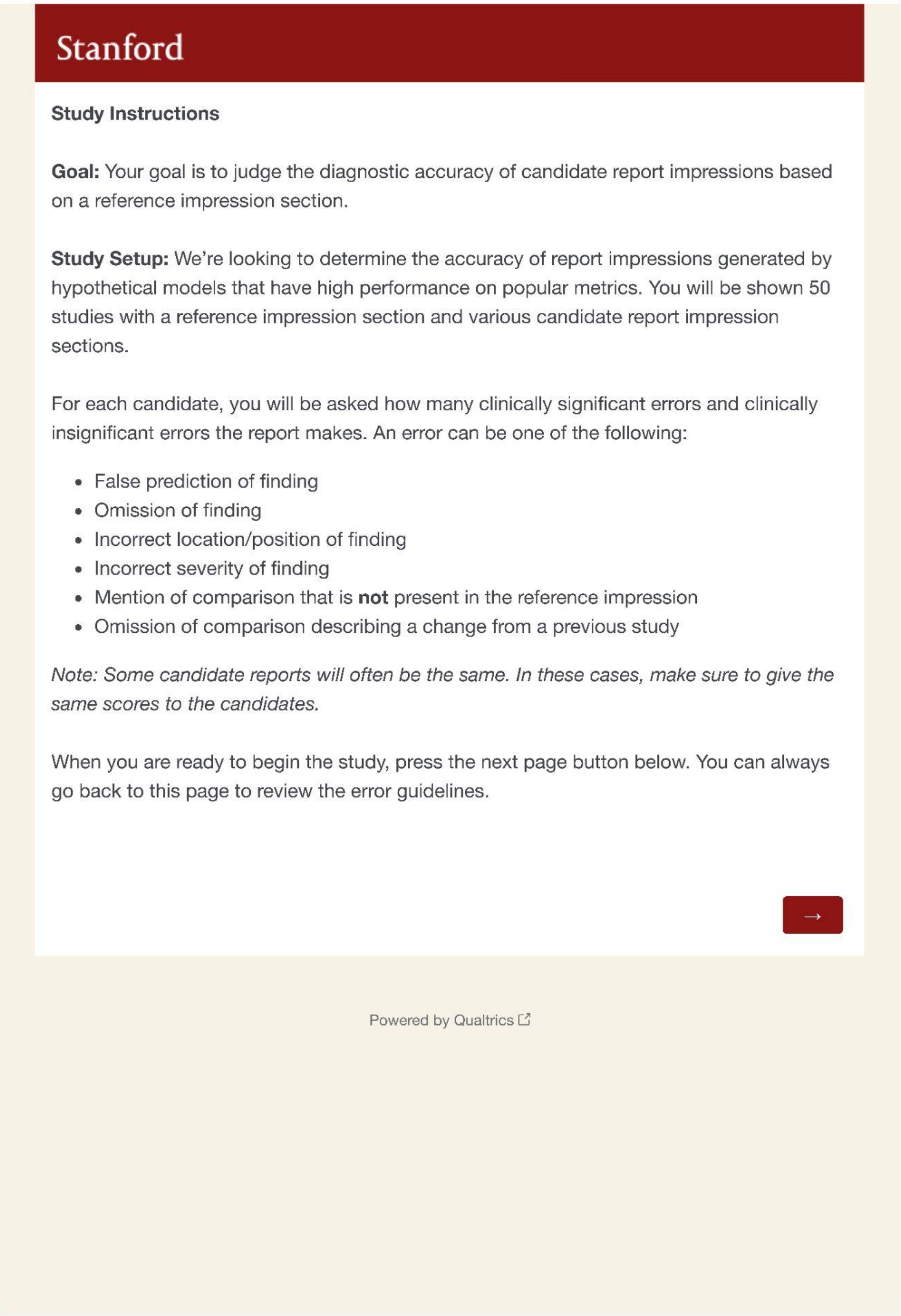
Patterns, Volume *4*

Supplemental information

Evaluating progress in automatic chest X-ray radiology report generation

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar

Supplemental information



Study Instructions

Goal: Your goal is to judge the diagnostic accuracy of candidate report impressions based on a reference impression section.

