

# “Is That Good?”

## Evaluating Things Is Hard

Willie Boag



@WilliamBoag

6.S897/HST.956  
FRI APR 26



# Class Announcements

- PS6 due last night (+ slack days)
- Hope projects are going well!
- Final Community Consulting on Monday April 29 (5-7 in Star)

# Overview

- Case Studies
  - Generating Radiology Reports from CXR Images
  - NLP for Predicting Readmission
- Evaluating Things is Hard in Healthcare (even without the ML)

# Reminder: What We've Said So Far

[@86](#) “How does one generally decide what evaluation metric to use (eg. AUC, Precision/Recall, just accuracy, etc)?”



**the instructors' answer**, *where instructors collectively construct a single answer*

In general, it depends :P

More specifically, you should get a clinical collaborator for whatever task you're working on. Have a conversation with that collaborator about what they want Machine Learning for and what it means to be doing a good job. For instance:

- We saw that for diabetes prediction, we cared about top-100 PPV because what mattered was whether we are delivering the intervention to the right people.
- Sometimes you're more interested in assessing the discriminative power of your model across thresholds (because it's not clear yet which threshold will be used downstream), and maybe AUC is better.

And in some sense, all of these metrics are just proxies for what we actually care about (i.e. improving care, helping patients, etc). If one believes that the ultimate goal is to make healthcare better, then you should pick your eval metrics such that they are the closest approximation to what it means to be "better" in the real world. Because while the gold standard might be deploying our tools in real life, in practice that can be a long & costly process. So it's easier to pick these standard metrics as a starting point. With luck, a conversation with your clinical collaborator will help you pick the metrics that are least wrong!

related: <https://lukeoakdenrayner.wordpress.com/2019/01/21/medical-ai-safety-doing-it-wrong/>

~ An instructor (David Sontag) endorsed this answer ~

# Case Study 1: Generating Radiology Reports



**Model**



the patient was imaged in a lordotic position, which distorts the mediastinal contours. within that limitation, the lungs are clear without consolidation or edema. the mediastinum is otherwise unremarkable. the cardiac silhouette is within normal limits for size. no effusion or pneumothorax is noted. no displaced fractures are evident.



**St. Michael's**

Inspired Care. Inspiring Science.

# We Made a Very Fancy Model!

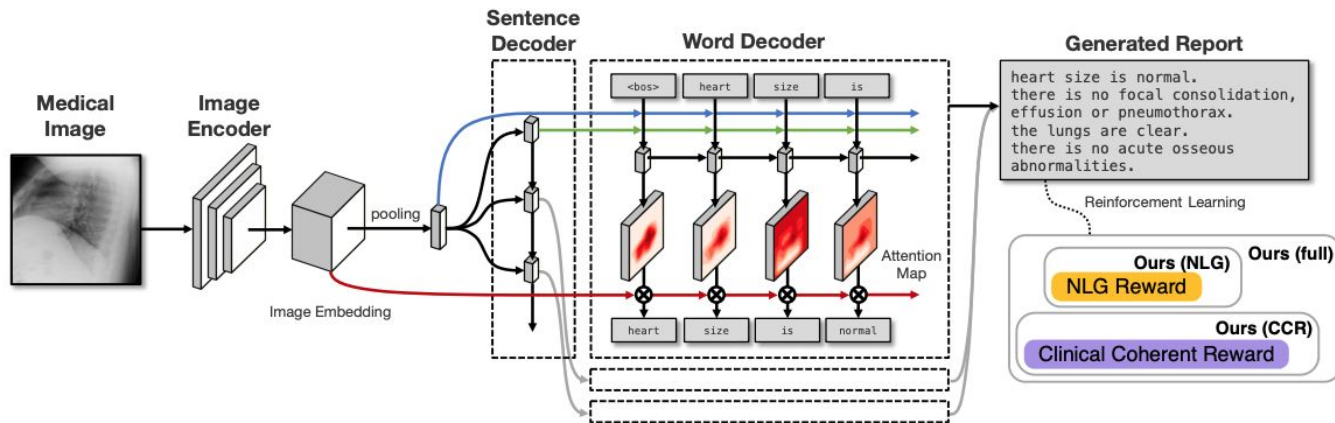


Figure 2: **The model for our proposed *Clinically Coherent Reward*.** Images are first encoded into image embedding maps, and a sentence decoder takes the pooled embedding to recurrently generate topics for sentences. The word decoder then generates the sequence from the topic with attention on the original images. NLG reward, clinically coherent reward, or combined, can then be applied as the reward for reinforcement policy learning.

Look how fancy it is!

# It Generates Great Output!



Ground Truth

cardiomegaly is moderate. bibasilar atelectasis is mild. there is no pneumothorax. a lower cervical spinal fusion is partially visualized. healed right rib fractures are incidentally noted.

Ours (full)

ap portable upright view of the chest. there is no focal consolidation, effusion, or pneumothorax. the cardiomeastinal silhouette is normal. imaged osseous structures are intact.

TieNet

pa and lateral views of the chest. there is mild enlargement of the cardiac silhouette. there is no pleural effusion or pneumothorax. there is no acute osseous abnormalities.



as compared to the previous radiograph, the monitoring and support devices are unchanged. unchanged bilateral pleural effusions, with a tendency to increase, and resultant areas of atelectasis. the air collection in the bilateral soft tissues is slightly decreased. unchanged right picc line. no definite evidence of pneumothorax.

as compared to the previous radiograph, the patient has received a nasogastric tube. the course of the tube is unremarkable, the tip of the tube projects over the middle parts of the stomach. there is no evidence of complication, notably no pneumothorax. the other monitoring and support devices are constant. constant appearance of the cardiac silhouette and of the lung parenchyma.

as compared to the previous radiograph, there is no relevant change. tracheostomy tube is in place. there is a layering pleural effusions. NAME bilateral pleural effusion and compressive atelectasis at the right base. there is no pneumothorax.

Obviously our output is better than the previous work.

