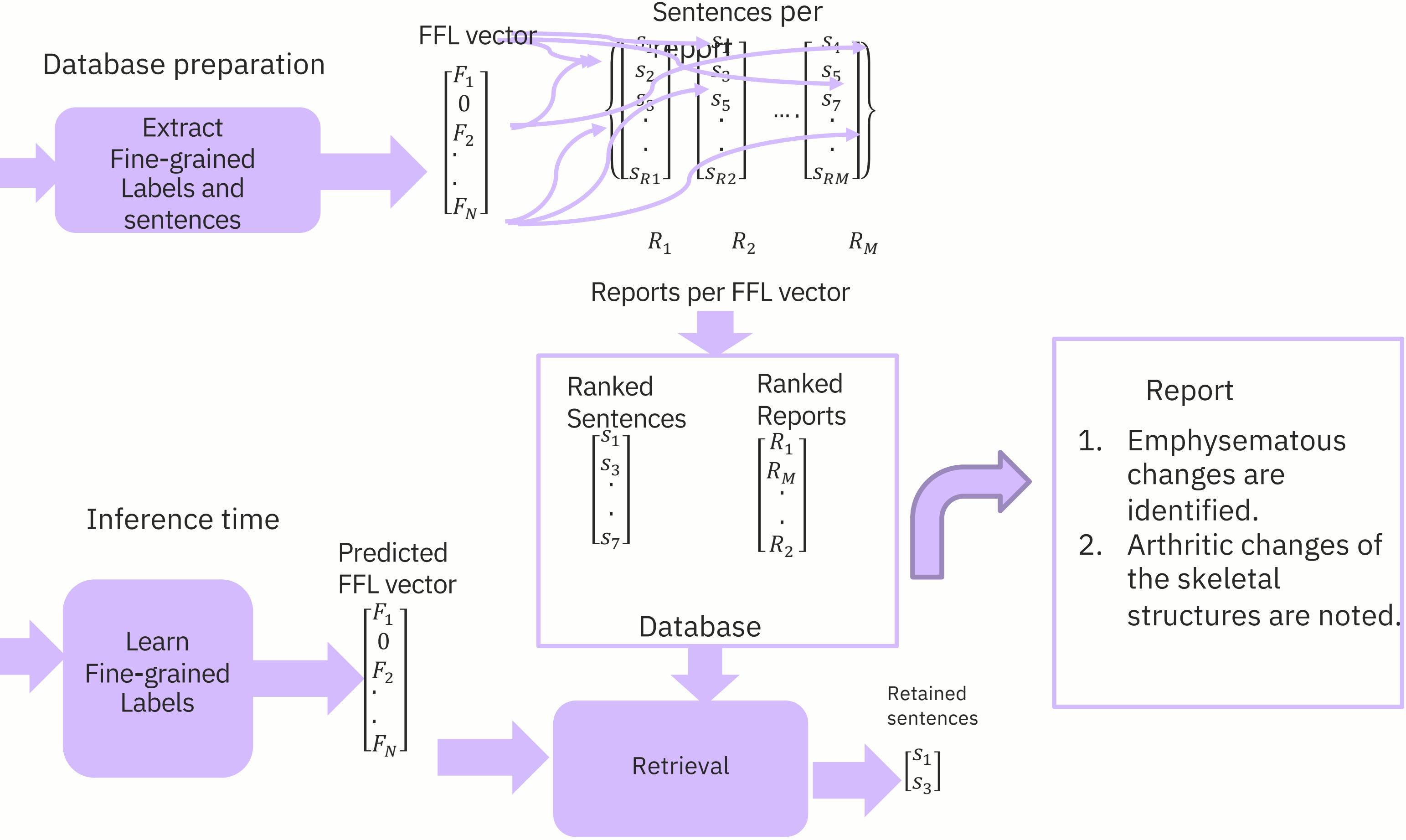
PROVIDERS: Doe, Jane Lynn MD
SIGNATURES: Doe, Jane Lynn MD

If you have questions or concerns regarding this report, feel free to contact us by phone at 555-555-5555, or by e-mail at contact@aplusradiology.com

Prior reports

Given chest X-ray





Database preparation

Extract Fine-grained Labels and sentences

FFL vector

$$\begin{bmatrix} F_1 \\ 0 \\ F_2 \\ . \\ . \\ F_N \end{bmatrix}$$

Sentences per report

$$\left\{ \begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ . \\ . \\ s_{R1} \end{bmatrix} \begin{bmatrix} s_1 \\ s_3 \\ s_5 \\ . \\ . \\ s_{R2} \end{bmatrix} \cdots \begin{bmatrix} s_4 \\ s_5 \\ s_7 \\ . \\ . \\ s_{RM} \end{bmatrix} \right\}$$

R_1 R_2 R_M

Reports per FFL vector

Ranked Sentences

$$\begin{bmatrix} s_1 \\ s_3 \\ . \\ . \\ s_7 \end{bmatrix}$$

Ranked Reports

$$\begin{bmatrix} R_1 \\ R_M \\ . \\ . \\ R_2 \end{bmatrix}$$

Database

Inference time

Learn Fine-grained Labels

Predicted FFL vector

$$\begin{bmatrix} F_1 \\ 0 \\ F_2 \\ . \\ . \\ F_N \end{bmatrix}$$

Retrieval

Retained sentences

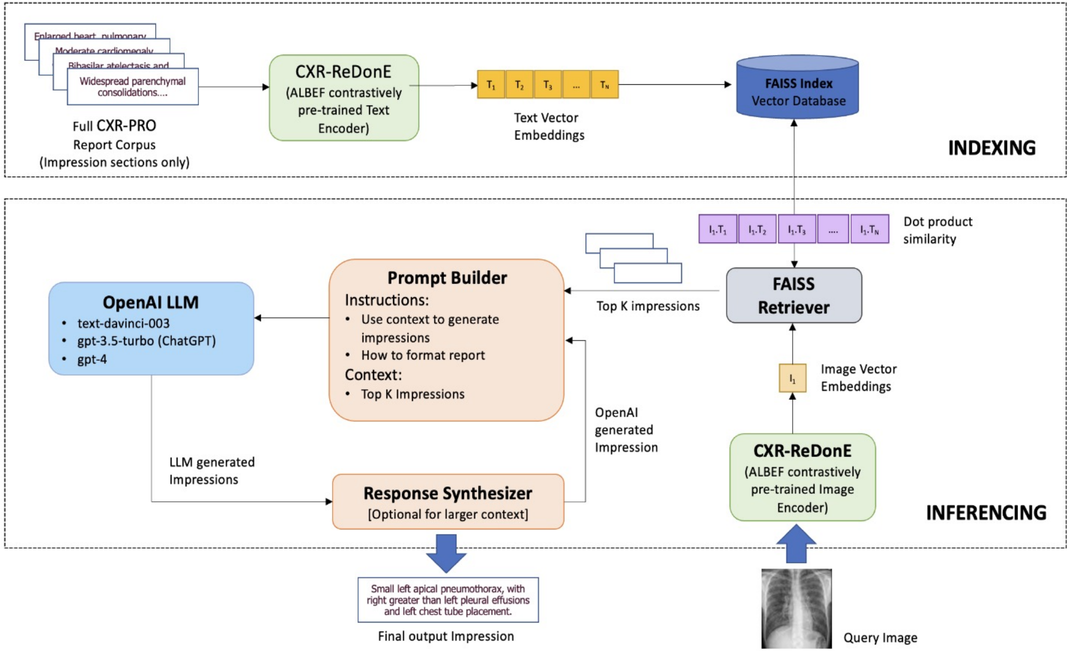
$$\begin{bmatrix} s_1 \\ s_3 \end{bmatrix}$$

Report

1. Emphysematous changes are identified.
2. Arthritic changes of the skeletal structures are noted.

RAG in CXR-RePaiR-Gen

- Working off CXR-DonE, using impression reports
- Designing & formatting prompts based on context
- RAG bridges knowledge gaps in healthcare
- Used text-davinci-003, GPT-3.5-turbo, GPT-4
    - Worked off CXR-ReDonE, CXR-RePaiR
- Contrastive X-ray-Report Pair Retrieval based Generation (CXR-RePaiR-Gen)

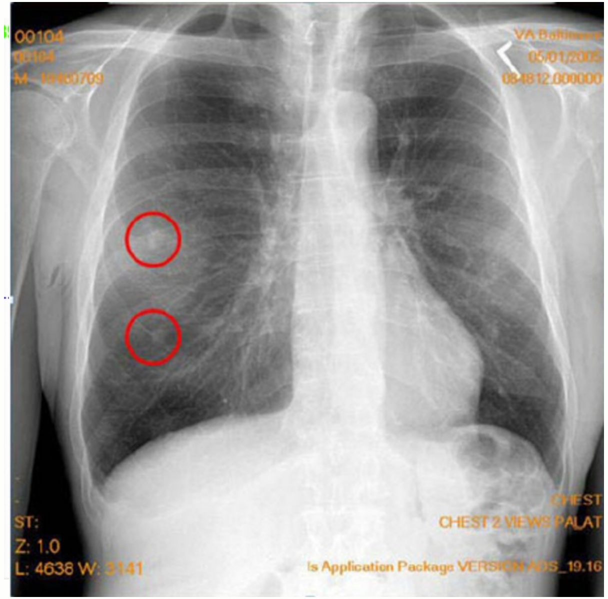

Enlarged heart, pulmonary
Moderate cardiomegaly
Bibasilar atelectasis and
Widespread parenchymal consolidations....
Full CXR-PRO Report Corpus (Impression sections only)
CXR-ReDonE (ALBEF contrastively pre-trained Text Encoder)
T1 T2 T3 ... TN
Text Vector Embeddings
FAISS Index Vector Database
INDEXING
I1.T1 I1.T2 I1.T3 .... I1.TN
Dot product similarity
FAISS Retriever
Top K impressions
Prompt Builder
Instructions:
Use context to generate impressions
How to format report
Context:
Top K Impressions
OpenAI LLM
text-davinci-003
gpt-3.5-turbo (ChatGPT)
gpt-4
Image Vector Embeddings
I1
OpenAI generated Impression
CXR-ReDonE (ALBEF contrastively pre-trained Image Encoder)
LLM generated Impressions
Response Synthesizer [Optional for larger context]
INFERENCING
Small left apical pneumothorax, with right greater than left pleural effusions and left chest tube placement.
Final output Impression
Query Image

Ranjit, M., Ganapathy, G., Manuel, R., & Ganu, T. (2023, May 5). *Retrieval augmented chest X-ray report generation using openai GPT models*. arXiv.org. https://arxiv.org/abs/2305.03660

Figure 1: We project all the text embeddings of sentences from radiology impression using a contrastively pretrained vision-language encoder (CXR-ReDonE) to a vector database index and retrieve the most matching sentences for an input image embedding using the same encoder model. The retrieved impression reports or sentences form the context of the prompt to the LLM along with instructions to generate the impression.

Chest X-ray Reporting – A case study

Chest X-ray



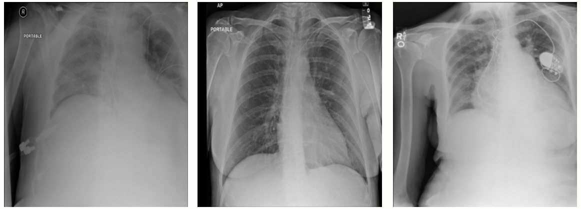
Automated Preliminary Read

There are atherosclerotic changes of the aorta.
There are calcified right hilar and mediastinal lymph nodes.
Arthritic changes of the skeletal structures are noted.

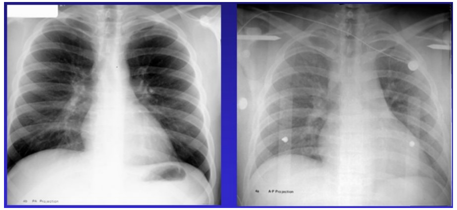
- ❖ Can AI produce an automated report?
- ❖ Can AI be as accurate as the radiologists?

How does a radiologist interpret chest X-rays?

Technical Assessment

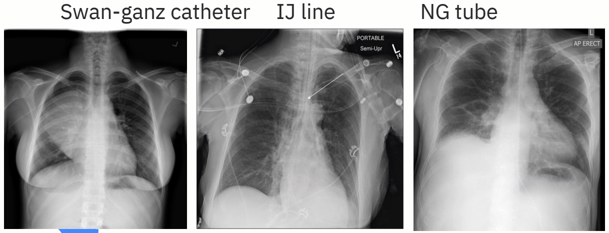


Viewpoint and position assessment



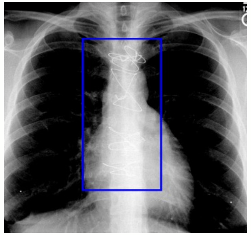
Frontal or lateral?
AP, PA or AP portable?

Any lines and tubes?

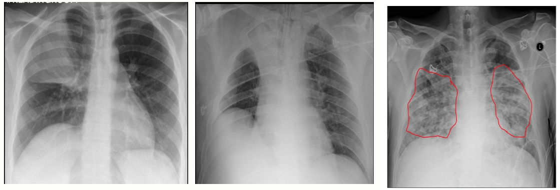


Are they properly positioned?

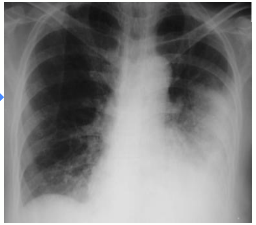
Any devices or artifacts?



Any anatomical abnormalities?



Any disease?



Generated report

Lung Findings : Lungs are clear.
No evidence of pleural effusion, pneumothorax or pulmonary edema.

Mediastinum findings :
Cardio-mediastinal silhouette is normal.

Impression: No Consolidation
No Pleural Effusion
No Pneumothorax
No pulmonary edema.