Study Setup: We're looking to determine the accuracy of report impressions generated by hypothetical models that have high performance on popular metrics. You will be shown 50 studies with a reference impression section and various candidate report impression sections.

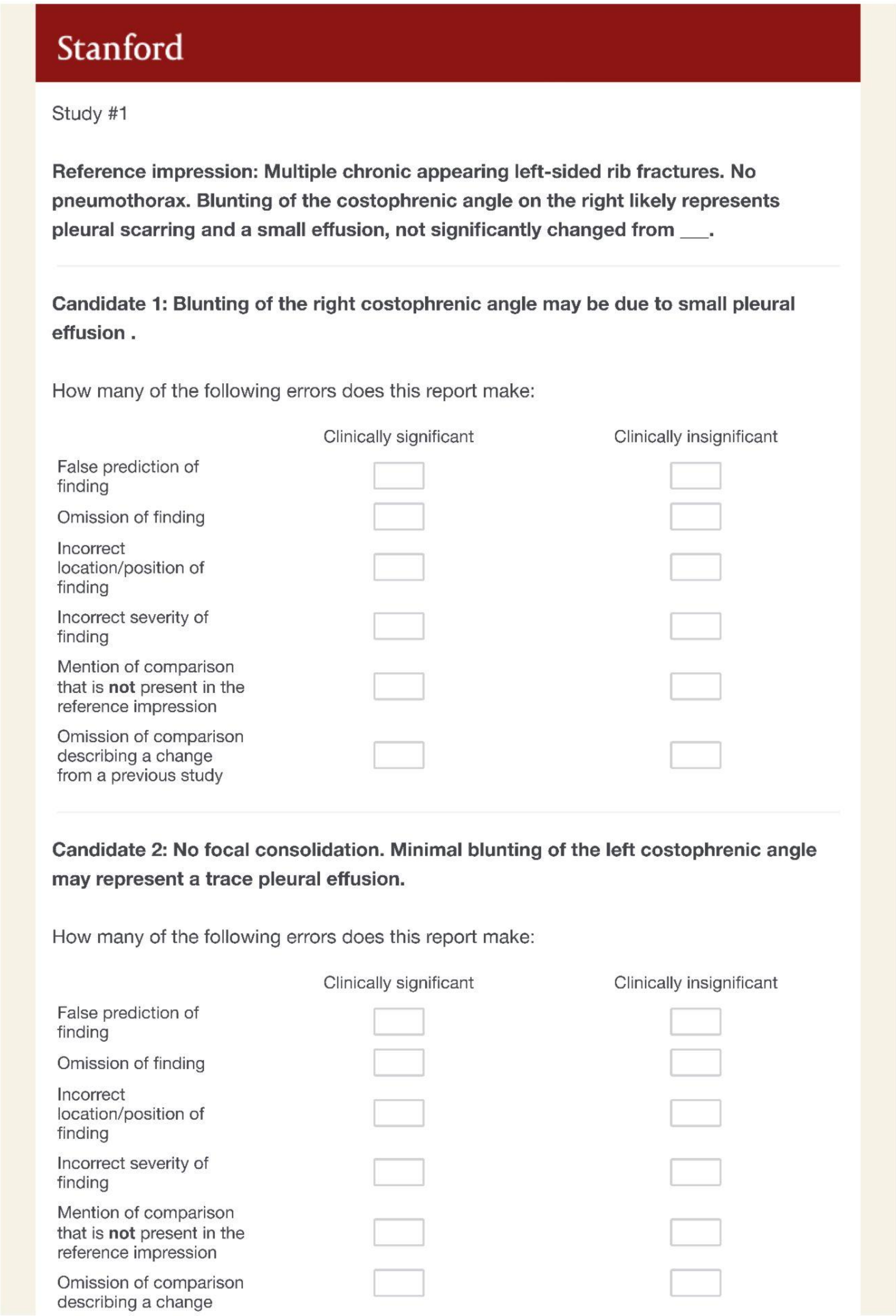
For each candidate, you will be asked how many clinically significant errors and clinically insignificant errors the report makes. An error can be one of the following:

- False prediction of finding
- Omission of finding
- Incorrect location/position of finding
- Incorrect severity of finding
- Mention of comparison that is **not** present in the reference impression
- Omission of comparison describing a change from a previous study

Note: Some candidate reports will often be the same. In these cases, make sure to give the same scores to the candidates.