# Oh Wait, I Got That Backwards.



cardiomegaly is moderate. bibasilar atelectasis is mild. there is no pneumothorax. a lower cervical spinal fusion is partially visualized. healed right rib fractures are incidentally noted.

ap portable upright view of the chest. there is no focal consolidation, effusion, or pneumothorax. the cardiomeastinal silhouette is normal. imaged osseous structures are intact.

pa and lateral views of the chest. there is mild enlargement of the cardiac silhouette. there is no pleural effusion or pneumothorax. there is no acute osseous abnormalities.



as compared to the previous radiograph, the monitoring and support devices are unchanged. unchanged bilateral pleural effusions, with a tendency to increase, and resultant areas of atelectasis. the air collection in the bilateral soft tissues is slightly decreased. unchanged right picc line. no definite evidence of pneumothorax.

as compared to the previous radiograph, the patient has received a nasogastric tube. the course of the tube is unremarkable, the tip of the tube projects over the middle parts of the stomach. there is no evidence of complication, notably no pneumothorax. the other monitoring and support devices are constant. constant appearance of the cardiac silhouette and of the lung parenchyma.

as compared to the previous radiograph, there is no relevant change. tracheostomy tube is in place. there is a layering pleural effusions. NAME bilateral pleural effusion and compressive atelectasis at the right base. there is no pneumothorax.

Ours is still better, though... 😄



# Wow! It Is Doing Great!

| Model       | Natural Language |              |              |              |              |              | Clinical     |
|-------------|------------------|--------------|--------------|--------------|--------------|--------------|--------------|
|             | CIDEr            | ROUGE        | BLEU-1       | BLEU-2       | BLEU-3       | BLEU-4       | Accuracy     |
| Ours (NLG)  | <b>1.153</b>     | <b>0.307</b> | <b>0.352</b> | <b>0.223</b> | <b>0.153</b> | <b>0.104</b> | 0.834        |
| Ours (CCR)  | 0.956            | 0.284        | 0.294        | 0.190        | 0.134        | 0.094        | <b>0.868</b> |
| Ours (full) | 1.046            | <b>0.306</b> | 0.313        | 0.206        | 0.146        | <b>0.103</b> | <b>0.867</b> |

Probably...

What does a CIDEr score of 1.15 mean?

# NLP Language Generation Eval Metrics

- BLEU (ngrams)
- CIDEr (ngrams)
- METEOR (ngrams+soft sim)
- ROUGE (ngrams)
- TER / TERp (edit distance)
- BADGER (compression)
- etc

|                               |             |             |
|-------------------------------|-------------|-------------|
| HYPOTHESIS: I went for a walk | BADGER=0.88 | BLEU2=0.75  |
|                               | TERp=0.31   |             |
| REFERENCE: I went for a swim  | BLEU3=0.67  | METEOR=0.36 |

# BLEU: bilingual evaluation understudy (Papineni, 2001)

- **Purpose:** Early attempt for automatic Machine Translation

- **Operationalizing:**

- Precision metric; “modified n-gram precision”
- Metric: Geometric mean of the n-gram precisions
- Formula: left; where:  $Count_{clip} = \min(Count, Max\_Ref\_Count)$

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-gram} \in C} Count_{clip}(n\text{-gram})}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-gram}' \in C'} Count(n\text{-gram}')}.$$

- **Details (Ignore for Intuition, but matter for Implementation):**

- “Modification”: “clipped” count
- Overcoming inflated precisions due to sentence length: multiplicative brevity penalty factor.

- **Intuitions:**

- Because it was an early attempt, it is both (1) simple and (2) widely used
- Drawbacks: More references produces higher scores. (Should not compare scores for candidates evaluated on corpora of different lengths.)

# CIDEr: Consensus-based Image Description Evaluation (Vedantam, 2014)

- **Purpose:** Image description evaluation that correlates with human judgement
  - Motivation: Develop an evaluation protocol that is “human-like”, since using humans in studies is expensive and timely.

- **How It Works:**

- Each sentence/document is represented with tf-idf n-grams
- computed using the average cosine similarity between the candidate sentence and the reference sentences
- Formula:

$$\text{CIDEr}_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{\mathbf{g}^n(c_i) \cdot \mathbf{g}^n(s_{ij})}{\|\mathbf{g}^n(c_i)\| \|\mathbf{g}^n(s_{ij})\|},$$

- **Intuition:**

- Assumes many (20-50) reference sentences per image/prediction. The many references is what creates the “consensus”

# Scores Revisited

The Real Question: But is a CIDEr score of 1.15 good? 🤔

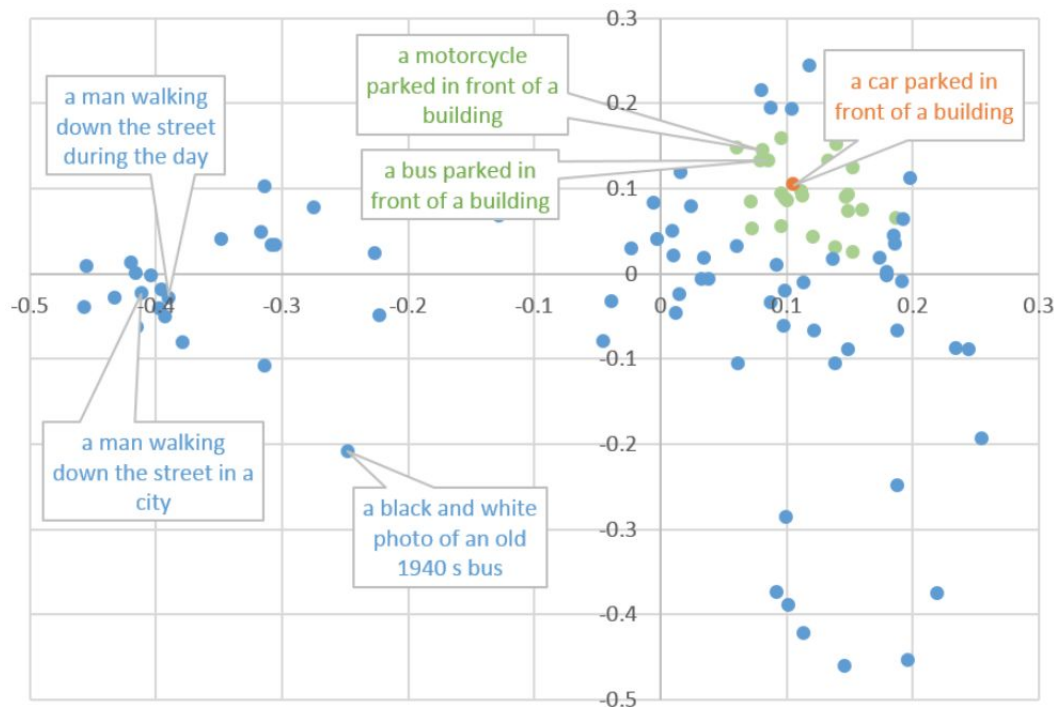
| Model       | Natural Language |              |              |              |              |              | Clinical     |
|-------------|------------------|--------------|--------------|--------------|--------------|--------------|--------------|
|             | CIDEr            | ROUGE        | BLEU-1       | BLEU-2       | BLEU-3       | BLEU-4       | Accuracy     |
| Ours (NLG)  | <b>1.153</b>     | <b>0.307</b> | <b>0.352</b> | <b>0.223</b> | <b>0.153</b> | <b>0.104</b> | 0.834        |
| Ours (CCR)  | 0.956            | 0.284        | 0.294        | 0.190        | 0.134        | 0.094        | <b>0.868</b> |
| Ours (full) | 1.046            | <b>0.306</b> | 0.313        | 0.206        | 0.146        | <b>0.103</b> | <b>0.867</b> |