Problems addressed

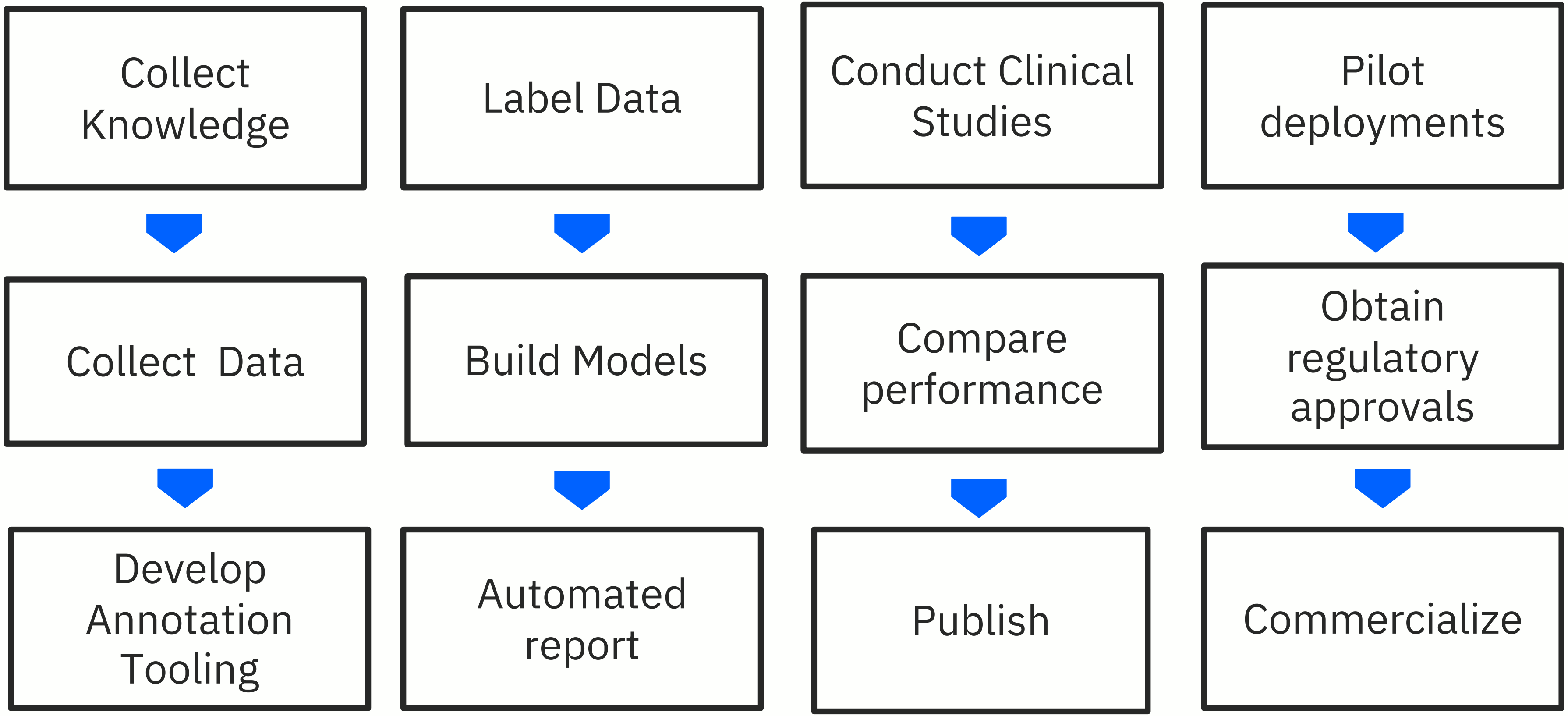
Radiologists

- → Catalog all findings in chest X-rays
- → Gather benchmark dataset
- → Establish ground truth
- → Record radiologists reads
- → Establish evaluation metrics
- → Benchmark radiologists' performance

Machine

- → Assemble training datasets
- → Label datasets
- → Build machine learning models
- → Record machine reads
- → Compare performance

Data Science Problems



Collect Knowledge

Collect Data

Develop Annotation Tooling

Label Data

Build Models

Automated report

Conduct Clinical Studies

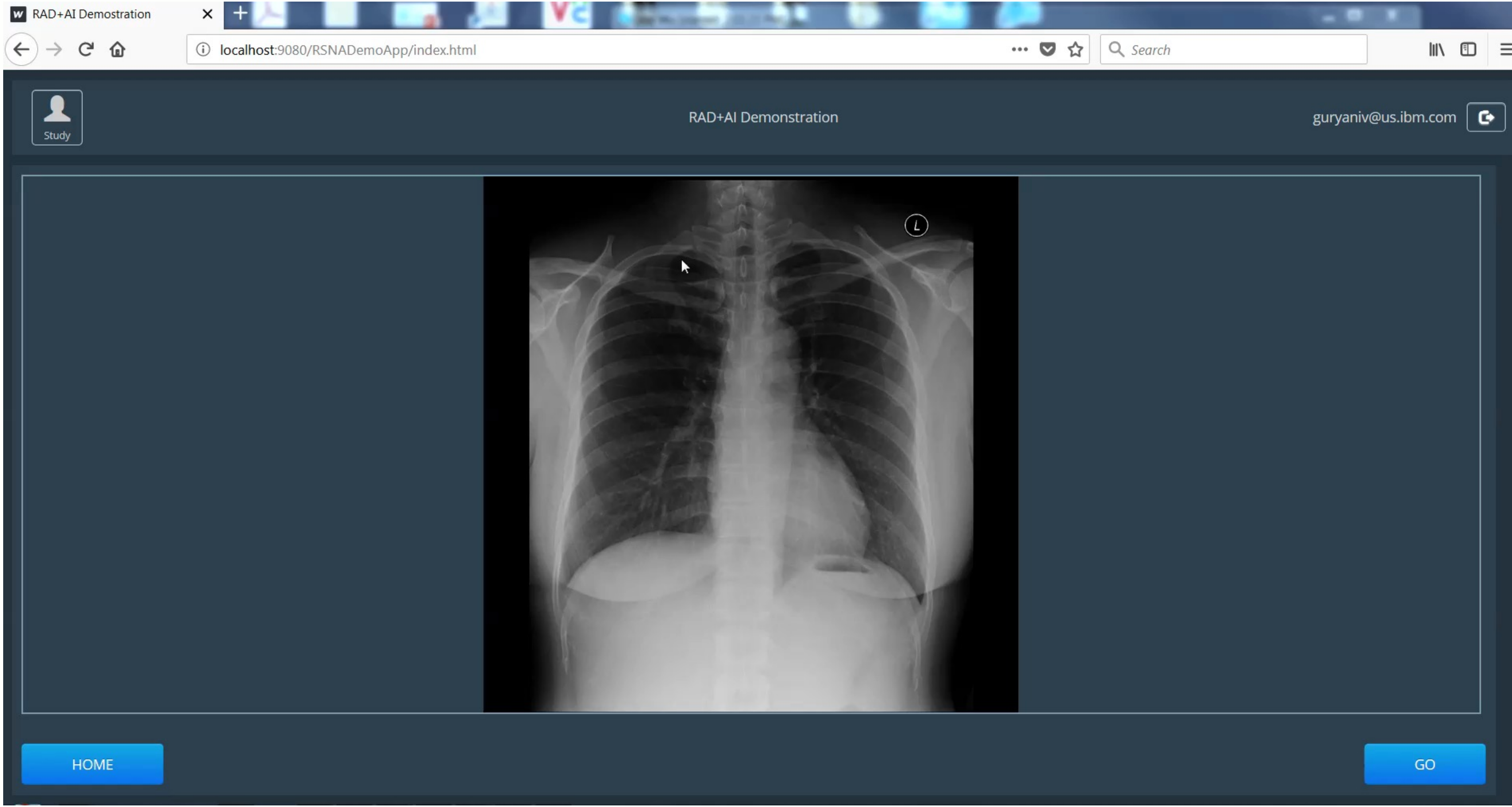
Compare performance

Publish

Pilot deployments

Obtain regulatory approvals

Commercialize


RAD+AI Demostration
localhost:9080/RSNADemoApp/index.html
Search
Study
RAD+AI Demonstration
guryaniv@us.ibm.com
L
HOME
GO

Knowledge Curation: Cataloging all possible findings in chest X-rays

Largest assembly of chest X-ray findings!

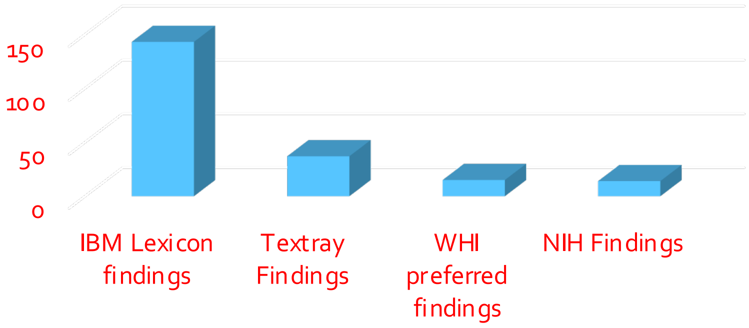
237 discrete findings

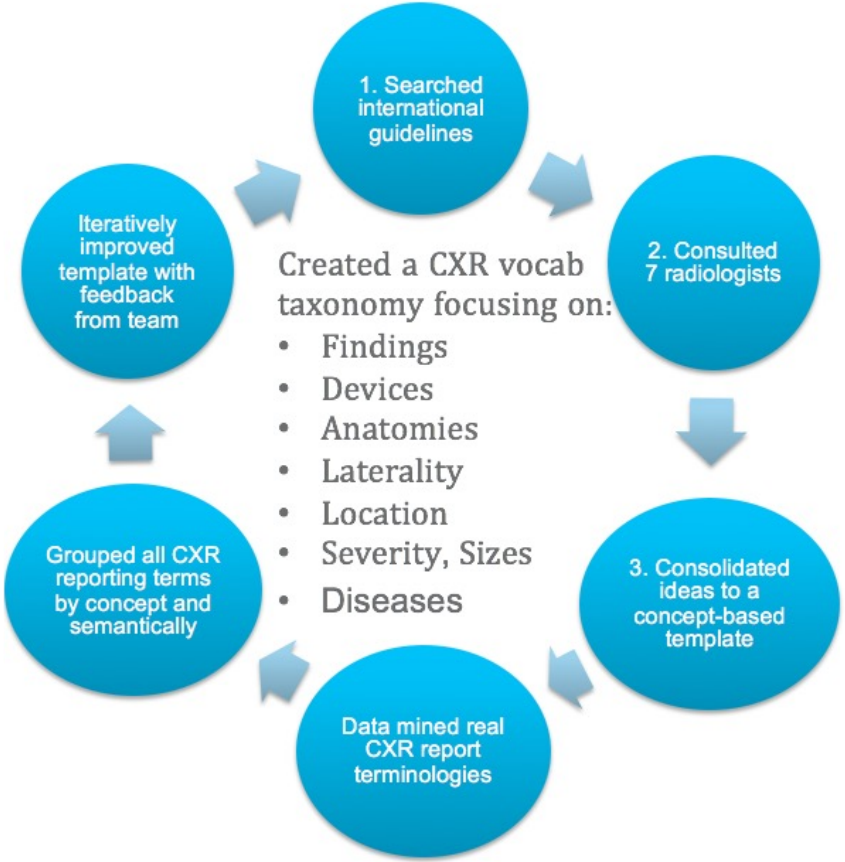
(99 anatomical findings, 79 disease, 26 technical assessment, 22 tubes and lines/finding, 7 device, 4 views)

Validated from

- Textbooks
- Fleischner guidelines
- UMLS
- Radiology education
- Radiologists board
- Over 200,000 radiology reports

Comparison of Findings





1. Searched international guidelines

2. Consulted 7 radiologists

3. Consolidated ideas to a concept-based template

Data mined real CXR report terminologies

Grouped all CXR reporting terms by concept and semantically

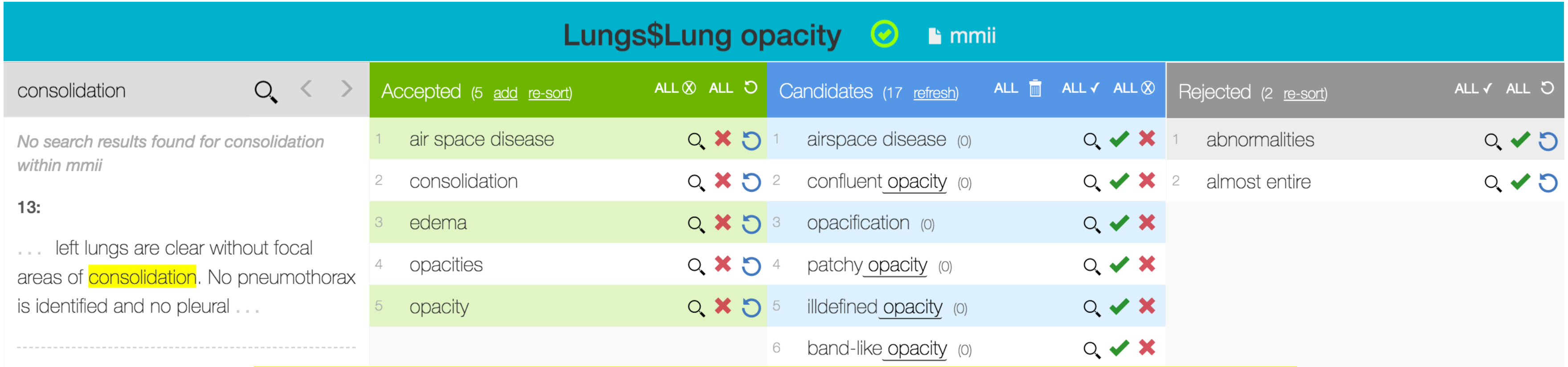
Iteratively improved template with feedback from team

Created a CXR vocab taxonomy focusing on:

- Findings
- Devices
- Anatomies
- Laterality
- Location
- Severity, Sizes
- Diseases

IBM DLA Tool Accelerated Bottom-up Vocabulary Curation

- Step 1: Start from a broad concept – e.g. lung opacity



Lungs$Lung opacity ✓ mmii

consolidation

No search results found for consolidation within mmii

13:

. . . left lungs are clear without focal areas of consolidation. No pneumothorax is identified and no pleural . . .

Accepted (5 add re-sort) ALL ⊗ ALL ↺

1	air space disease
2	consolidation
3	edema
4	opacities
5	opacity

Candidates (17 refresh) ALL 🗑 ALL ✓ ALL ⊗

1	airspace disease	(0)
2	confluent opacity	(0)
3	opacification	(0)
4	patchy opacity	(0)
5	illdefined opacity	(0)
6	band-like opacity	(0)

Rejected (2 re-sort) ALL ✓ ALL ↺

1	abnormalities
2	almost entire

[6] Anni Coden, Daniel Gruhl, Neal Lewis, et al. Spot the drug! an unsupervised pattern matching method to extract drug names from very large clinical corpora. In 2012 IEEE Second International Conference on HISB, pages 33–39. IEEE, 2012.