When you are ready to begin the study, press the next page button below. You can always go back to this page to review the error guidelines.

Figure S1. Radiologist evaluation survey instructions and interface on Qualtrics.



Stanford

Study #1

Reference impression: Multiple chronic appearing left-sided rib fractures. No pneumothorax. Blunting of the costophrenic angle on the right likely represents pleural scarring and a small effusion, not significantly changed from ___.

Candidate 1: Blunting of the right costophrenic angle may be due to small pleural effusion .

How many of the following errors does this report make:

	Clinically significant	Clinically insignificant
False prediction of finding	□	□
Omission of finding	□	□
Incorrect location/position of finding	□	□
Incorrect severity of finding	□	□
Mention of comparison that is **not** present in the reference impression	□	□
Omission of comparison describing a change from a previous study	□	□

Candidate 2: No focal consolidation. Minimal blunting of the left costophrenic angle may represent a trace pleural effusion.

How many of the following errors does this report make:

	Clinically significant	Clinically insignificant
False prediction of finding	□	□
Omission of finding	□	□
Incorrect location/position of finding	□	□
Incorrect severity of finding	□	□
Mention of comparison that is **not** present in the reference impression	□	□
Omission of comparison describing a change	□	□

Figure S2. Interface for evaluating a pair of a test report (denoted as "Reference Impression") and a metric-oracle report (denoted as "Candidate 1"). The survey asks radiologists to input the number of clinically significant and insignificant errors for six error categories.