# Baseline: Random



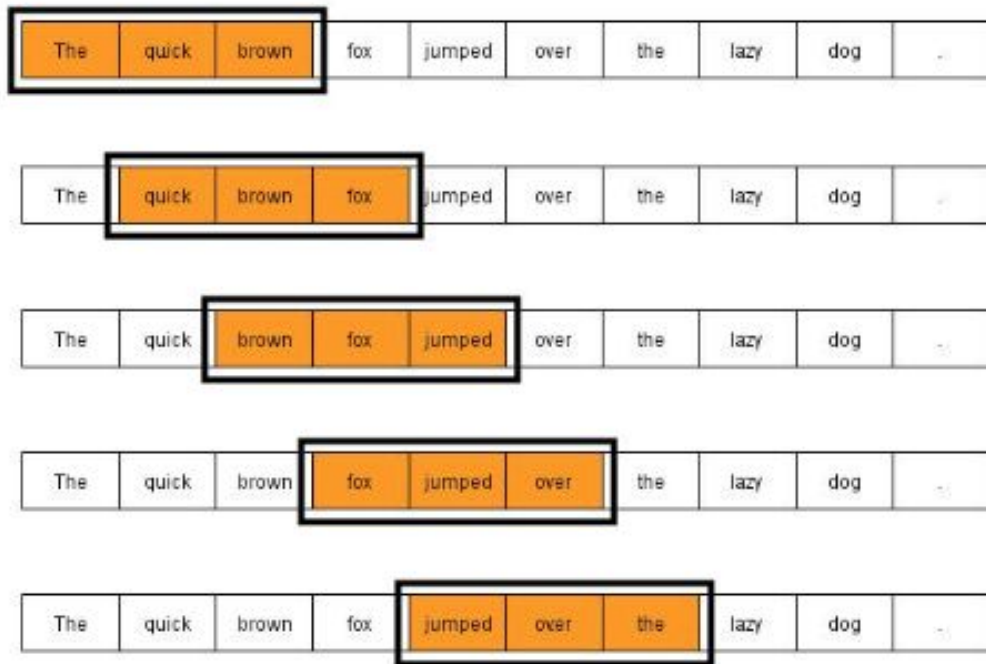
- Pick some random image from the train set.
- Return the caption of that image.
- Expectation: This will be irrelevant but grammatical

# Baseline: Nearest Neighbor



- Extract features from image using ConvNet.
- Find the closest image from the train set.
- Return the caption of that image.
- Expectation: This will be grammatical and sort of relevant.

# Baseline: 3-gram




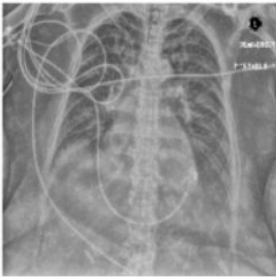
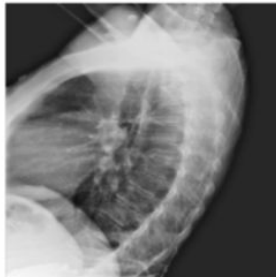
- Extract features from image using ConvNet
- Find the 100 closest images from the train set
- Fit a trigram from those similar reports.
- Generate a new report from that language model.
- Expectation: potentially ungrammatical, but specific to the image.



# Examples

Hard to read, but we can “generate” examples using each method for every image.

We know that the random method is bad. But is it always easy to tell whether KNN or 3-gram is “better” ?

| Image     |    |    |   |
|-----------|---|--|---|
| Reference | pa and lateral views of the chest demonstrate the lungs are well-expanded and clear. the cardiomeastinal silhouette is normal. there is no pleural effusion or pneumothorax.  | in comparison with the study of DATE, the monitoring and support devices are in essentially unchanged position. there is again mild enlargement of the cardiac silhouette with pulmonary edema and bilateral layering pleural effusions.   | the patient was imaged in a lordotic position, which distorts the mediastinal contours. within that limitation, the lungs are clear without consolidation or edema. the mediastinum is otherwise unremarkable. the cardiac silhouette is within normal limits for size. no effusion or pneumothorax is noted. no displaced fractures are evident. |
| 3-gram    | pa and lateral views of the chest . there is no pleural effusion , or pleural effusion or pneumothorax . the cardiomeastinal silhouette is within normal limits . lungs are essentially clear . no acute osseous abnormality . levoconvex scoliosis of the chest were obtained . low lung volumes . there are no pleural effusion or pneumothorax is seen . the mediastinal and hilar contours are normal . | et tube , enteric tube tip is difficult to assess the status of the intra-aortic balloon pump , with mild increase in pulmonary outflow tract remain unchanged . at the level of the exam is a moderate left-sided pleural effusion with bibasilar <u>pelural</u> fluid and atelectasis . there are no acute bony abnormality .  | the lungs are mildly hyperinflated but clear , the cardiomeastinal silhouette is normal . mediastinal and hilar contours are normal . imaged osseous structures are intact .  |
| KNN       | left pleural tube is in stable position. there has been a slight increase in the left pleural effusion with increased atelectasis at the left base. there is a stable left apical pneumothorax and atelectasis at the right base. cardiomeastinal and hilar contours are stable. there is no focal consolidation concerning for pneumonia.  | on the first radiograph, obtained at 1249, there was malposition of the <u>dobbhoff</u> catheter in the right bronchial system. no evidence of pneumothorax or other complications. on the radiograph performed at 1255, the <u>dobbhoff</u> catheter follows the course of the esophagus, with the tip in the proximal parts of the stomach. again, no complication such as pneumothorax is seen. | the lungs are well inflated and clear. the cardiomeastinal silhouette, hilar contours, and pleural surfaces are normal. there is no pleural effusion or pneumothorax.   |