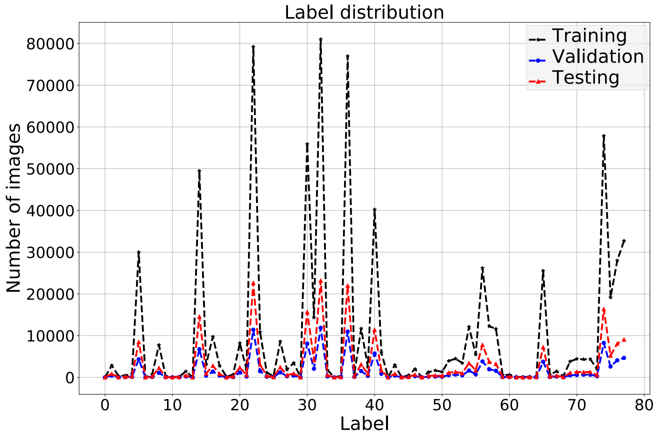
- Text Corpus: ~200,000 CXR reports (MIMIC-III) [6]
- Domain experts seed the "Accepted" terms with a few examples
- DLA tool proposes "Candidate" terms/phrases that occur in similar contexts
- Experts able to "Accept" or "Reject" candidate phrases efficiently – option to view examples in context
- Expanded to over 200 bottom-up curated lung opacity related terms/phrases in 30 minutes
- Chest X-ray lexicon - largest assembled for chest X-rays (11977 vocabulary terms, 237 lexical concepts, 78 core findings, 26 clinical categories)

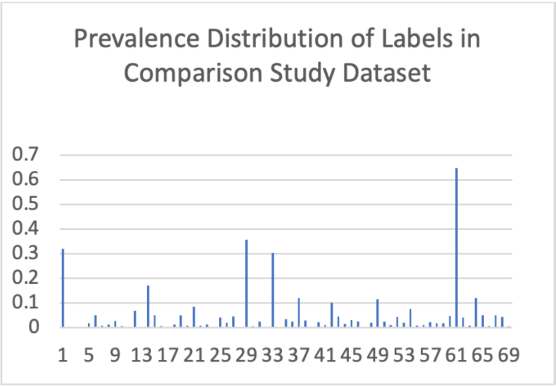
Core finding Labels

Finding Label	Finding Label	Finding Label
not otherwise specified opacity (pleural/parenchymal opacity)	mediastinal displacement	old fractures
linear/patchy atelectasis	increased reticular markings/ild pattern	subcutaneous air
pleural effusion or thickening	dislocation	elevated hemidiaphragm
normal anatomically	dilated bowel	superior mediastinal mass/enlargement
enlarged cardiac silhouette	osteotomy changes	sub-diaphragmatic air
pulmonary edema/hazy opacity	new fractures	pneumomediastinum
consolidation	shoulder osteoarthritis	cyst/bullae
not otherwise specified calcification	elevated humeral head	hydropneumothorax
pneumothorax	azygous fissure (benign)	spinal degenerative changes
lobar/segmental collapse	contrast in the gi or gu tract	calcified nodule
fracture	other internal post-surgical material	lymph node calcification
mass/nodule (not otherwise specified)	sternotomy wires	bullet/foreign bodies
hyperaeration	cardiac pacer and wires	other soft tissue abnormalities
degenerative changes	msk or spinal hardware	enteric tubes
vascular calcification	low lung volumes	incorrect placement
tortuous aorta	rotated	central intravascular lines: incorrectly positioned
multiple masses/nodules	lungs otherwise not fully included	enteric tubes: incorrectly positioned
vascular redistribution	lungs obscured by overlying object or structure	coiled/kinked/fractured
enlarged hilum	apical lordotic	tubes in the airway: incorrectly positioned
scoliosis	apical kyphotic	hernia

Training Deep Learning Models - Multi-institutional Datasets

- Datasets came unlabeled and different incidence rates
  - NIH Hospitals
    - 30,805 patients
    - 112, 120 images (original)
    - **No reports – re-read ~17000 images**
  - MIMIC-CXR
    - 63,478 patients
    - 473,056 images
    - 206,754 reports
  - Indiana
    - 2964 images & reports (Benchmarking text analytics)
  - Deccan Hospital (normal/abnormal studies)
    - 10,000 images & reports



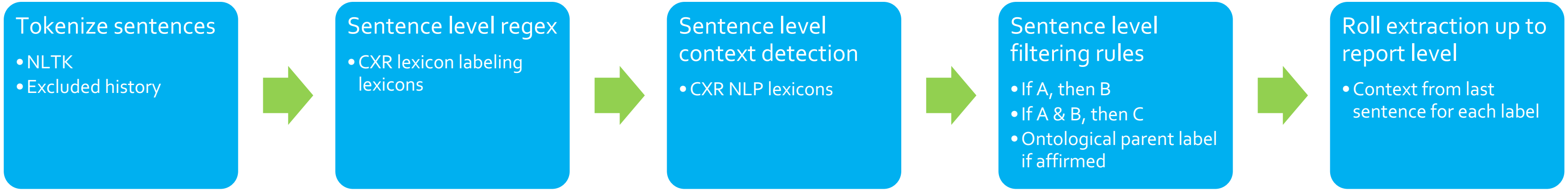


Labeling images from reports

- Manual labeling not an option for large scale image datasets
- Automated labeling methods needed from radiology reports
  - Existing algorithms for labeling have precision and recall issues due to tolerance to variations in spoken ways to describe the findings in reports
  - Our approaches:
    - Detecting coarse-grained findings
      - Vocabulary-driven concept extraction algorithm shipped in WHI products (Java)
      - A new simpler and higher precision python implementation lexical concept identification algorithm
    - Detecting fine-grained findings
      - Required natural language parse of the sentences.

Labeling process	Number of images labeled	Time taken
Manual labeling	36,554	4 months
Automatic labeling from text	587,058	4 days (7-10 days with verification)

Coarse NLP pipeline using CXR Lexicon



Tokenize sentences
- NLTK
- Excluded history

Sentence level regex
- CXR lexicon labeling lexicons

Sentence level context detection
- CXR NLP lexicons

Sentence level filtering rules
- If A, then B
- If A & B, then C
- Ontological parent label if affirmed

Roll extraction up to report level
- Context from last sentence for each label

1. History: ___M with right upper quadrant pain, nausea, vomiting, serosanguineous drainage from JP drainage → Excluded

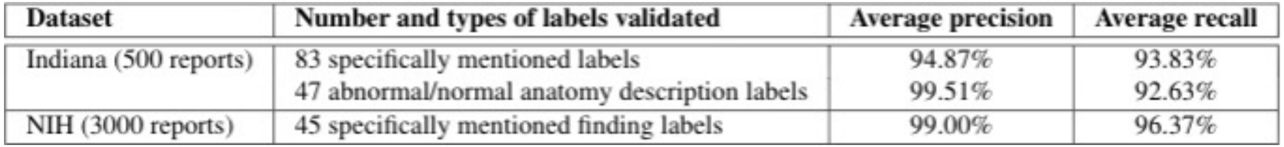
   " semantic category | context | label "

2. Heart size is enlarged. → nlp | yes | abnormal & subnatomy | yes | cardiac silhouette, anatomicalfinding | yes | enlarge cardiac silhouette
3. Lungs are clear. → nlp | yes | normal & majorstructure | yes | lungs
4. No focal consolidation, pleural effusion or pneumothorax is demonstrated.

   → anatomicalfinding | no | consolidation, anatomicalfinding | no | pleural effusion, anatomicalfinding | no | pneumothorax
5. A left subclavian PICC line is present.

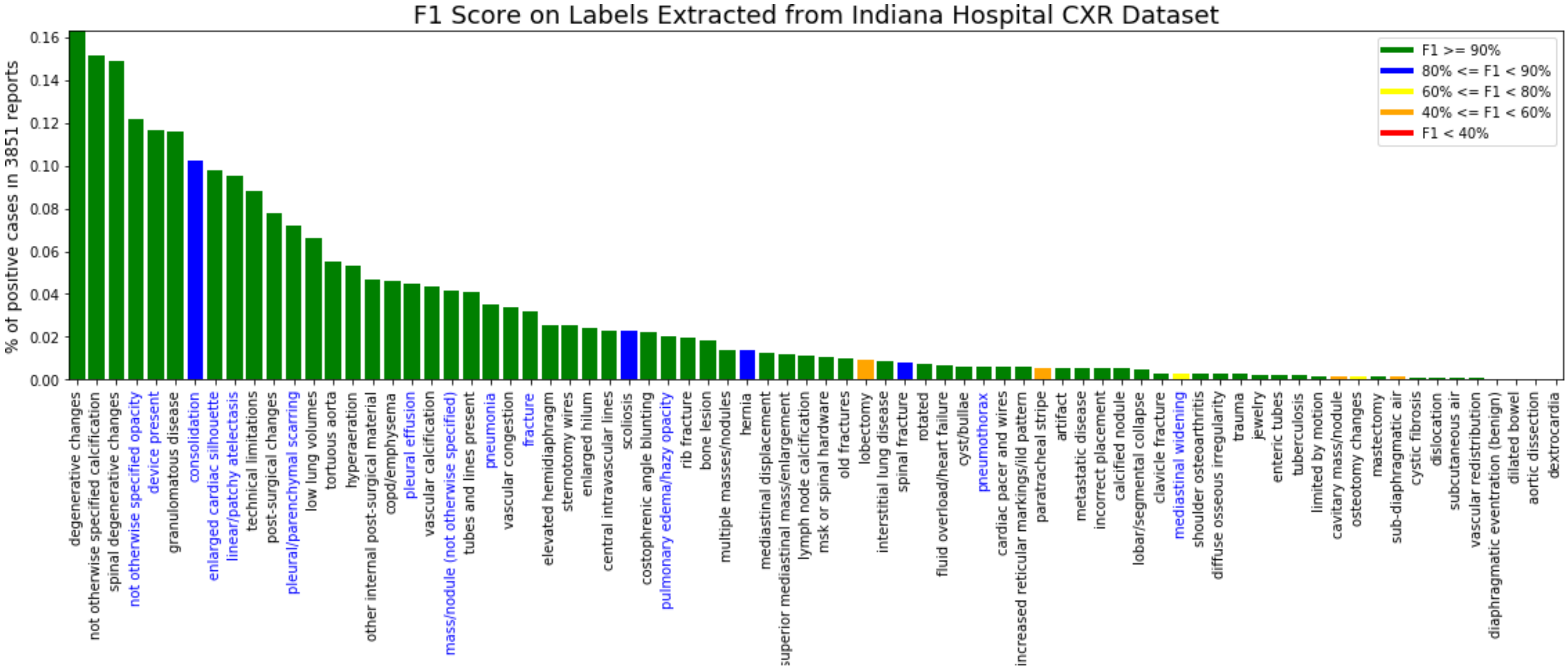
   → tubesandlines | yes | central intravascular lines, tubesandlines | yes | tubes or lines present

Evaluation on two unseen report corpus



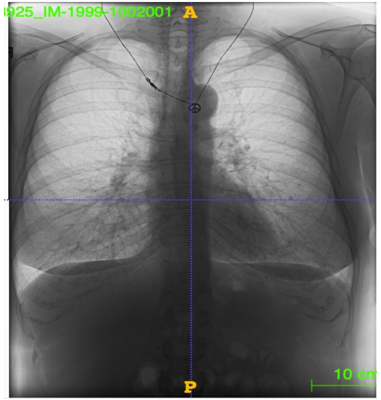
Dataset	Number and types of labels validated	Average precision	Average recall
Indiana (500 reports)	83 specifically mentioned labels 47 abnormal/normal anatomy description labels	94.87% 99.51%	93.83% 92.63%
NIH (3000 reports)	45 specifically mentioned finding labels	99.00%	96.37%

NLP performance: F1 Scores per Label



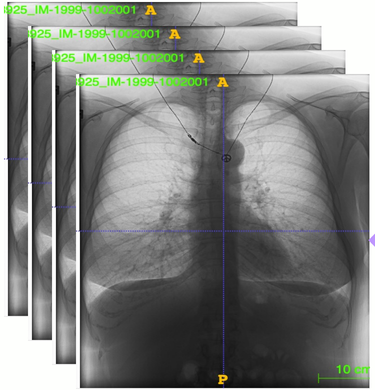
Extracting fine-grained findings

- Core findings are not sufficient for automated reporting.
- Need fine-grained descriptions
  - Anatomy affected
  - Sub-anatomy
  - Location
  - Laterality
  - Severity
  - Size
  - character
  - Shape
  - Correlation
  - Procedure
  - Measure
  - Cause
  - Symptom
  - Hedge
  - Adjectives
  - Other POS
- **How many fine-grained findings are there?**

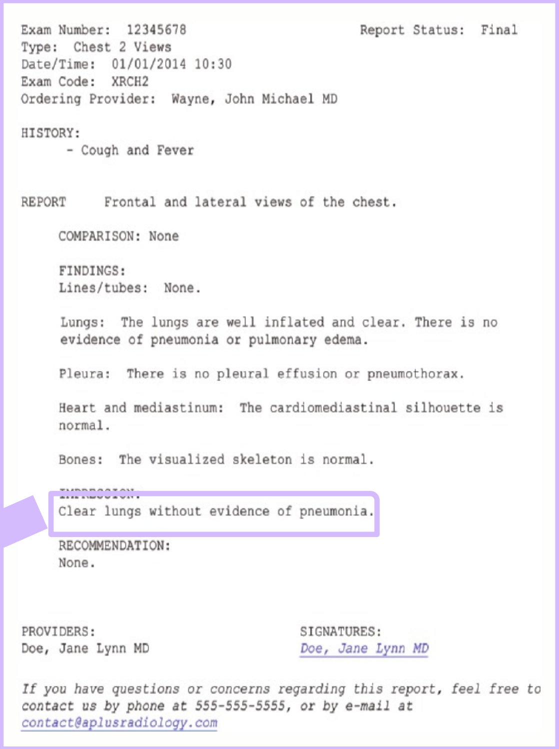


There are atherosclerotic changes of the aorta.
There are calcified right hilar and mediastinal lymph nodes.
Arthritic changes of the skeletal structures are noted.

Labeling images from reports



Associated reports



Exam Number: 12345678 Report Status: Final
Type: Chest 2 Views
Date/Time: 01/01/2014 10:30
Exam Code: XRCH2
Ordering Provider: Wayne, John Michael MD

HISTORY:
- Cough and Fever

REPORT Frontal and lateral views of the chest.

COMPARISON: None

FINDINGS:
Lines/tubes: None.

Lungs: The lungs are well inflated and clear. There is no evidence of pneumonia or pulmonary edema.

Pleura: There is no pleural effusion or pneumothorax.

Heart and mediastinum: The cardiomediastinal silhouette is normal.

Bones: The visualized skeleton is normal.

IMPRESSION:
Clear lungs without evidence of pneumonia.

RECOMMENDATION:
None.