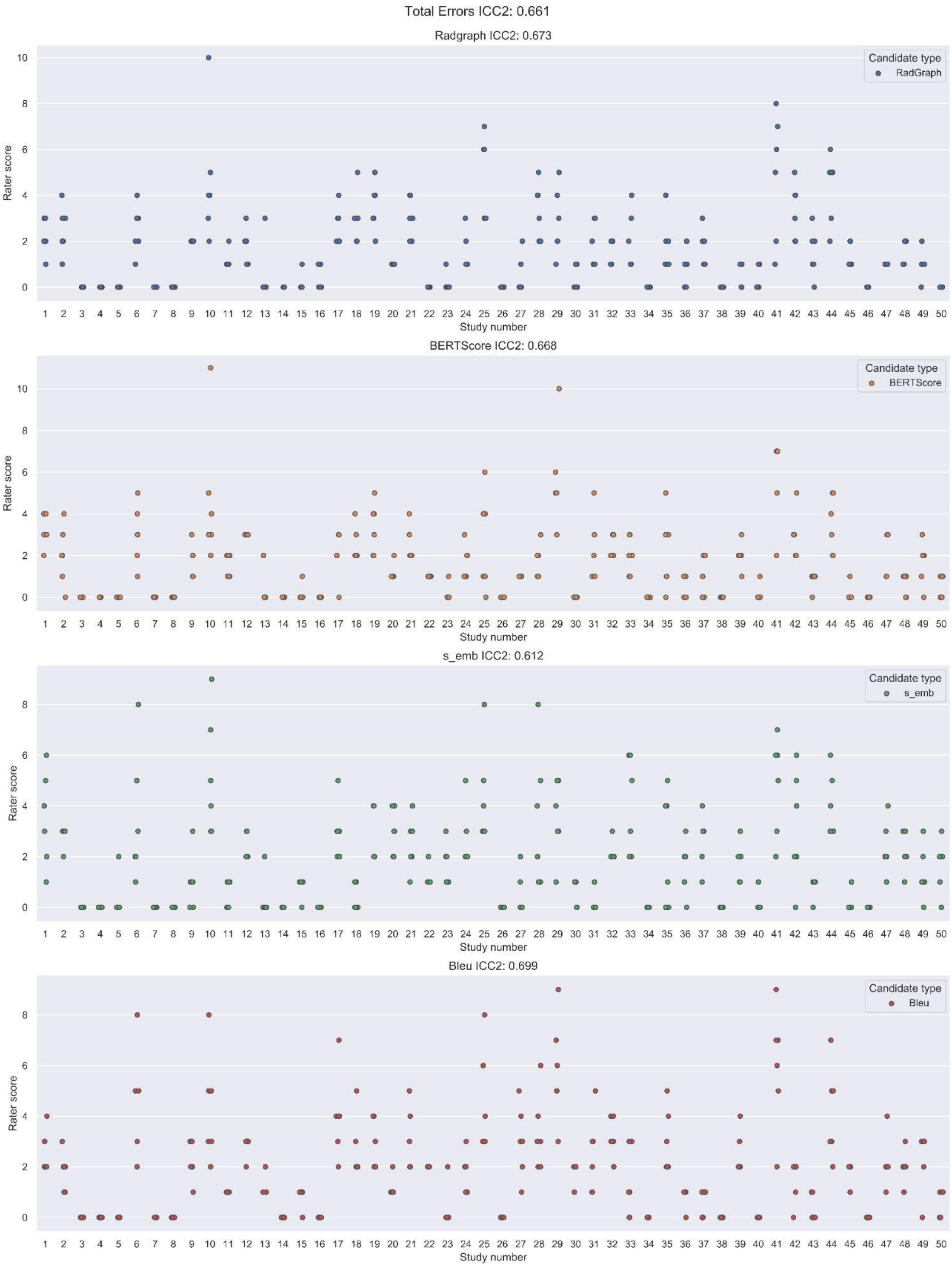


Figure S3. Dotplot of the radiologist total error scores on the 50 studies and corresponding intraclass correlation. Candidate scores are split up by metric-oracle method. Each dot represents a single radiologist's score for a candidate report.

Table S1. Coverage of pathologies, as determined by the CheXpert labels in MIMIC-CXR, for the 50 randomly sampled reports in the radiologist experiment. The counts listed for “No Finding” refer to explicit labels for “No Finding”: namely 1.0 for positive mentions, 0.0 for negative mentions and -1.0 for uncertain mentions. There are edge cases where explicit labels for “No Finding” were suppressed, for instance if a pathology not included in the CheXpert label set was mentioned. If we also count studies where no other pathology had a positive label along with explicit positive labels for “No Finding” then we have 15 counts, instead of 14.

	Positive mentions	Negative mentions	Uncertain mentions
Atelectasis	8	0	2
Cardiomegaly	9	2	1
Consolidation	3	1	1
Edema	6	4	2
Enlarged Cardiomediastinum	1	0	0
Fracture	2	0	0
Lung Lesion	3	0	2
Lung Opacity	16	0	0
No Finding	14	0	0
Pleural Effusion	9	5	1
Pleural Other	3	0	0
Pneumonia	2	4	9
Pneumothorax	1	6	0
Support Devices	8	0	0

Table S2. Per-radiologist Kendall rank correlation coefficient (tau-b) values quantifying metric-radiologist alignment.