# The Crux: The Evaluation Metrics are Crummy

| Model  | BLEU-1       | BLEU-2  | BLEU-3  | BLEU-4  | CIDEr | CheXpert Accuracy | CheXpert Precision | CheXpert F1  |
|--------|--------------|---------|---------|---------|-------|-------------------|--------------------|--------------|
| Random | 0.265        | 0.137   | 0.070   | 0.036   | 0.570 | 0.770             | 0.146              | 0.148        |
| 1-gram | 0.196        | < 0.001 | < 0.001 | < 0.001 | 0.348 | 0.742             | 0.206              | 0.174        |
| 2-gram | 0.194        | 0.098   | 0.043   | 0.013   | 0.404 | 0.764             | 0.225              | 0.193        |
| 3-gram | 0.206        | 0.107   | 0.057   | 0.031   | 0.435 | 0.782             | 0.225              | 0.185        |
| KNN    | <b>0.305</b> | 0.171   | 0.098   | 0.057   | 0.755 | 0.818             | 0.253              | <b>0.258</b> |

(expected) The CheXpert scores of trigrams are better than the random baseline.

**(Surprising!) Random sentences achieve better BLEU and CIDEr scores than generating sentences with trigrams.**

BLEU and CIDEr care more about grammaticality (and other surface level similarities) than about whether the text is correctly describing the x-ray.

# CheXpert Sentence Labeler

Not bad. We have some quibbles with it (e.g. rule-based, negation is hard, doesn't capture uncertainty, etc), but by and large it measures clinical accuracy.

Should we care about readability at all? Is word salad sufficient?

|   | Observation       | Labeler Output |
|---|-------------------|----------------|
| 1. <i>unremarkable</i> <u>cardiomediastinal silhouette</u>  | No Finding        | 0              |
|   | Enlarged Cardiom. |                |
|   | Cardiomegaly      |                |
|   | Lung Opacity      |                |
| 2. diffuse <u>reticular pattern</u> , which can be seen with an atypical <u>infection</u> <b>or</b> chronic fibrotic change. <i>no</i> focal <u>consolidation</u> . | Lung Lesion       | 1              |
|   | Edema             |                |
|   | Consolidation     |                |
|   | Pneumonia         |                |
| 3. <i>no</i> <u>pleural effusion</u> or <u>pneumothorax</u>   | Atelectasis       | 0              |
|   | Pneumothorax      |                |
|   | Pleural Effusion  |                |
|   | Pleural Other     |                |
| 4. mild degenerative changes in the lumbar spine and old right rib <u>fractures</u> .   | Fracture          | 1              |
|   | Support Devices   |                |
|   |                   |                |

# But I Thought CIDEr Was Really Good... ?

Maybe for describing YouTube videos.

Maybe when you have 20-50 reference sentences per image.



**A cow is standing in a field.**

**A cow with horns and long hair covering its face stands in a field.**

**A cow with hair over its eyes stands in a field.**

This horned creature is getting his picture taken.

A furry animal with horns roams on the range.



**Mike has a baseball and Jenny has a basketball.**

**Jenny is holding a basketball and Mike is holding a baseball.**

**Jenny is playing with a basketball and Mike is playing with a baseball.**

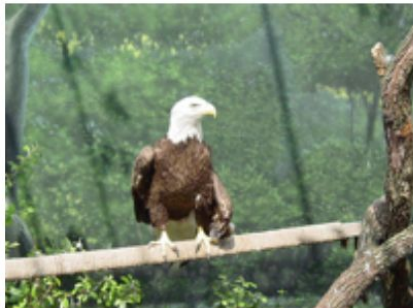
Jenny brought a bigger ball than Mike.

Mike is sad that Jenny is leaving in five days.

Figure 1: Images from our PASCAL-50S (left) and ABSTRACT-50S (right) datasets with a subset of corresponding (human) sentences. Sentences shown in **bold** are representative of the consensus descriptions for these images. We propose to capture such descriptions with our evaluation protocol.

# How Are Metrics “Validated” ?

1. Create some new, “better” metric
2. Have humans annotate some task
3. Show that your metric agrees with human judgment better than the other metrics (e.g. your metric ranks the sentences better than other metrics)

|   |   |   |
|---|---|---|
|  | <p><b>Reference Sentences</b></p> <p><b>R1:</b> A bald eagle sits on a perch.</p> <p><b>R2:</b> An american bald eagle sitting on a branch in the zoo.</p> <p><b>R3:</b> Bald eagle perched on piece of lumber.</p> <p>...</p> <p><b>R50:</b> A large bird standing on a tree branch.</p> | <p><b>Candidate Sentences</b></p> <p><b>C1:</b> An eagle is perched among trees.</p> <p><b>C2:</b> A picture of a bald eagle on a rope stem.</p> <p><b>Triplet Annotation</b></p> <p><i>Which of the sentences, B or C, is more similar to sentence A?</i></p> <p><b>Sentence A :</b> Anyone from R1 to R50</p> <p><b>Sentence B :</b> C1</p> <p><b>Sentence C :</b> C2</p> |
|---|---|---|

(a) (b) (c)

Figure 2: Illustration of our triplet annotation modality. Given an image (a), with reference sentences (b) and a pair of candidate sentences (c, top), we match them with a reference sentence one by one to form triplets (c, bottom). Subjects are shown these 50 triplets on Amazon Mechanical Turk and asked to pick which sentence (B or C) is more similar to sentence A.

# What Do We Need?

## Clinical judgments!

That might mean a few different things.

- We have some thoughts/ideas to propose.
- We'd love your feedback of other things we could be looking at.

Ultimately, we want:

1. Clinicians to make some sort of annotation which boils down to saying “this generated report is better than that generated report.”
2. We look at how well each metric (e.g. BLEU, CIDEr, etc) agrees with those rankings.
3. We make a better metric (e.g. using clinical concepts) that better agrees with docs.

## Naive Idea: Likert Scale

“1-to-10, how good is this generated report for this image?”

## Problems (Humans are inconsistent):

- Your 7 is my 5.
- Your 4 before lunch might be a 7 after lunch.



the patient was imaged in a lordotic position, which distorts the mediastinal contours. within that limitation, the lungs are clear without consolidation or edema. the mediastinum is otherwise unremarkable. the cardiac silhouette is within normal limits for size. no effusion or pneumothorax is noted. no displaced fractures are evident.