PROVIDERS: SIGNATURES:
Doe, Jane Lynn MD Doe, Jane Lynn MD

If you have questions or concerns regarding this report, feel free to contact us by phone at 555-555-5555, or by e-mail at contact@aplusradiology.com

<u>Clear</u> <u>lungs</u> without evidence of <u>pneumonia.</u>

Anatomical finding

Anatomy

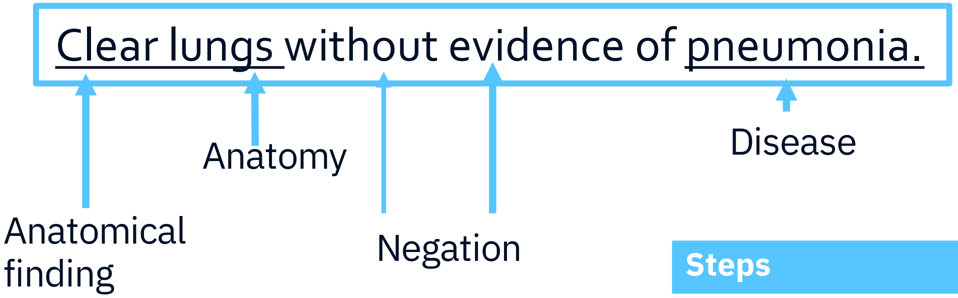
Negation

Disease

Fine-grain modifiers:

- Anatomy affected, Sub-anatomy, Location, Laterality, Severity, Size, Shape, Character, Correlation, Cause, Symptom, Hedge

Phrasal grouping – FFL patterns



Clear lungs without evidence of pneumonia.

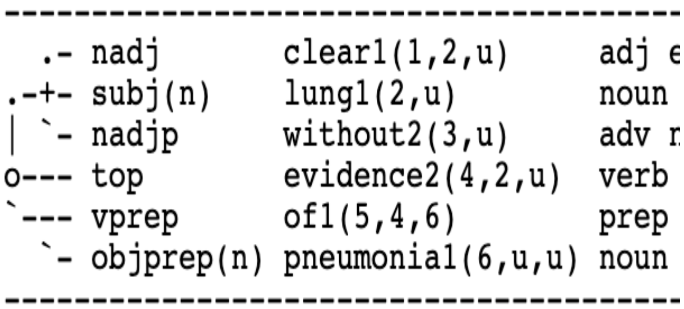
Anatomical finding

Anatomy

Negation

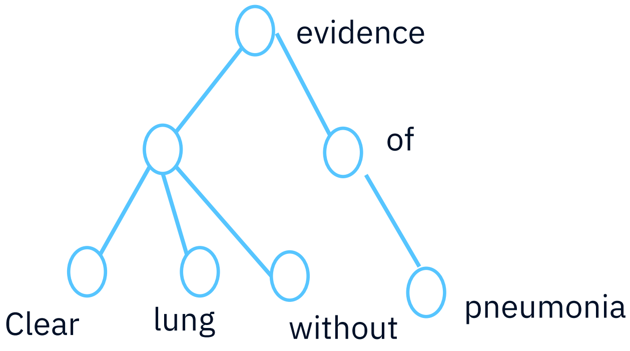
Disease

ESG Parser Output



```
---------------------------------------
  .- nadj       clear1(1,2,u)     adj e
.-+- subj(n)    lung1(2,u)        noun
| `- nadjp      without2(3,u)     adv r
o--- top        evidence2(4,2,u)  verb
`--- vprep      of1(5,4,6)        prep
  `- objprep(n) pneumonia1(6,u,u) noun
---------------------------------------
```

Dependency parse tree



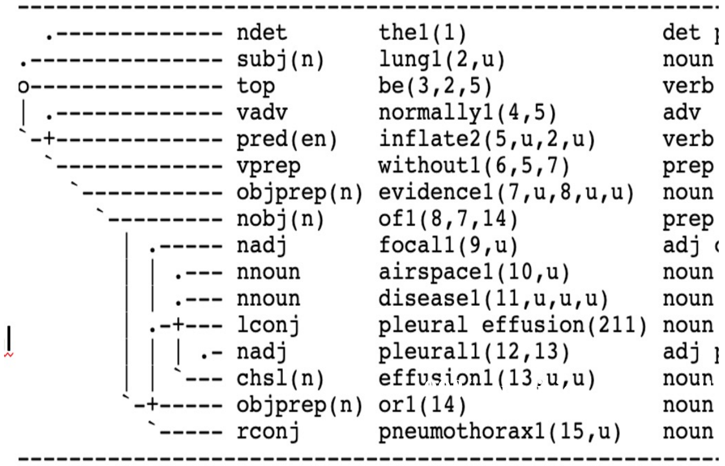
Steps	Action
Initial groups given by dependency parser	[(1,2,u)]-> clear lung -> (core group)
Phrasal grouping using connected component analysis	[(2,u)(4,2,u)(5,4,6)(6,u,u))]-> lung evidence of pneumonia -> (core group)
Negation detection	[(3,u)] -> without -> (negation span, helper group)
Assembled FFL patterns	**anatomical finding|no|clear lung|lung||||clear disease|no|pneumonia|lung|**

FFL pattern F=< T|N|C|M* >

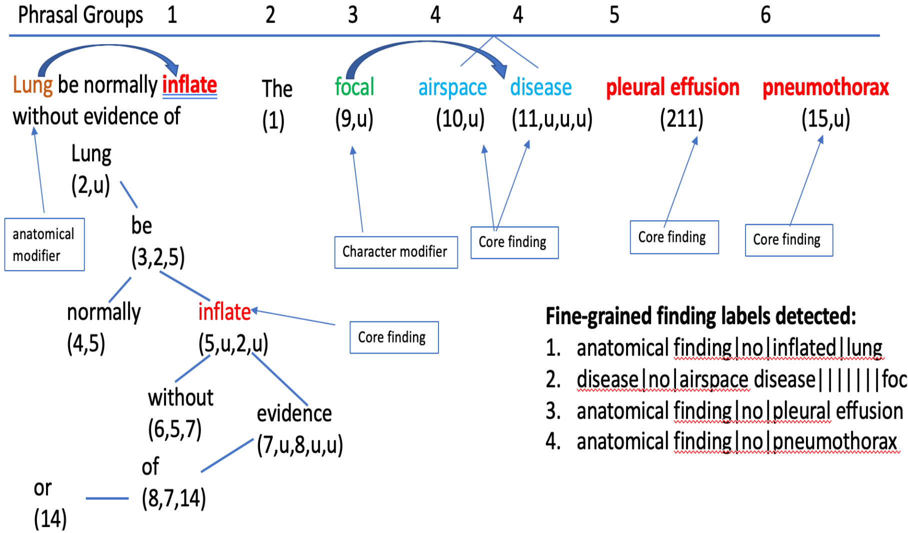
Natural language analysis of the sentence using ESG parser ("Deep parsing in Watson", M.Mc Cord, J.W. Murdock, B.K. Boguraev, IBM Journal of Research and Development (Volume: 56 , Issue: 3.4 , May-June 2012)

Fine-grained finding extraction – Another example

The lungs are normally inflated without evidence of focal airspace disease, pleural effusion, or pneumothorax.



```
---------------------------------------------------------
  .------------- ndet        the1(1)              det
.--------------- subj(n)     lung1(2,u)           noun
o--------------- top         be(3,2,5)            verb
| .------------- vadv        normally1(4,5)       adv
`-+------------- pred(en)    inflate2(5,u,2,u)    verb
  `------------- vprep       without1(6,5,7)      prep
    `----------- objprep(n)  evidence1(7,u,8,u,u) noun
      `--------- nobj(n)     of1(8,7,14)          prep
        | .----- nadj        focal1(9,u)          adj
        | | .--- nnoun       airspace1(10,u)      noun
        | | .--- nnoun       disease1(11,u,u,u)   noun
        | .-+--- lconj       pleural effusion(211) noun
        | | | .- nadj        pleural1(12,13)      adj
        | | `--- chsl(n)     effusion1(13,u,u)    noun
        `-+----- objprep(n)  or1(14)              noun
          `----- rconj       pneumothorax1(15,u)  noun
---------------------------------------------------------
```

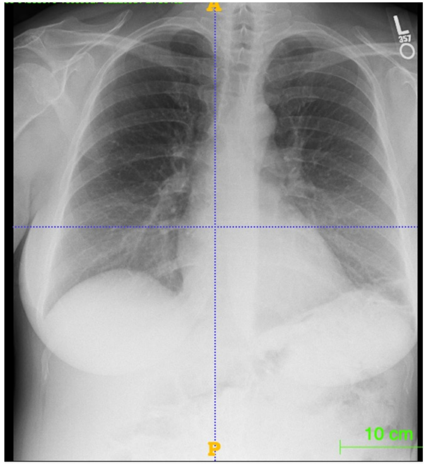


Fine-grained finding labels detected:

1. anatomical finding|no|inflated|lung
2. disease|no|airspace disease||||||focal
3. anatomical finding|no|pleural effusion
4. anatomical finding|no|pneumothorax

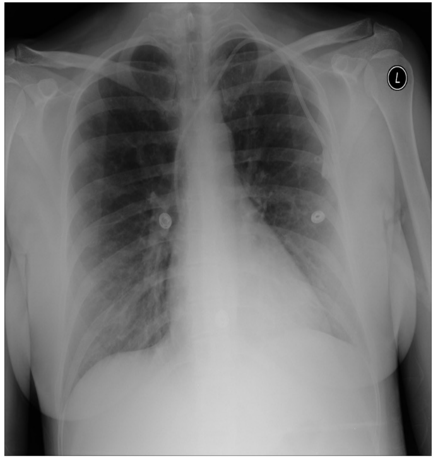
This work won the best paper award (Homer Warner Award) at AMIA 2020

Fine-grained description labels for images



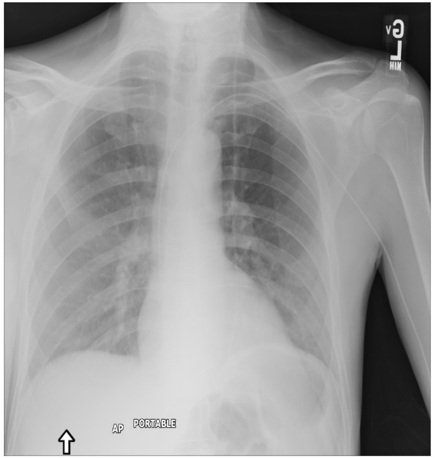
anatomicalfinding|yes|atelectasis|left lungs;;lungs;;basal||left lungs|left;;basal||left

Left basal atelectasis.



disease|yes|pneumonia|lung|lower lobe|||right lower lobe||right lower lobe|||suspicious

Right lower lobe heterogenous opacities suspicious for pneumonia.



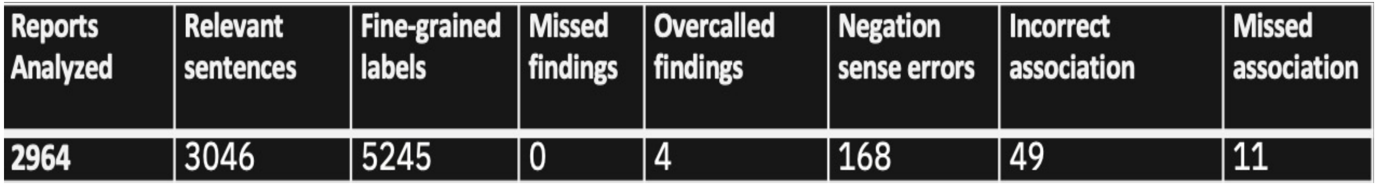
anatomicalfinding|yes|diaphragm elevated|lungs;;other soft tissues||||left

Elevated left hemidiaphragm.

Label extraction accuracy:

- Less than 4% error overall
- Less than 1% error in modifier associations
- Most are negation detection errors

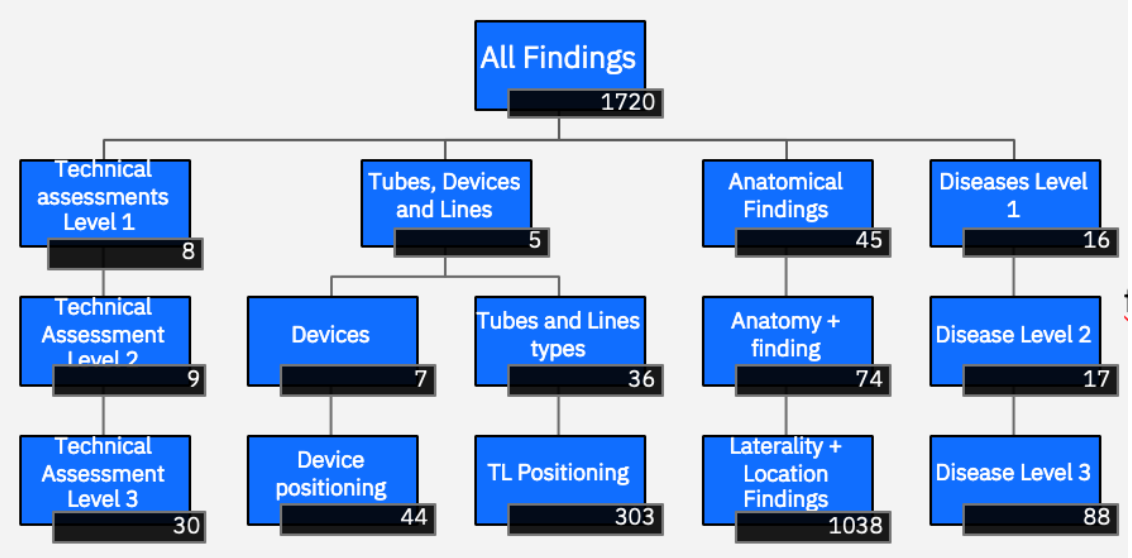
Label Extraction Accuracy



Reports Analyzed	Relevant sentences	Fine-grained labels	Missed findings	Overcalled findings	Negation sense errors	Incorrect association	Missed association
2964	3046	5245	0	4	168	49	11

Characterizing all fine-grained findings in chest X-rays

- Catalogued large number of FFL patterns from over 220,000 reports
- Auto-labeled over 500,000 images in a few hours
- Retained labels with at least 100 images
- Largest collection of fine-grained finding labels assembled!



All Findings 1720

Technical assessments Level 1 8

Technical Assessment Level 2 9

Technical Assessment Level 3 30

Tubes, Devices and Lines 5

Devices 7

Device positioning 44

Tubes and Lines types 36

TL Positioning 303

Anatomical Findings 45

Anatomy + finding 74

Laterality + Location Findings 1038

Diseases Level 1 16

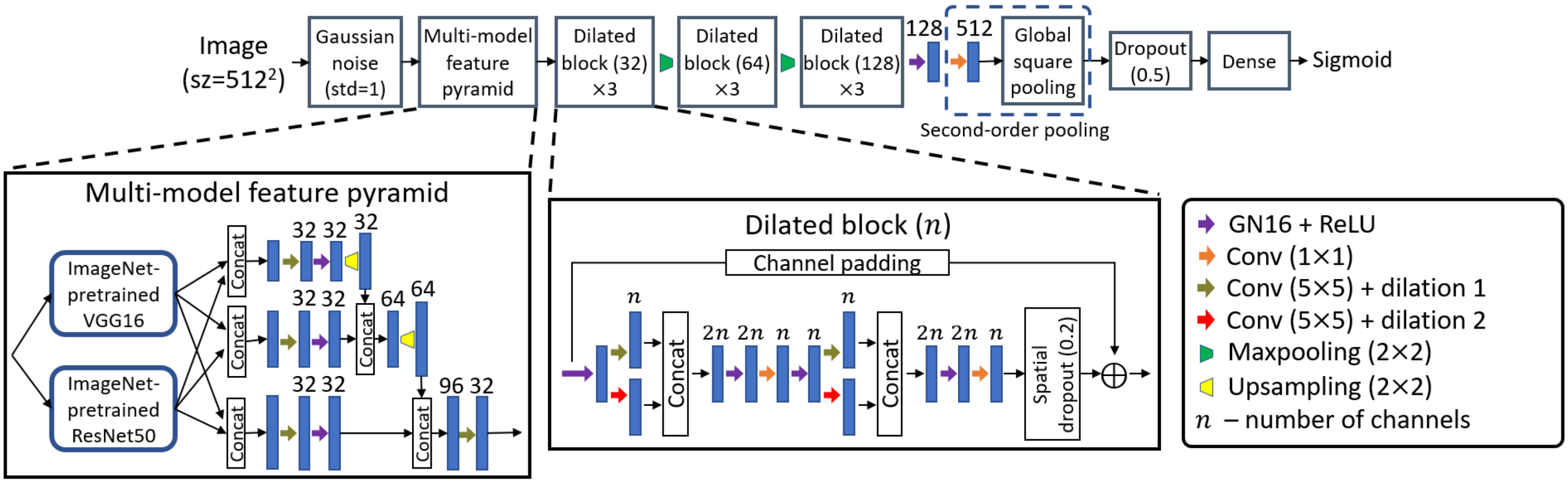
Disease Level 2 17

Disease Level 3 88

Reports analyzed	Unique sentences analyzed	FFL labels extracted	Expanded findings	CFL	Retained FFL	Spanned coverage
232964	203, 938	102,135	1720	78	**457**	83%

In Proc. AMIA'20

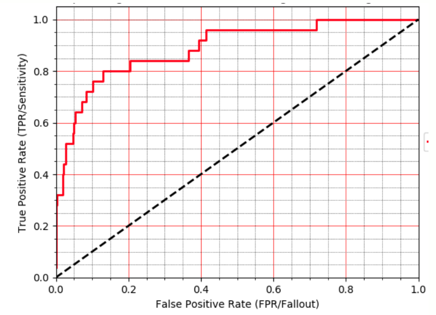
Building a single fine-grained deep learning model for all findings



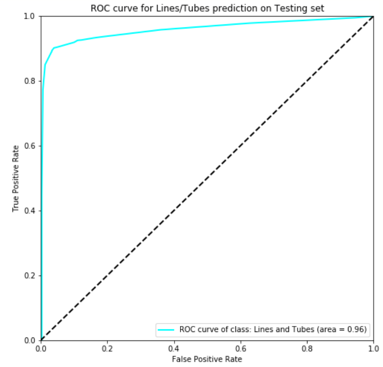
In JAMA’20

Building Deep Learning Models

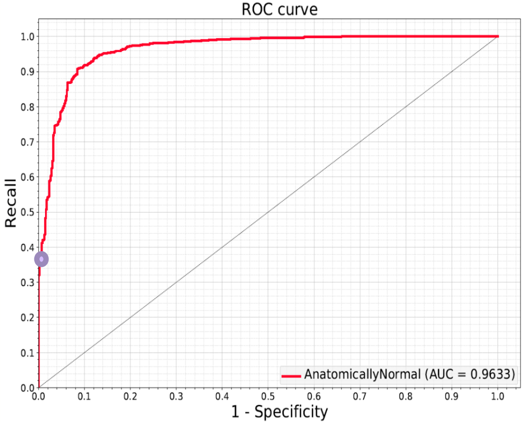
Technical Assessment (SPIE 2019)

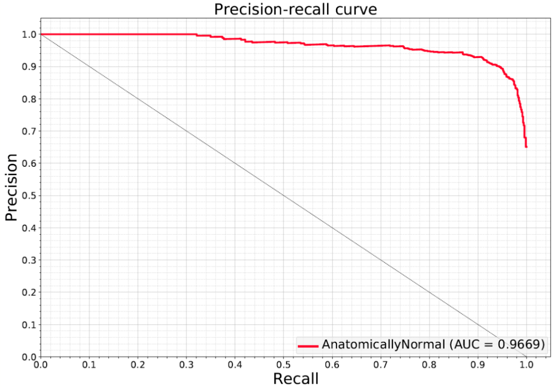


Tubes and Lines Classifier (MICCAI 2019)

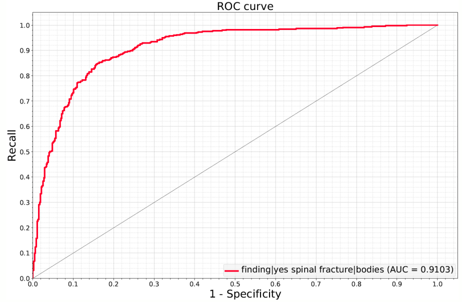


Normal/Abnormal (ISBI'2020)

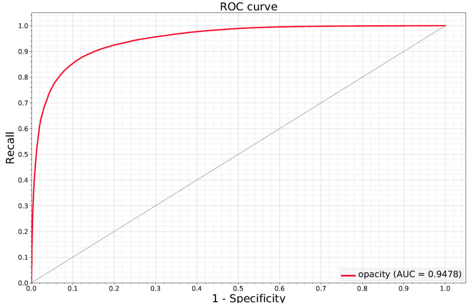




Spine Fracture Model



Opacity Model



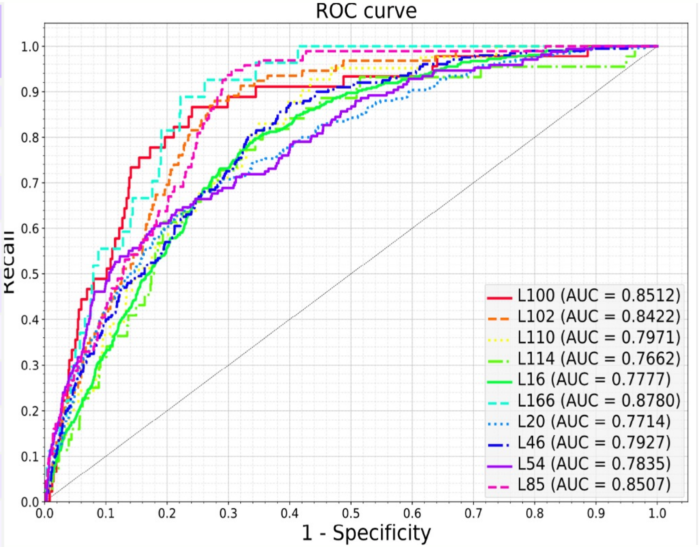
V. Subramanian et al, "Automated detection and type classification of central venous catheters in chest X-rays," Proc. MICCAI 2019, pp. 522-530

Building Models for all Core and Fine-grained Findings

Coarse and fine-grained models built with similar architecture followed by late fusion

Dataset	Train	Validate	Number of findings	Test set	Average AUC	Weighted Average AUC	Models
MIMIC-4 + NIH	249,286	35,822	78	70,932	0.81	0.84	Core finding model
MIMIC-4 + NIH	75613	10,615	457	20,941	0.73	0.73	Fine-grained model

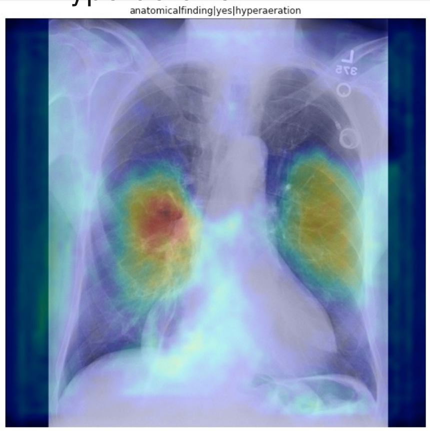
Label code	Samples	FFL semantics	AUC
L166	123	Mild dextroscoliosis	0.878
L85	464	Enteric tube at diaphragm	0.851
L100	215	Slight pulmonary edema	0.851
L102	478	Nasogastric tube in the stomach	0.842
L110	257	Elevated right hemidiaphragm	0.797
L46	1069	Moderate bilateral pleural effusions	0.793
L54	837	Pneumothorax in the right pleura	0.784
L16	2154	Moderate pulmonary edema	0.778
L20	1400	Moderate pleural effusion	0.771
L114	247	Moderate pleural effusion in the right lower lobe	0.766



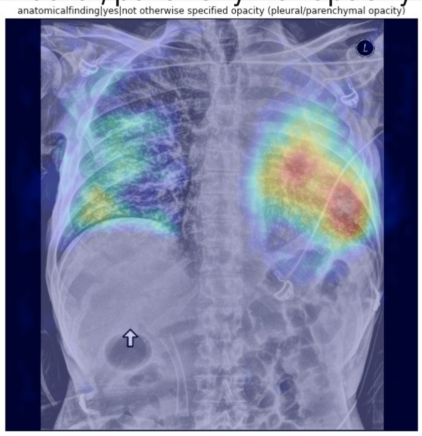
Syeda-Mahmood, T., Wong, K., Wu, J. T., Jadhav, A., & Boyko, O. (2021). Extracting and Learning Fine-Grained Labels from Chest Radiographs. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, *2020*, 1190–1199.

Disease Localization

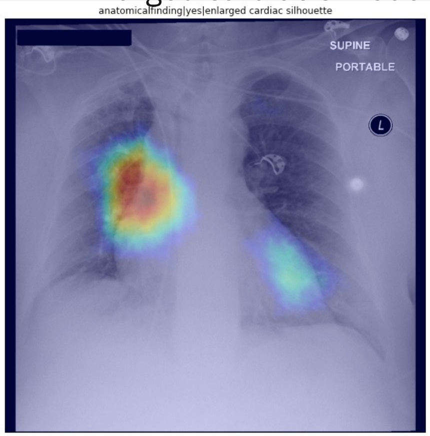
Hyperaeration



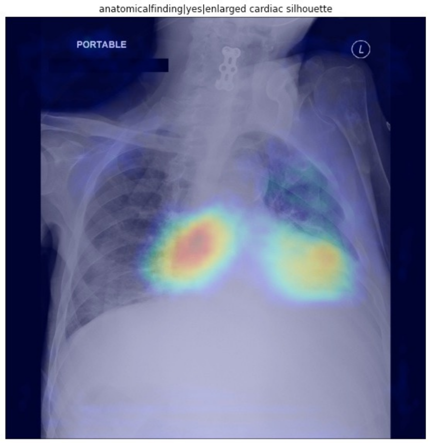
Pleural/parenchymal opacity


anatomicalfinding|yes|not otherwise specified opacity (pleural/parenchymal opacity)
L