### How good is the generated report for this image?

the worst

(1)

○

(2)

○

(3)

O

(4)

Q

(5)

C

(6)

O

(7)

○

(8)

Q

perfect

(9)

C

# Potential Annotation: Ranking

Here is an image and 4 generated reports trying to describe that image. Rank them by how good each report is.

Benefit:

- People are better at ranking A vs B than trying to assign “goodness” scores to A and B separately.



Rank

1

B) as compared to the previous radiograph , there has been no significant interval change in the right ij catheter . there is no evidence of a right pleural effusions . no longer visualized .

3

C) cardiac , mediastinal and hilar contours are normal . pulmonary vasculature is normal . apart from subsegmental atelectasis in the left lower lobe , the lungs are clear . no focal consolidation , pleural effusion or pneumothorax is present .

2

D) the , portable left has to be the side borderline size now the wires filling left . extensive in lucent is partial . interstitial pulmonary cm congestion with cephalad chest the have evidence since fluid right developing and constant , and tip

4

E) as the lungs .



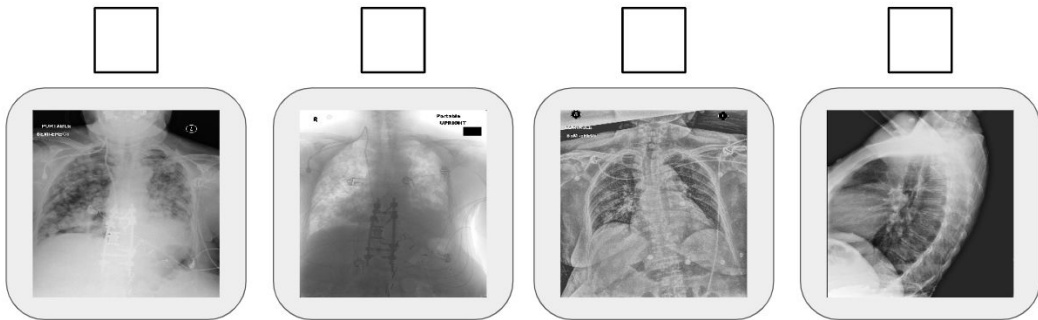
# Potential Annotation: Content Selection

Here is a generated report and four images. Which image do you think the report was trying to describe?

as compared to the previous radiograph , there has been no significant interval change in the right ij catheter . there is no evidence of a right pleural effusions . no longer visualized .

Benefit:

- Doesn't leave any room for "I didn't like this sentence because it had bad grammar"



# Potential Annotation: Edit Distance for Radiologist

Populate the report with the generated text and compute the edit distance between the draft and the final report.



## Starting Draft for Radiologist

cardiac and hilar contours are normal . pulmonary vasculature is abnormal . apart from subsegmental atelectasis in the left lower lobe , the lungs are clear . pleural effusion or pneumothorax is present .

## Final Radiologist-Written Report

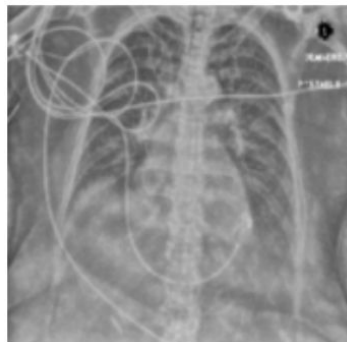
cardiac , mediastinal , and hilar contours are normal . pulmonary vasculature is ~~abnormal~~ normal . apart from subsegmental atelectasis in the left lower lobe , the lungs are clear . no focal consolidation , - pleural effusion or pneumothorax is present .

Benefit:

- Easy to measure/compute.
- Translates naturally to how doing well would result in a bonafide improvement for radiologists.

# Potential Annotation: Simulated Use Case

We might care more about how the report is used (as opposed to how easy it is to generate).



## To ED Doc:

When you order a Chest X-Ray, and this report comes back, what decision would you make in response?

In comparison with the study of DATE , the monitoring and support devices are in essentially unchanged position . there is again mild enlargement of the cardiac silhouette with pulmonary edema and bilateral layering pleural effusions .



Intervention 1



Intervention 2



Intervention 3



Intervention 4

Benefit:

- The ultimate goal of ordering the Chest X-Ray is to impact care in some way; how would the ordering doctor actually use the information?
- What aspects of the generated report do/don't change how the doctors respond?

# Next Steps

We are working with some radiologists to decide what annotation task would be best for evaluating whether our models are getting better.

They will then provide us with annotations & we will devise a new eval metric which agrees with their judgments better than BLEU or CIDEr do.

UToronto is also **interested in deploying the pipeline of CXR technologies** (starting with image classification and potentially including report generation).

- This could allow for A/B testing as the gold standard for which models work best.
- We can then use that data to see how much you can trust our metric to generalize to performance (at least within their hospital system).
- Interesting: do radiologists & ED docs agree on what makes a “good” report?

# Case Study 2: Psychiatric Readmission from Notes



DATE OF ADMISSION: MM/DD/YYYY DATE OF DISCHARGE: MM/DD/YYYY

**DISCHARGE (DIAGNOSES):**  
AXIS I: Bipolar disorder, depressed, with psychotic features, symptoms in remission.  
AXIS II: Delivered. AXIS III: None. AXIS IV: Moderate.  
AXIS V: Global assessment of functioning 65 on discharge.

**REASON FOR ADMISSION:** The patient was admitted with a chief complaint of suicidal ideation. The patient was brought to the hospital after his guidance counselor found a note the patient wrote, which detailed who he was giving away his possessions to if he dies. Three weeks ago, he got pushed into a corner at school and threatened to shoot himself and others with a gun. The patient was suspended for that remark.

**PROCEDURES AND TREATMENT:**  
1. Individual and group psychotherapy.  
2. Psychopharmacologic management.  
3. Family therapy conducted by social work department with the patient and the patient's family.

**HOSPITAL COURSE:** The patient responded well to individual and group psychotherapy, milieu therapy and medication management. As stated, family therapy was conducted.

**DISCHARGE ASSESSMENT:** At the time of discharge, the patient is alert and fully oriented. Mood euthymic. Affect broad range. He denies any suicidal or homicidal ideation. IQ is at baseline. Memory intact. Insight and judgment good.

**PLAN:** The patient may be discharged as he no longer poses a risk of harm towards himself or others. The patient will continue on the following medications: Miltaine LA 60 mg q.a.m. The patient will follow up with Dr. Doe for medication management and Dr. Smith for psychotherapy. All other discharge orders per the psychiatrist, as arranged by social work.



Model

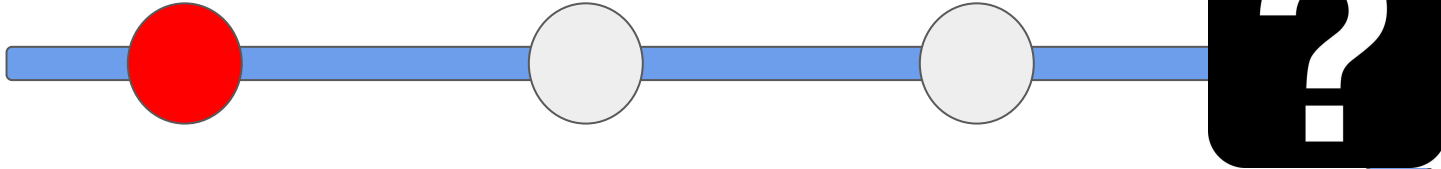
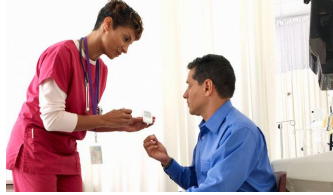


180-day Readmission?

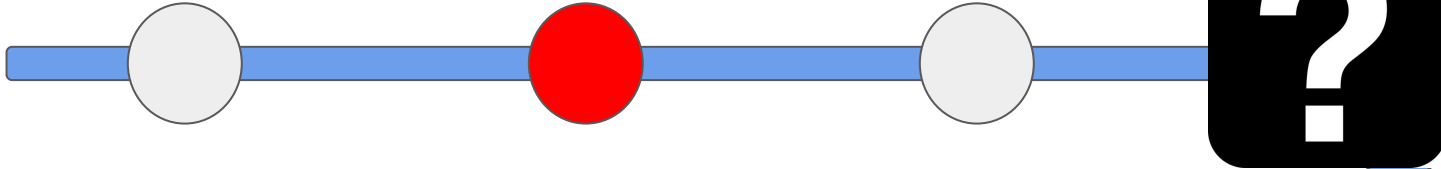
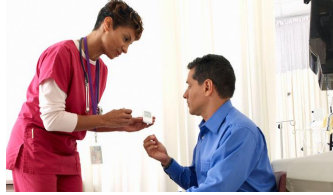


MASSACHUSETTS  
GENERAL HOSPITAL

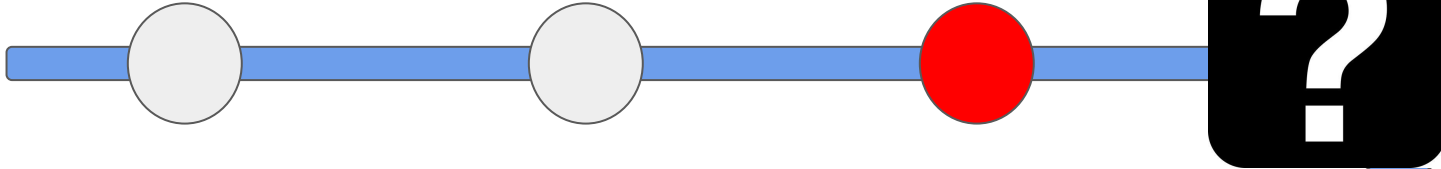
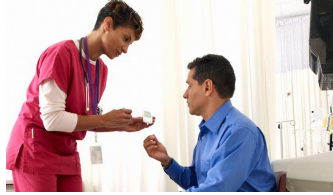
# The Context: Readmission To What?



# The Context: Readmission To What?



# The Context: Readmission To What?





# Readmission problem

- Readmissions are costly and often indicate someone at high-risk was sent home too soon.
- 40-50% of psychiatric patients are readmitted within one year
  - 45% in MGH dataset

**Goal: Can we identify high-risk patients before they are sent home too soon?**

If so, their doctors might be able to get them the resources they need to prevent cases from becoming much larger problems down the line.

# An Example Note

**DATE OF ADMISSION:** MM/DD/YYYY

**DATE OF DISCHARGE:** MM/DD/YYYY

**DISCHARGE DIAGNOSES:**

AXIS I: Bipolar disorder, depressed, with psychotic features, symptoms in remission.

AXIS II: Deferred.

AXIS III: None.

AXIS IV: Moderate.

AXIS V: Global assessment of functioning 65 on discharge.

**REASON FOR ADMISSION:** The patient was admitted with a chief complaint of suicidal ideation. The patient was brought to the hospital after his guidance counselor found a note the patient wrote, which detailed who he was giving away his possessions to if he dies. Three weeks ago, he got pushed into a corner at school and threatened to shoot himself and others with a gun. The patient was suspended for that remark.

**PROCEDURES AND TREATMENT:**

1. Individual and group psychotherapy.
2. Psychopharmacologic management.
3. Family therapy conducted by social work department with the patient and the patient's family.

**HOSPITAL COURSE:** The patient responded well to individual and group psychotherapy, milieu therapy and medication management. As stated, family therapy was conducted.

**DISCHARGE ASSESSMENT:** At the time of discharge, the patient is alert and fully oriented. Mood euthymic. Affect broad range. He denies any suicidal or homicidal ideation. IQ is at baseline. Memory intact. Insight and judgment good.

**PLAN:** The patient may be discharged as he no longer poses a risk of harm towards himself or others. The patient will continue on the following medications; Ritalin LA 60 mg q.a.m., The patient will follow up with Dr. Doe for medication management and Dr. Smith for psychotherapy. All other discharge orders per the psychiatrist, as arranged by social work.