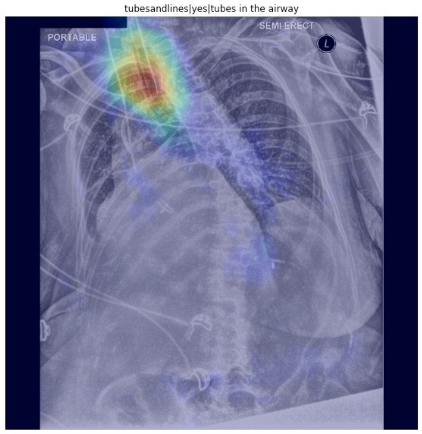
Enlarged cardiac silhouette



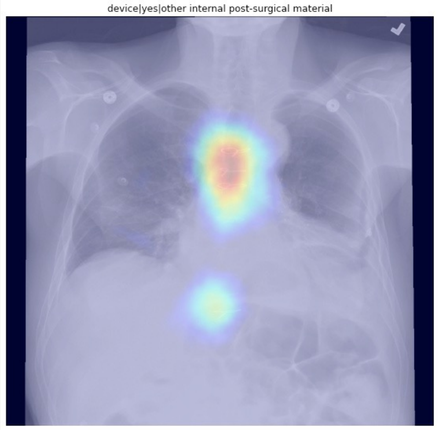
Low lung volumes



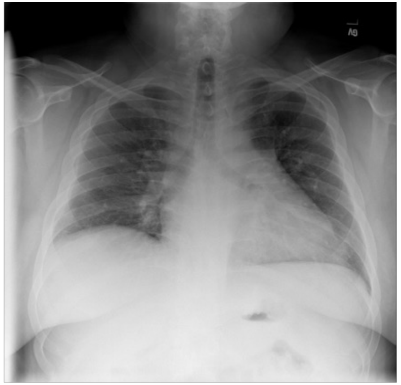
Tubes in the airway



Post-surgical material



Chest X-ray report generation

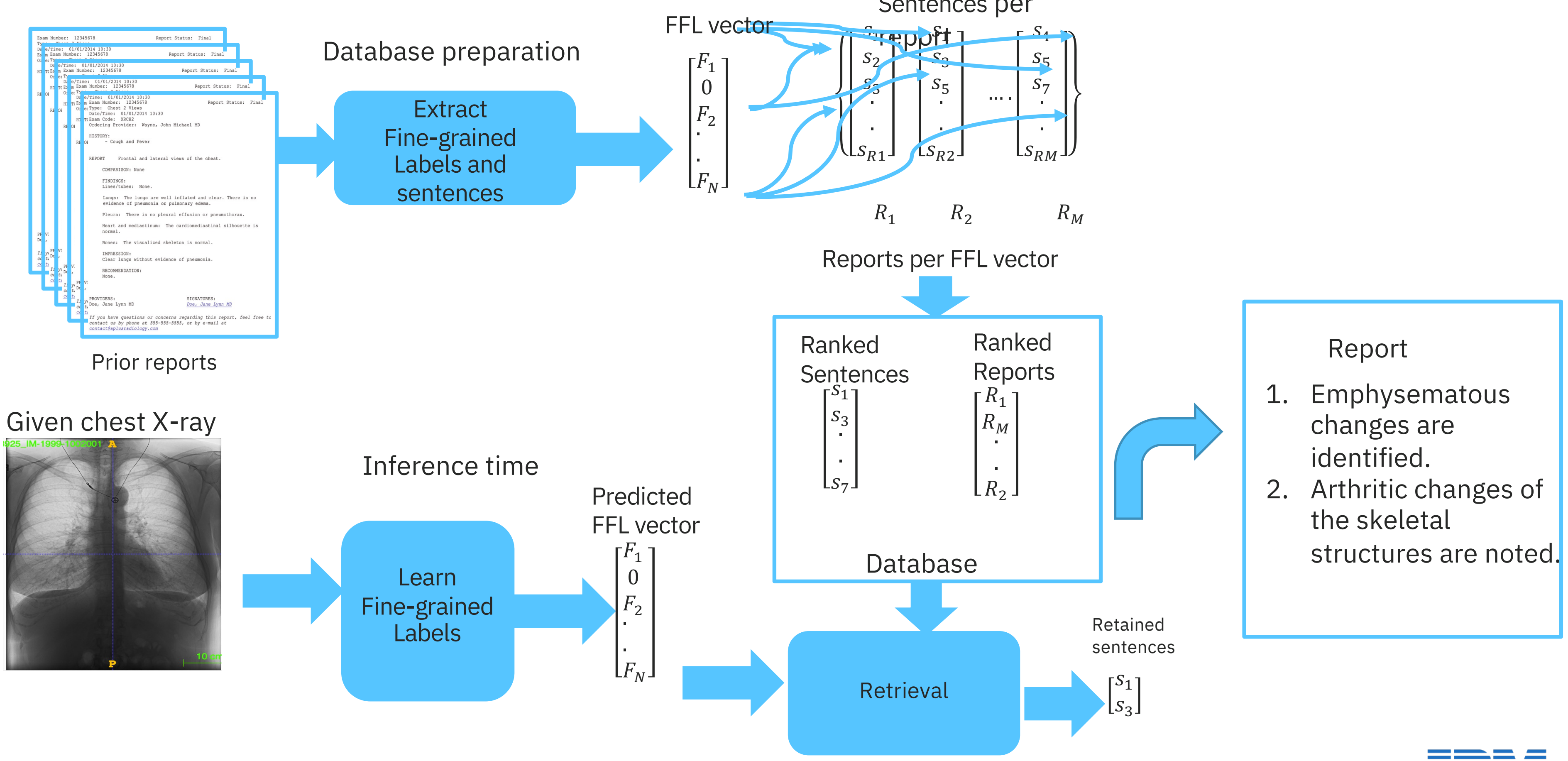


Coarse Findings	Fine-grained Findings	Report
Elevated hemidiaphragm	Elevated right hemidiaphragm mild	The right hemidiaphragm is mildly elevated.
Cardiomegaly	Cardiomegaly	Enlarged cardiac silhouette.

- Combine deep learning with document retrieval methods
  - Generate a database of sentences from prior reports that capture the findings
  - Use document retrieval techniques to rank reports and their associated sentences to match predicting findings
  - Assemble the report from ranked and edited sentences.

In MICCAI 2020

Automatic Report Generation – RAG approach



Exam Number: 12345678 Report Status: Final

Type: Chest 2 Views

Date/Time: 01/01/2014 10:30

Exam Code: XRCH2

Ordering Provider: Wayne, John Michael MD

HISTORY:
- Cough and Fever

REPORT Frontal and lateral views of the chest.

COMPARISON: None

FINDINGS:
Lines/tubes: None.

Lungs: The lungs are well inflated and clear. There is no evidence of pneumonia or pulmonary edema.

Pleura: There is no pleural effusion or pneumothorax.

Heart and mediastinum: The cardiomediastinal silhouette is normal.

Bones: The visualized skeleton is normal.

IMPRESSION:
Clear lungs without evidence of pneumonia.

RECOMMENDATION:
None.

PROVIDERS: Doe, Jane Lynn MD

SIGNATURES: Doe, Jane Lynn MD

If you have questions or concerns regarding this report, feel free to contact us by phone at 555-555-5555, or by e-mail at contact@aplusradiology.com

Prior reports

Database preparation

Extract Fine-grained Labels and sentences

FFL vector

$$\begin{bmatrix} F_1 \\ 0 \\ F_2 \\ \cdot \\ \cdot \\ F_N \end{bmatrix}$$

Sentences per report

$$\begin{Bmatrix} \begin{pmatrix} \begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ \cdot \\ \cdot \\ s_{R1} \end{bmatrix} & \begin{bmatrix} s_1 \\ s_3 \\ s_5 \\ \cdot \\ \cdot \\ s_{R2} \end{bmatrix} & \cdots & \begin{bmatrix} s_4 \\ s_5 \\ s_7 \\ \cdot \\ \cdot \\ s_{RM} \end{bmatrix} \end{pmatrix} \end{Bmatrix}$$

R_1 R_2 R_M

Reports per FFL vector

Database

Ranked Sentences

$$\begin{bmatrix} s_1 \\ s_3 \\ \cdot \\ \cdot \\ s_7 \end{bmatrix}$$

Ranked Reports

$$\begin{bmatrix} R_1 \\ R_M \\ \cdot \\ \cdot \\ R_2 \end{bmatrix}$$

Given chest X-ray

Inference time

Learn Fine-grained Labels

Predicted FFL vector

$$\begin{bmatrix} F_1 \\ 0 \\ F_2 \\ \cdot \\ \cdot \\ F_N \end{bmatrix}$$

Retrieval

Retained sentences

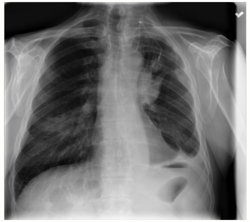
$$\begin{bmatrix} s_1 \\ s_3 \end{bmatrix}$$

Report

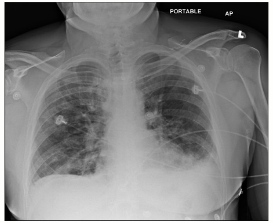
1. Emphysematous changes are identified.
2. Arthritic changes of the skeletal structures are noted.

Automated Report Generation Results

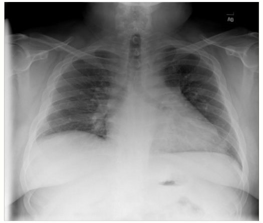
Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
Vis-Att [20]	0.39	0.25	0.16	0.11	0.16	0.32
MM-Att [25]	0.46	0.35	0.27	0.19	0.27	0.36
KERP [12]	0.48	0.32	0.22	0.16	-	0.33
Template-based	0.28	0.29	0.32	0.27	0.35	0.34
Clinical Accurate [15]	0.35	0.22	0.15	0.10	-	0.45
Co-Att [9]	0.51	0.39	0.30	0.25	0.21	0.44
Jiebo Luo [28]	0.53	0.37	0.31	0.25	0.34	0.45
CFL-only	0.49	0.39	0.36	0.32	0.48	0.52
FFL+CFL (ours)	**0.56**	**0.51**	**0.50**	**0.49**	**0.55**	**0.58**



00019000_000.png



00005567_000



00011569_000

	00019000_000.png	00005567_000	00011569_000
Ground Truth	overall impression : Left hilar opacity may represent primary lung mass. Left hilar opacity. Left port.	Small left effusion. Pleuroparenchymal opacities at the left lung base. Wires external to patient. Surgical clips superior to the left clavicle.	The right hemidiaphragm is mildly elevated. Overall impression : cardiomegaly.
Ours	Left perihilar opacity. lung mass. Left port noted.	Left pleural effusion. Left lung opacities. External tubing. Surgical clips near left clavicle.	Elevated right hemidiaphragm. Enlarged cardiac silhouette.

- Tested against a ground truth dataset of 2964 unique reports from the Indiana collection.
- Ground truth statements were extracted by Indiana reports from Findings and Impression sections.
- The sentences of reports generated by machine were compared to those of the ground truth using a variety of scores

In T. Syeda-Mahmood et al., “Chest X-ray Report Generation through Fine-Grained Label Learning, in Proc. MICCAI’2020

Clinical studies performed

- Normal/abnormal discrimination studies
  - Deccan hospital retrospective study
  - Deccan hospital prospective study
- Discrete preliminary read studies
  - Consensus ground truth generation
- Preliminary report quality Turing study
  - Blind observations of reports generated by unknown sources

Field pilot study - Deccan Hospital

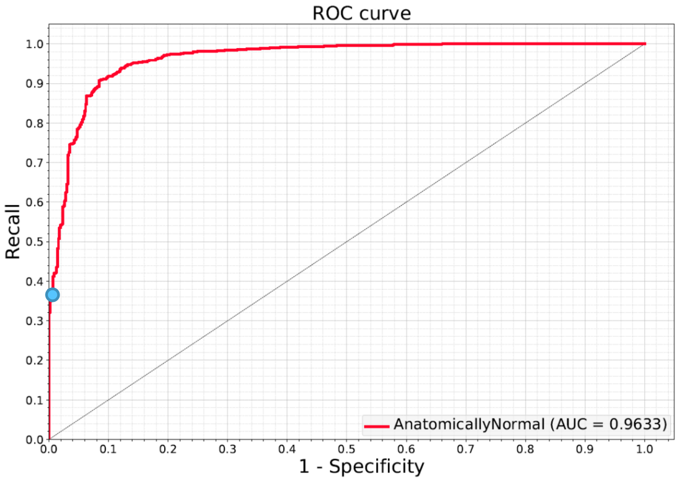
- Deccan Hospital
  - a 600 bed hospital in Hyderabad, India
- Period of study:
  - Retrospective Data collection:
    - October 2018-March 2019
  - Prospective data collection
    - October 2019 – March 2020
- Selection of Patients:
  - All patients being seen in Deccan hospital for whom chest X-rays are being taken
  - All patients signed consent form
  - Chest X-ray study performed
  - Clinical data is recorded and followed up

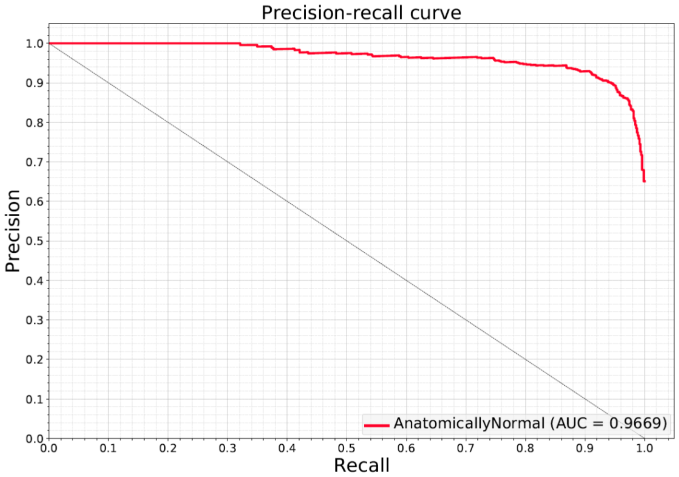
Verification method of assessment of normal read

- No clinical follow-up
- Patient discharged with no follow-ups
- Patient not readmitted or revisits within 30 days.
- As validated by pulmonologist

Testing on independent data, three radiologist consensus

- Total test set size 1749
- The radiologist panel agreed on normal-abnormal labels for **1271** images in this Deccan test set.
- Performance of our model on this triple consensus images:
  - Area under ROC curve: 0.96
  - Area under PR curve: 0.97
  - **We can detect one third of the normal images without a single false negative!**
- Note: If two out of three majority vote is considered as the ground truth label, then all 1749 images can be used for testing. The area under ROC curve in this case was 0.92 and 25% of normal can be filtered out.

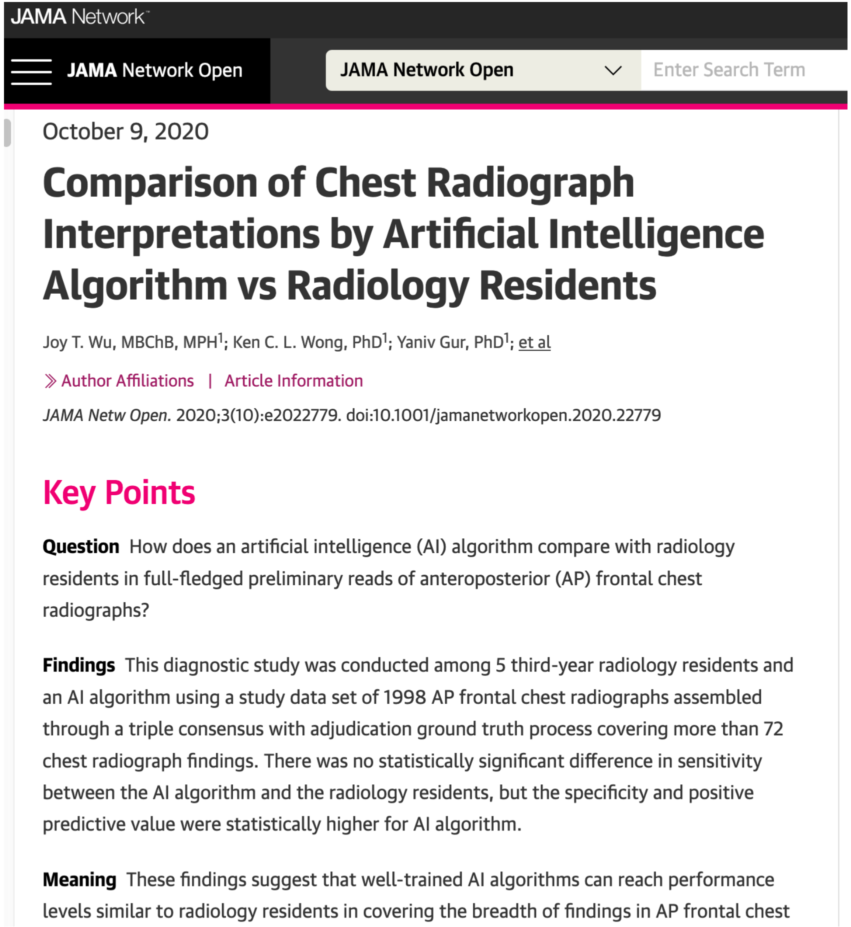




In Proc. ISBI 2020

Discrete label comparison study

- Experiments
  - 5 radiology residents
  - 1998 chest X-rays
  - 400 images per radiology resident
- Measures
  - Label-based AUC comparisons
  - ROC curve placements
  - Kappa scores
  - Average Image-based precision and recall
  - Variance between radiologists



October 9, 2020

Comparison of Chest Radiograph Interpretations by Artificial Intelligence Algorithm vs Radiology Residents