# Results

These are the AUCs... but are we doing a good job?

|                                | LR   | SVC  | XGB  | MLP  |
|--------------------------------|------|------|------|------|
| Demographics                   | 0.67 | 0.67 | 0.68 | 0.66 |
| TF-IDF                         | 0.68 | 0.68 | 0.68 | 0.63 |
| LDA-75                         | 0.69 | 0.68 | 0.67 | 0.67 |
| TF-IDF + demographics          | 0.69 | 0.69 | 0.69 | 0.63 |
| LDA-75 + demographics          | 0.7  | 0.7  | 0.69 | 0.67 |
| TF-IDF + LDA-75 + demographics | 0.7  | 0.7  | 0.69 | 0.63 |

**Table 3.** Overall 180-day readmission performance.

# Results

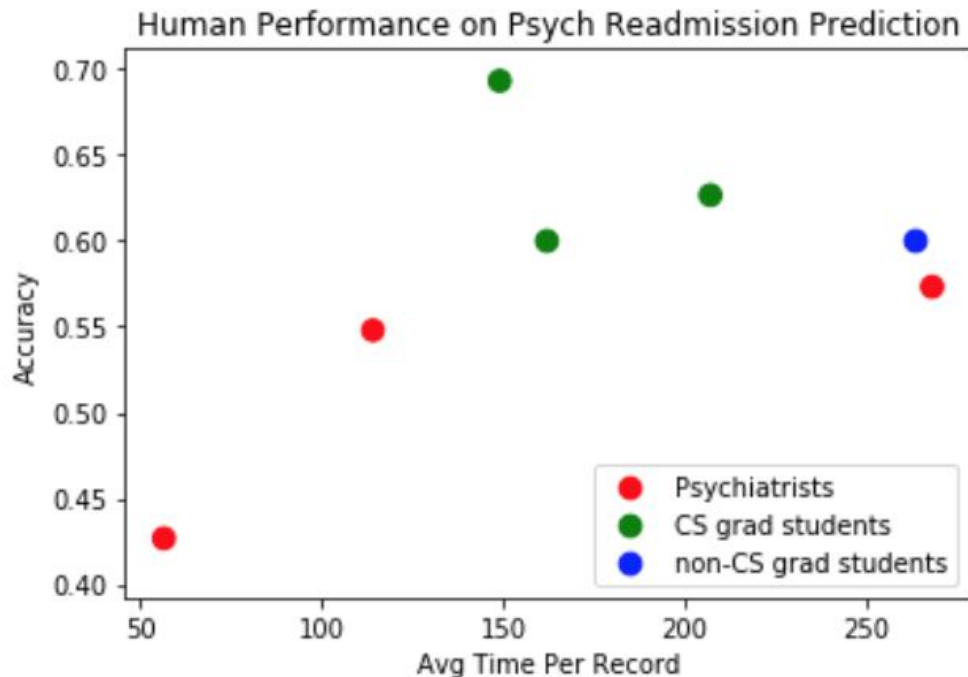
These are the AUCs... but are we doing a good job?

|                                | LR   | SVC  | XGB  | MLP  |
|--------------------------------|------|------|------|------|
| Demographics                   | 0.67 | 0.67 | 0.68 | 0.66 |
| TF-IDF                         | 0.68 | 0.68 | 0.68 | 0.63 |
| LDA-75                         | 0.69 | 0.68 | 0.67 | 0.67 |
| TF-IDF + demographics          | 0.69 | 0.69 | 0.69 | 0.63 |
| LDA-75 + demographics          | 0.7  | 0.7  | 0.69 | 0.67 |
| TF-IDF + LDA-75 + demographics | 0.7  | 0.7  | 0.69 | 0.63 |

**Table 3.** Overall 180-day readmission performance.

**Red Flag:** There's so little spread between just demographics info (.67) and the beefiest model (0.70).

# How Well Do Humans Do On The Task?



7 human annotators tried reading a note and guessing whether 180d readmission.

**(Surprising!)** Grad students were better at predicting readmission than Psychiatrists were (lol).

But how would ML do  
On the same 75 notes?

# How Well Do Humans Do On The Task?

**Table 5.** Annotator performance for baseline predictions and learning phase.

| Annotator          | Baseline<br>Balanced<br>Accuracy | Baseline<br>F1 | Baseline<br>Speed<br>(sec/note) | Learning<br>Balanced<br>Accuracy | Learning<br>F1 | Learning<br>Speed<br>(sec/note) | Total<br>Balanced<br>Accuracy | Total<br>F1  | Total<br>Speed<br>(sec/note) |
|--------------------|----------------------------------|----------------|---------------------------------|----------------------------------|----------------|---------------------------------|-------------------------------|--------------|------------------------------|
| Grad Student 1     | 0.597                            | 0.471          | 249                             | 0.490                            | 0.457          | 119                             | 0.530                         | 0.348        | 162                          |
| Grad Student 2     | 0.615                            | 0.545          | 158                             | <b>0.776</b>                     | <b>0.649</b>   | 144                             | <b>0.718</b>                  | <b>0.610</b> | 149                          |
| Grad Student 3     | 0.503                            | 0.400          | 170                             | 0.669                            | 0.545          | 287                             | 0.610                         | 0.481        | 207                          |
| Grad Student 4     | <b>0.632</b>                     | <b>0.593</b>   | 315                             | 0.605                            | 0.457          | 237                             | 0.627                         | 0.516        | 263                          |
| Psychiatrist 1     | 0.581                            | 0.522          | 170                             | 0.565                            | 0.432          | 167                             | 0.573                         | 0.459        | 114                          |
| Psychiatrist 2     | 0.441                            | 0.364          | 65                              | 0.401                            | 0.256          | 52                              | 0.417                         | 0.295        | 56                           |
| Psychiatrist 3     | 0.455                            | 0.250          | 368                             | 0.486                            | 0.231          | 218                             | 0.477                         | 0.238        | 268                          |
| non-MD (avg)       | <b>0.587</b>                     | <b>0.502</b>   | 223                             | <b>0.635</b>                     | <b>0.527</b>   | 197                             | <b>0.621</b>                  | <b>0.489</b> | 195                          |
| Psychiatrist (avg) | 0.492                            | 0.379          | 201                             | 0.484                            | 0.306          | 146                             | 0.489                         | 0.331        | 146                          |

# How Well Do Rules/ML Do On The Task?

|                                 |                   |       |
|---------------------------------|-------------------|-------|
| Reminder:                       | Balanced Accuracy | F1    |
| Top individual human:           | 0.718             | 0.610 |
| Top individual human (learning) | 0.776             | 0.649 |
| Non-MD Average:                 | 0.621             | 0.489 |