Joy T. Wu, MBChB, MPH[1]; Ken C. L. Wong, PhD[1]; Yaniv Gur, PhD[1]; et al

» Author Affiliations | Article Information

JAMA Netw Open. 2020;3(10):e2022779. doi:10.1001/jamanetworkopen.2020.22779

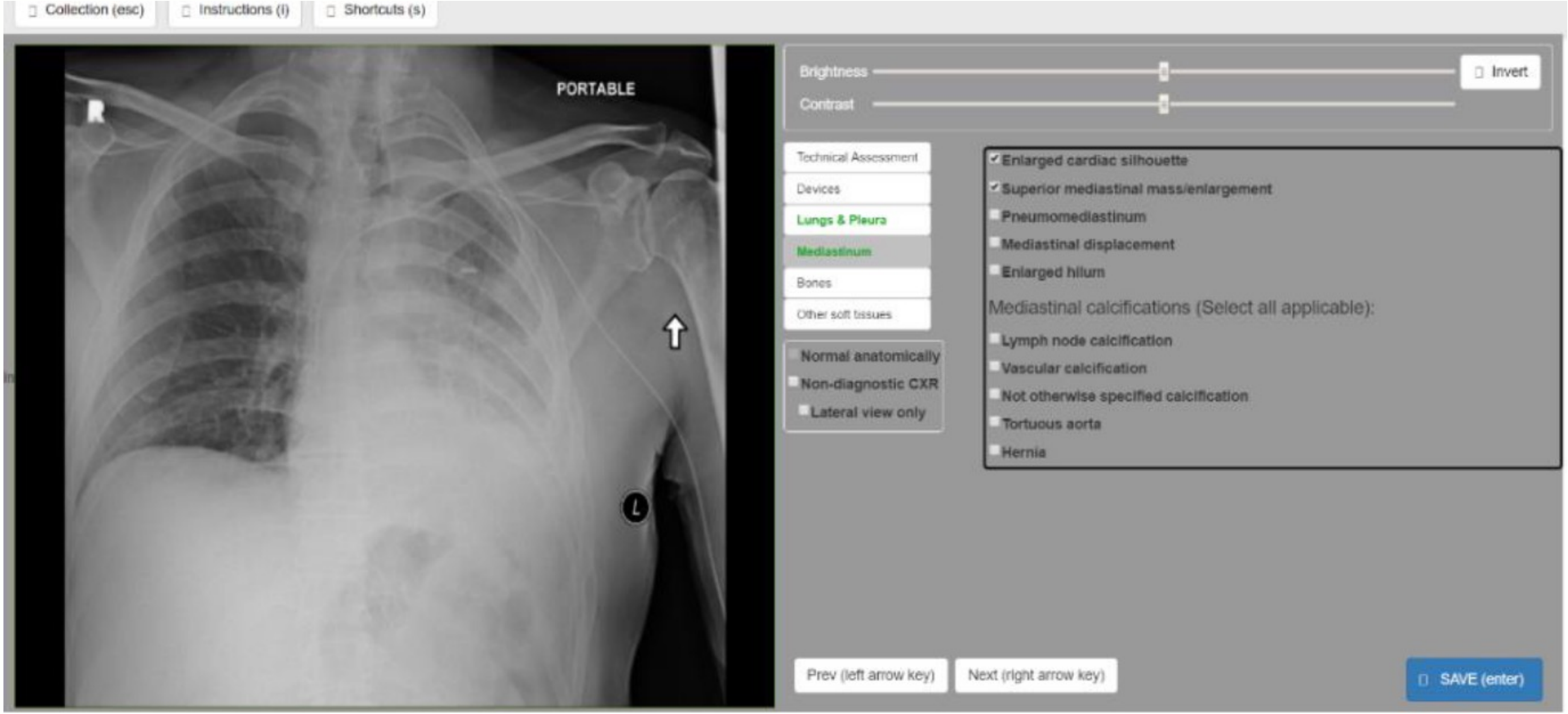
Key Points

Question How does an artificial intelligence (AI) algorithm compare with radiology residents in full-fledged preliminary reads of anteroposterior (AP) frontal chest radiographs?

Findings This diagnostic study was conducted among 5 third-year radiology residents and an AI algorithm using a study data set of 1998 AP frontal chest radiographs assembled through a triple consensus with adjudication ground truth process covering more than 72 chest radiograph findings. There was no statistically significant difference in sensitivity between the AI algorithm and the radiology residents, but the specificity and positive predictive value were statistically higher for AI algorithm.

Meaning These findings suggest that well-trained AI algorithms can reach performance levels similar to radiology residents in covering the breadth of findings in AP frontal chest

Proving the capability of AI – Recording the radiologists performance



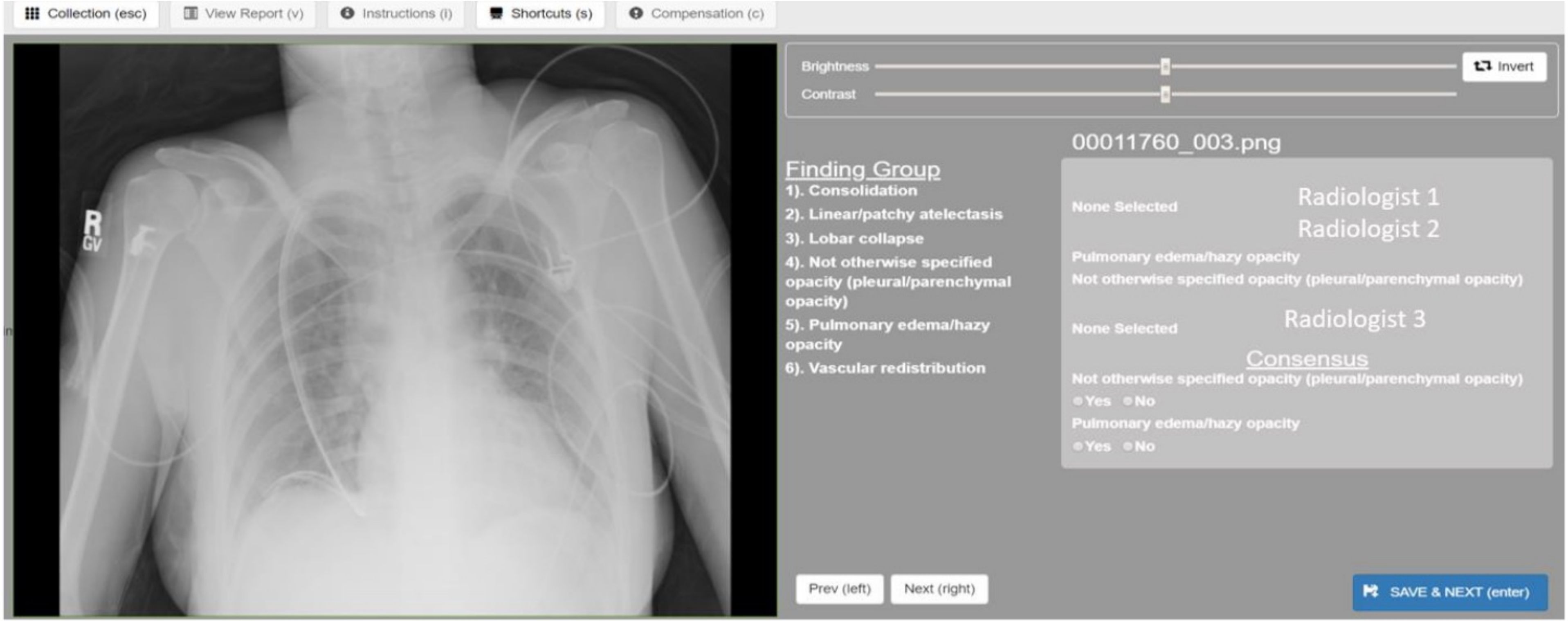
- □ Normal anatomically
- □ Non-diagnostic CXR
- □ Lateral view only

- ☑ Enlarged cardiac silhouette
- ☑ Superior mediastinal mass/enlargement
- □ Pneumomediastinum
- □ Mediastinal displacement
- □ Enlarged hilum

Mediastinal calcifications (Select all applicable):

- □ Lymph node calcification
- □ Vascular calcification
- □ Not otherwise specified calcification
- □ Tortuous aorta
- □ Hernia

Proving the capability of AI – Triple Consensus Studies



00011760_003.png

Finding Group

1). Consolidation

2). Linear/patchy atelectasis

3). Lobar collapse

4). Not otherwise specified opacity (pleural/parenchymal opacity)

5). Pulmonary edema/hazy opacity

6). Vascular redistribution

None Selected — Radiologist 1

Radiologist 2

Pulmonary edema/hazy opacity

Not otherwise specified opacity (pleural/parenchymal opacity)

None Selected — Radiologist 3

Consensus

Not otherwise specified opacity (pleural/parenchymal opacity)

○ Yes ○ No

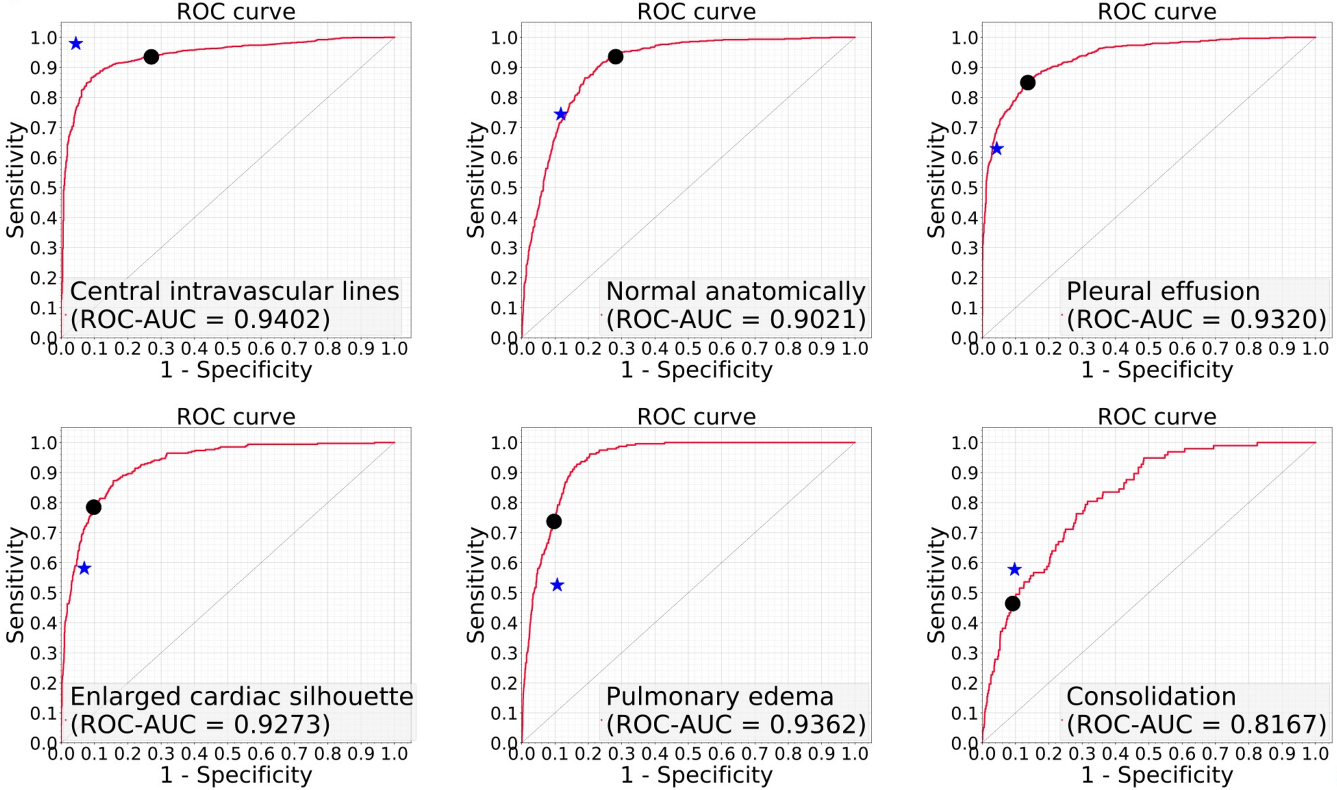
Pulmonary edema/hazy opacity

○ Yes ○ No

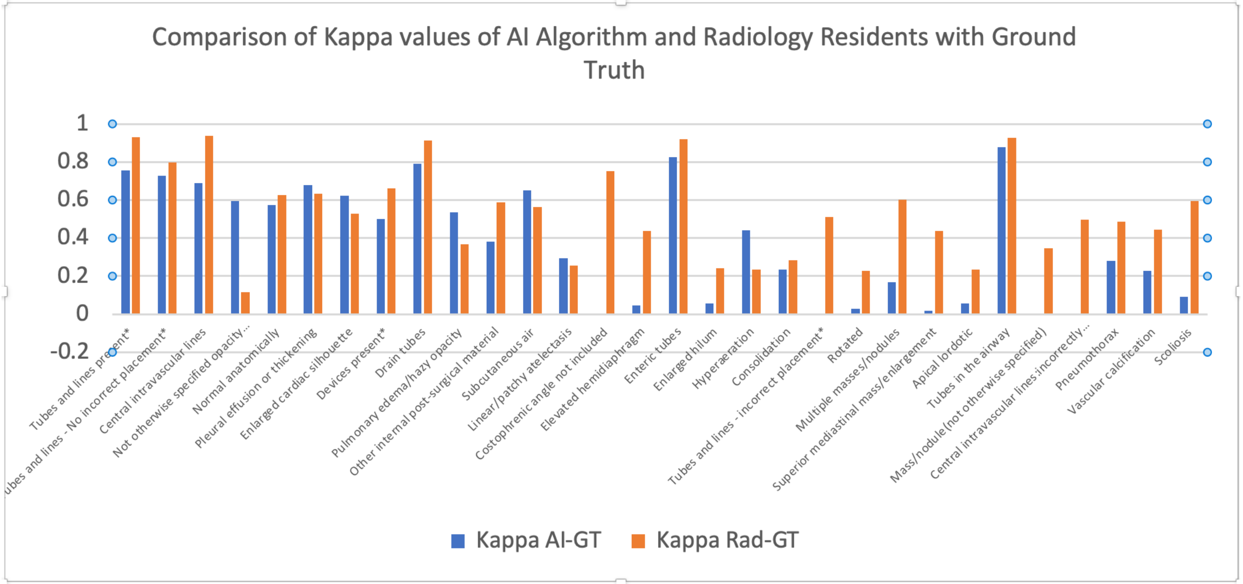
AUC Comparisons

Finding label	Number of images in the comparison study dataset	Interpret difficulty	AUC in Comparison Study Dataset	DL Label-based PPV	DL Label-based sensitivity	DL Label-based specificity	Rads Label-based PPV	Rads Label-based sensitivity	Rads label-based specificity
Central intravascular lines	1296	medium	0.865	0.864	0.935	0.729	0.976	0.979	0.956
Not otherwise specified opacity (pleural/parenchymal opacity)	713	low	0.787	0.695	0.818	0.801	0.719	0.122	0.974
Normal anatomically	637	medium	0.932	0.608	0.936	0.718	0.748	0.744	0.882
Pleural effusion or thickening	604	low	0.940	0.729	0.849	0.863	0.862	0.629	0.956
Enlarged cardiac silhouette	339	low	0.902	0.621	0.785	0.902	0.631	0.581	0.931
Drain tubes	240	low	0.927	0.746	0.904	0.958	0.888	0.963	0.984
Pulmonary edema/hazy opacity	236	low	0.936	0.504	0.737	0.903	0.397	0.525	0.893
Other internal post-surgical material	228	low	0.997	0.503	0.395	0.950	0.540	0.794	0.913
Subcutaneous air	203	low	0.817	0.671	0.704	0.961	0.946	0.429	0.997
Linear/patchy atelectasis	168	low	0.997	0.322	0.405	0.922	0.235	0.661	0.802
Costophrenic angle not included	149	medium	0.836	1.000	0.000	1.000	0.829	0.718	0.988
Elevated hemidiaphragm	137	low	0.976	0.444	0.029	0.997	0.508	0.445	0.968
Enteric tubes	100	high	0.978	0.832	0.840	0.991	0.921	0.930	0.996
Enlarged hilum	100	high	0.583	0.250	0.040	0.994	0.355	0.220	0.979
Hyperaeration	100	low	0.917	0.440	0.510	0.966	0.450	0.180	0.988
Consolidation	97	low	0.869	0.205	0.464	0.908	0.232	0.577	0.903
Rotated	92	low	0.513	0.167	0.022	0.995	0.196	0.489	0.903
Multiple masses/nodules	91	low	0.744	0.264	0.154	0.980	0.540	0.736	0.970
Superior mediastinal mass/enlargement	88	high	0.757	0.200	0.011	0.998	0.471	0.455	0.976
Apical lordotic	87	high	0.72	0.190	0.046	0.991	0.421	0.184	0.988
Tubes in the airway	86	low	0.773	0.884	0.884	0.995	0.951	0.907	0.998
Mass/nodule (not otherwise specified)	79	high	0.944	0.000	0.000	0.998	0.348	0.405	0.969
Central intravascular lines - incorrectly positioned	78	low	0.753	1.000	0.000	1.000	0.365	0.949	0.933
Pneumothorax	66	high	0.682	0.250	0.409	0.958	0.463	0.561	0.978
Vascular calcification	62	low	0.611	0.476	0.161	0.994	0.769	0.323	0.997
Scoliosis	59	low	0.666	0.600	0.051	0.999	0.587	0.627	0.987