**Table 6.** Rule-based and ML-based performance on the same 75 patients (for comparison against humans).

| Paradigm | Method                              | Accuracy     | F1           |
|----------|-------------------------------------|--------------|--------------|
| n/a      | Random (1000 trials)                | 0.498        | 0.378        |
| Rule     | Always No-Readmit                   | 0.693        | 0.000        |
| Rule     | Always Readmit                      | 0.307        | 0.469        |
| Rule     | Num Prior Admits > 12               | 0.693        | 0.531        |
| Rule     | Num Psych Dx in Prior 12 months > 5 | 0.667        | <b>0.603</b> |
| ML       | SVM (demographics)                  | 0.689        | 0.380        |
| ML       | SVM (demographics+tfidf)            | <b>0.732</b> | 0.454        |
| ML       | SVM (demographics+tfidf+LDA)        | 0.692        | 0.561        |

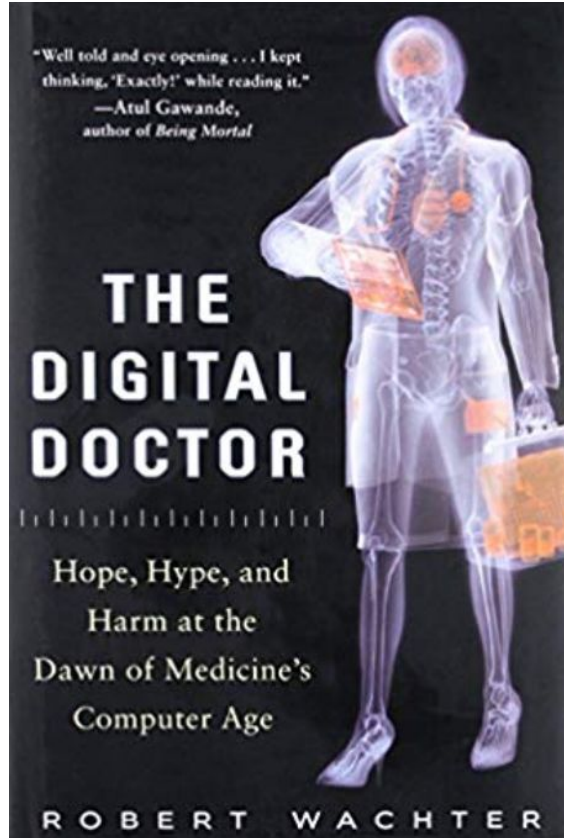
Note: Large difference between ML models when we look at F1 instead of AUC.

# Evaluating Health

- “Health” is tricky, because the best case scenario is that no one needs to come to the hospital.
- You can evaluate based on process or outcomes.
- Last week, we saw how hard it is to know if you’re treating sepsis well.



# Meaningful Use



The HITECH Act (2009) paid \$30B to incentivize hospitals to go digital.

The EHR needed to be “certified”, which was a low bar of boxes to tick.

This book by Dr. Wachter looks at some of the problems doctors face from EHRs created without focusing on the user experience:

- Alarm fatigue
- Notes copy/paste
- Constant Interruption / Context Switching
- Trusting typos because the computer “must be right”

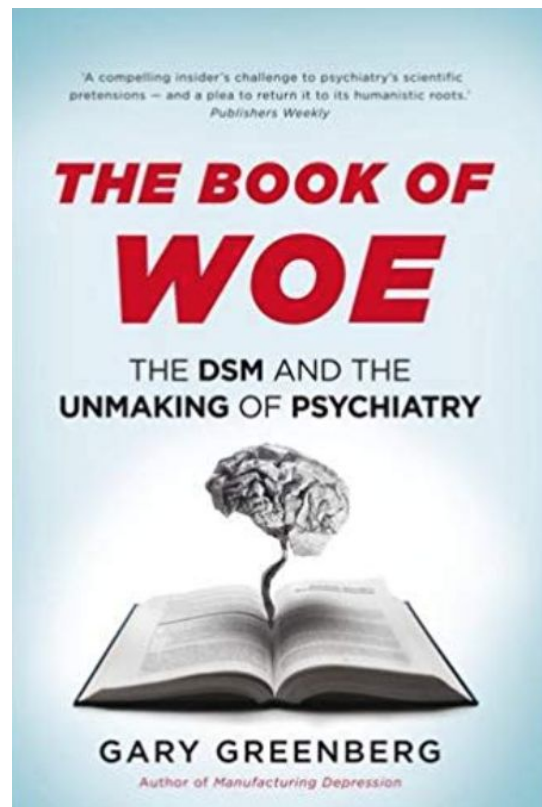
# Mental Health

The modern history of psychiatry shows that we're trying our best but are really not yet able to say we have a good understanding.

Unlike with physical health, there are almost no mental health disorders that have a biomarker which let us confirm correctness.

The standard DSM rules are the result of some committee of people reading papers and making their best guesses.

*Disclaimer: Haven't read this book in a year. Don't remember it as well.*



# Obesity

BMI screening followup

“We have to measure these things because it is important. The measurement will get better over time.”

->

“This is too hard to measure. We would like to track obesity, but we don’t have a good way to measure it, so we can’t.”

# Process vs Outcomes

|     | Outcomes  | Process   |
|-----|---|---|
| Pro | <ul style="list-style-type: none"><li>- Less ambiguous to measure.</li><li>- Ultimately one of the main goals of care.</li></ul>  | <ul style="list-style-type: none"><li>- “Partial credit” if you did a good job but still got a bad outcome (i.e. shouldn’t penalize you for being unlucky).</li></ul> |
| Con | <ul style="list-style-type: none"><li>- Attributing impact under a rare/sparse event (e.g. thousands of decisions are made before the patient dies/survives).</li></ul> | <ul style="list-style-type: none"><li>- It’s hard to know that the established process is correct/best.</li></ul>   |

# Thanks!

Anonymous Feedback appreciated!

<https://whatiswrongwith.me/willie>



### Starting Draft for Radiologist

cardiac and hilar contours are normal . pulmonary vasculature is abnormal . apart from subsegmental atelectasis in the left lower lobe , the lungs are clear . pleural effusion or pneumothorax is present .

### Final Radiologist-Written Report

cardiac , **mediastinal** , and hilar contours are normal . pulmonary vasculature is ~~abnormal~~ **normal** . apart from subsegmental atelectasis in the left lower lobe , the lungs are clear . **no focal consolidation** , **-** pleural effusion or pneumothorax is present .



## To ED Doc:

**When you order a Chest X-Ray, and this report comes back, what decision would you make in response?**

In comparison with the study of DATE , the monitoring and support devices are in essentially unchanged position . there is again mild enlargement of the cardiac silhouette with pulmonary edema and bilateral layering pleural effusions .



Intervention 1



Intervention 2



Intervention 3



Intervention 4