Comparison with radiologists



Comparison with radiologists

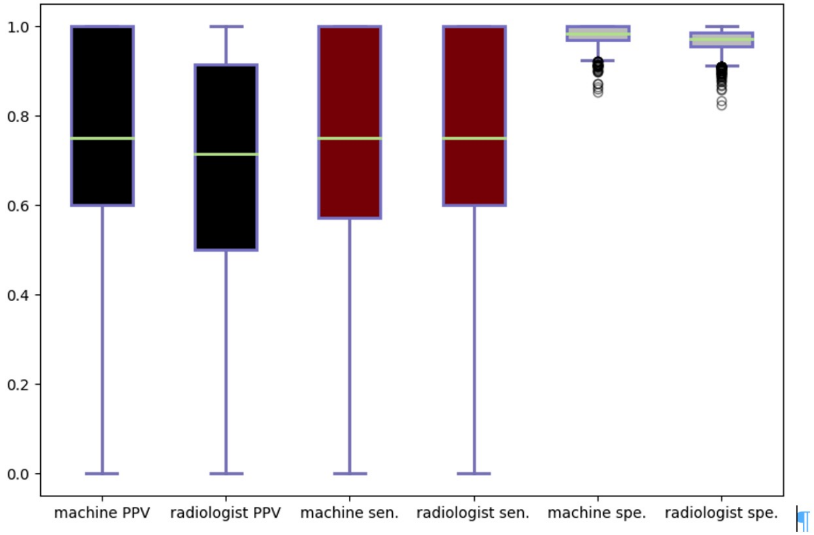


Comparison of Kappa values of AI Algorithm and Radiology Residents with Ground Truth

Method	Number of Images	Number of findings	Average image-based PPV	Average image-based sensitivity	Average image-based specificity
Resident 1	399	72	0.594 [0.567, 0.621]	0.688 [0.662,0.716]	0.958 [0.955, 0.962]
Resident 2	399	72	0.722 [0.697,0.748]	0.743 [0.719,0.768]	0.975 [0.972, 0.977]
Resident 3	400	72	0.704 [0.678,0.731]	0.729 [0.704,0.754]	0.971 [0.968,0.974]
Resident 4	400	72	0.648 [0.623,0.674]	0.685 [0.659,0.711]	0.967 [0.964,0.969]
Resident 5	400	72	0.743 [0.714,0.766]	0.755 [0.729,0.780]	0.975 [0.972,0.977]

Overall results

- No statistically significant difference in sensitivity between AI and Rads
- But AI is statistically better on PPV and specificity than Rads!
- Rads do better on more difficult findings, e.g. masses, pneumothorax



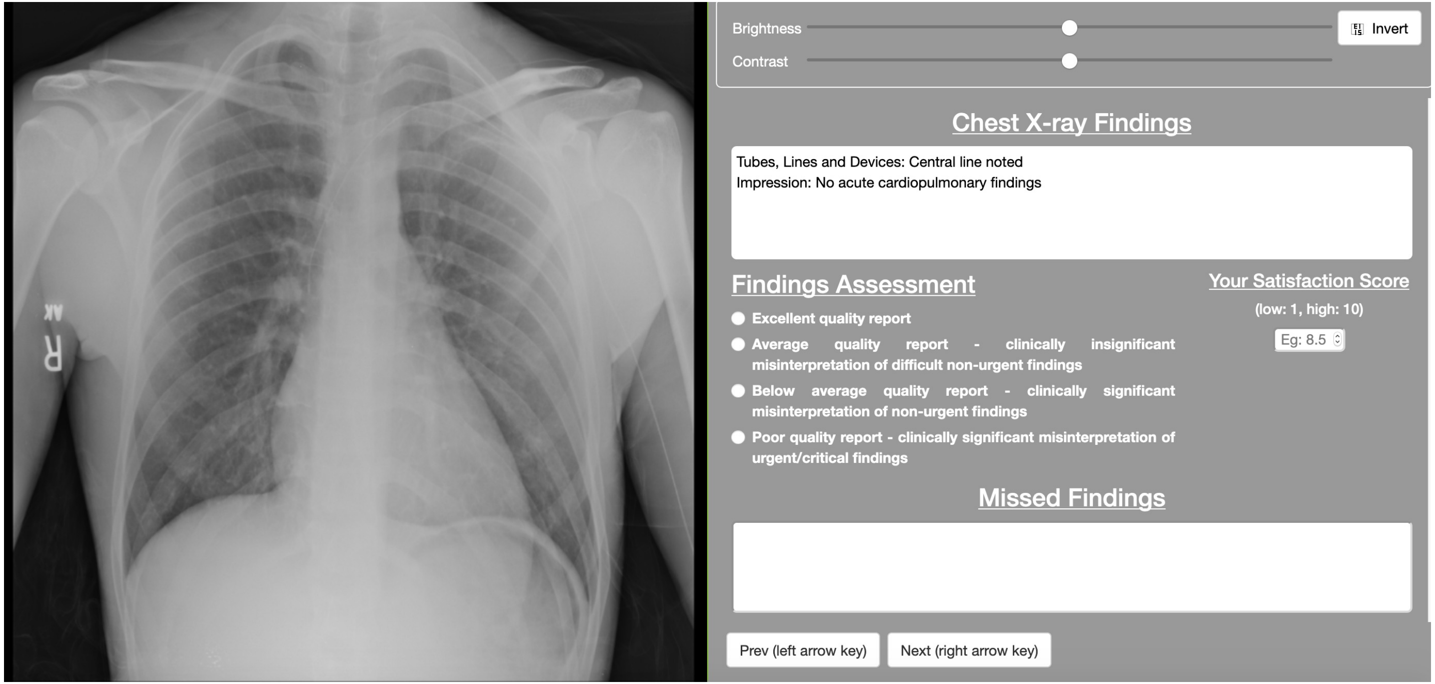
Method	Images tested	Total number of findings tested	Average image-based PPV	Average image-based sensitivity	Average Image-based specificity
R3 Residents	1998	72	68.2%	72%	97.3%
Algorithm	1998	72	73%	71.6%	98%
P-value (AI vs Rads)			*p<0.001*	*P=0.662*	*P<0.001*

Findings performance

AI Outperformed radiologists	Similar Performance of AI and Radiologists	Radiologists outperformed AI
Not otherwise specified opacity (pleural/parenchymal opacity)	Tubes and lines present	Scoliosis
Pleural effusion or thickening	Tubes in the airway	Enlarged hilum
Enlarged cardiac silhouette	Enteric tubes	Rotated
Pulmonary edema/hazy opacity	Drain tubes	Costophrenic angle not included
Subcutaneous air	tubesandlines - no incorrect placement	Elevated hemidiaphragm
Hyperaeration	Device present	Superior mediastinal mass/enlargement
	Normal anatomically	Apical Lordotic
	Other internal post-surgical material	Mass/nodule (not otherwise specified)
	Pneumothorax	Central vascular lines – incorrectly positioned
	Linear/patchy atelectasis	Multiple masses and nodules
	Central intravascular lines	Tubes and lines – incorrect placement
	Consolidation	
	Vascular calcification	

Turing Study – Comparing read performance

- Jointly with University of Maryland, Baltimore County
- First rolled out at RSNA 2018
- Latest released at AI Symposium for Biomedical Imaging Across Scales, February 2020
- Read quality scores from American College of Radiology



Chest X-ray Findings

Tubes, Lines and Devices: Central line noted
Impression: No acute cardiopulmonary findings

Findings Assessment

○ Excellent quality report
○ Average quality report - clinically insignificant misinterpretation of difficult non-urgent findings
○ Below average quality report - clinically significant misinterpretation of non-urgent findings
○ Poor quality report - clinically significant misinterpretation of urgent/critical findings

Your Satisfaction Score

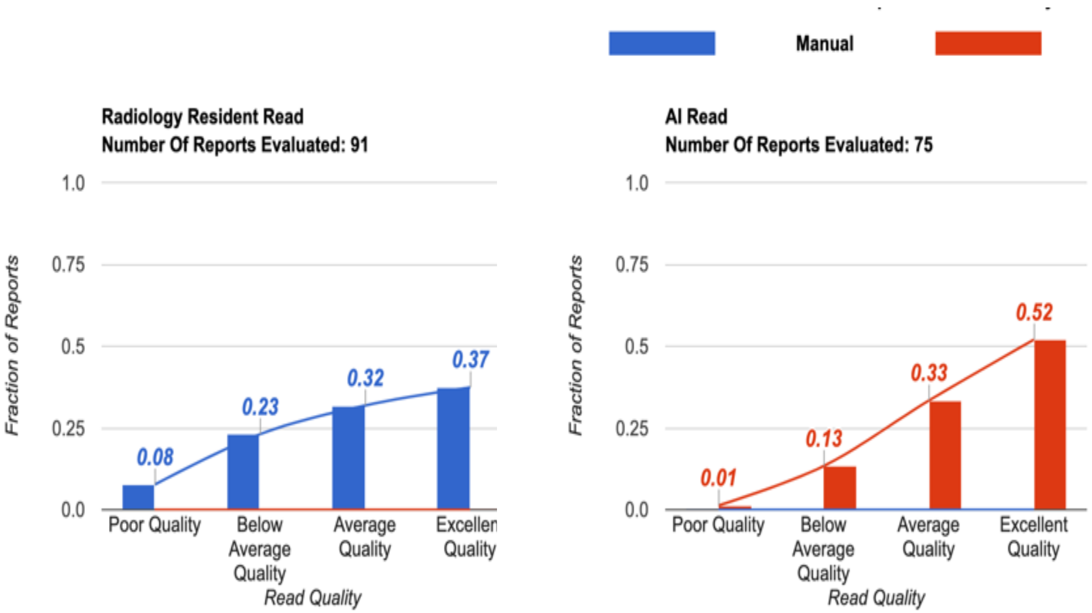
(low: 1, high: 10)

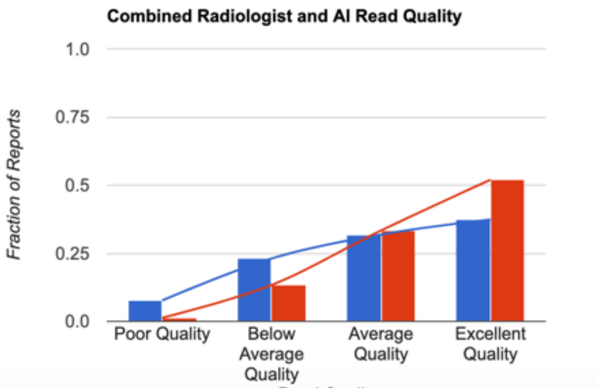
Eg: 8.5

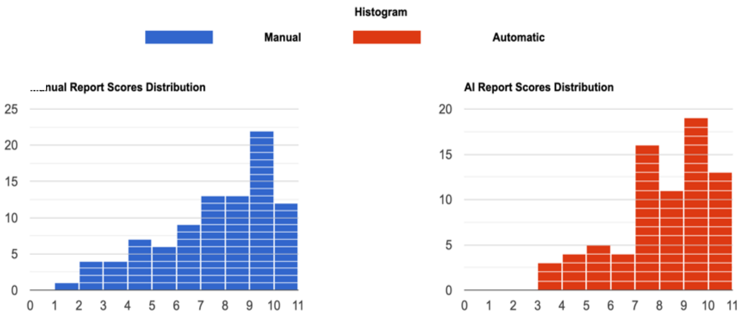
Missed Findings

Board certified radiologists evaluate blinded reports (written by residents and machines) for chest X-rays with triple consensus ground truth.

Turing Study – Results







Impression Scores

Score range 1-10 (best)

*Study being repeated at conferences

- 12 residents participated
- 166 reports evaluated in 1 hour live on site
- 3 senior radiologists acting as attendings
- Radiologists were blinded to the origin of reports
- None of the radiologists could tell which report came from machine or resident

Chest X-ray Automated Reporting - Summary

- Breadth & depth
  - Largest number of findings (237 core findings, nearly 2000 fine-grained)
  - Multi-hospital training data (4 hospitals)
  - Robustness across hospitals
  - State-of-the-art performance beats radiologist performance across widest variety of findings (image-based sensitivity, PPV, specificity)
- Proven best-in-class performance for specialized models
  - Normal/abnormal (AUC- 0.96)
  - Technical assessment (AUC – 0.93)
  - Tubes/lines (AUC -0.89)
  - All findings (Average AUC – 0.81)
  - Formal clinical studies conducted & IP secured
- Automatic report generation
  - Best-in-class BLEU score (0.56-0.58)
  - Semantic consistency in reporting
  - Virtually indistinguishable from templated reports selected manually