	Radiologist 1	Radiologist 2	Radiologist 3	Radiologist 4	Radiologist 5	Radiologist 6
BERTScore sig. and insig. errors	0.454 [95% CI 0.374 0.527]	0.441 [95% CI 0.362 0.517]	0.535 [95% CI 0.458 0.605]	0.442 [95% CI 0.348 0.521]	0.454 [95% CI 0.373 0.533]	0.511 [95% CI 0.424 0.590]
BERTScore sig. errors	0.410 [95% CI 0.322 0.490]	0.337 [95% CI 0.251 0.418]	0.540 [95% CI 0.469 0.600]	0.456 [95% CI 0.367 0.533]	0.419 [95% CI 0.342 0.497]	0.487 [95% CI 0.400 0.567]
RadGraph sig. and insig. errors	0.505 [95% CI 0.426 0.573]	0.499 [95% CI 0.429 0.566]	0.539 [95% CI 0.474 0.602]	0.491 [95% CI 0.415 0.554]	0.507 [95% CI 0.428 0.579]	0.451 [95% CI 0.364 0.526]
RadGraph sig. errors	0.474 [95% CI 0.398 0.556]	0.351 [95% CI 0.255 0.437]	0.540 [95% CI 0.476 0.596]	0.523 [95% CI 0.451 0.588]	0.499 [95% CI 0.424 0.566]	0.426 [95% CI 0.336 0.507]
BLEU sig. and insig. errors	0.398 [95% CI 0.307 0.483]	0.422 [95% CI 0.338 0.497]	0.475 [95% CI 0.392 0.550]	0.386 [95% CI 0.286 0.469]	0.412 [95% CI 0.319 0.492]	0.463 [95% CI 0.374 0.547]
BLEU sig. errors	0.345 [95% CI 0.248 0.434]	0.255 [95% CI 0.167 0.338]	0.472 [95% CI 0.389 0.542]	0.399 [95% CI 0.302 0.476]	0.357 [95% CI 0.264 0.444]	0.412 [95% CI 0.315 0.501]
CheXbert sig. and insig. errors	0.440 [95% CI 0.348 0.526]	0.424 [95% CI 0.332 0.509]	0.478 [95% CI 0.392 0.558]	0.503 [95% CI 0.420 0.581]	0.489 [95% CI 0.406 0.566]	0.451 [95% CI 0.364 0.530]
CheXbert sig. errors	0.392 [95% CI 0.300 0.475]	0.263 [95% CI 0.168 0.350]	0.451 [95% CI 0.362 0.529]	0.428 [95% CI 0.338 0.509]	0.411 [95% CI 0.330 0.490]	0.407 [95% CI 0.318 0.488]

Table S3(a). Radiologist evaluation of metric-oracles in terms of total number of errors in six error categories, averaged over 6 radiologists and 50 studies.

	Error 1	Error 2	Error 3	Error 4	Error 5	Error 6	Total
BLEU	0.807	0.550	0.113	0.133	0.140	0.097	1.840
CheXbert	0.597	0.443	0.227	0.197	0.150	0.093	1.707
BERTScore	0.477	0.523	0.183	0.153	0.077	0.113	1.527
RadGraph	0.427	0.573	0.147	0.160	0.077	0.110	1.493

Table S3(b). Radiologist evaluation of metric-oracles in terms of number of clinically significant errors in six error categories, averaged over 6 radiologists and 50 studies.

	Error 1	Error 2	Error 3	Error 4	Error 5	Error 6	Total
BLEU	0.607	0.353	0.087	0.093	0.107	0.077	1.323
CheXbert	0.430	0.263	0.193	0.176	0.107	0.077	1.247
BERTScore	0.363	0.310	0.147	0.117	0.053	0.083	1.073
RadGraph	0.300	0.343	0.133	0.143	0.053	0.070	1.043

Table S4(a). Multiple hypothesis testing outputs in terms of *total number of clinically significant and insignificant errors* in *false prediction of finding*. Significance of BLEU having a *more prominent failure mode* than BERTScore and RadGraph F1 in terms of *total number of clinically significant and insignificant errors* in *false prediction of finding*, as determined by the Benjamini-Hochberg Procedure with False Discovery Rate (FDR) of 1%.

	One-sided two-sample t test p-value	Benjamini-Hochberg Procedure critical value with False Discovery Rate (FDR) of 1%	Whether result is significant
BLEU > CheXbert	3.79e-3	2.50e-3	N
BLEU > BERTScore	9.50e-6	1.67e-3	Y
BLEU > RadGraph	1.07e-7	8.33e-4	Y
CheXbert > BLEU	9.96e-1	8.33e-3	N
CheXbert > BERTScore	7.65e-2	4.17e-3	N
CheXbert > RadGraph	6.39e-3	3.33e-3	N
BERTScore > BLEU	1.00e0	9.17e-3	N
BERTScore > CheXbert	9.51e-1	6.67e-3	N
BERTScore > RadGraph	2.24e-1	5.00e-3	N
RadGraph > BLEU	1.00e0	1.00e-2	N
RadGraph > CheXbert	9.94e-1	7.50e-3	N
RadGraph > BERTScore	7.76e-1	5.83e-3	N

Table S4(b). Multiple hypothesis testing outputs in terms of *clinically significant errors* in *false prediction of finding*. Significance of BLEU having a *more prominent failure mode* than BERTScore and RadGraph F1 in terms of *clinically significant errors* in *false prediction of finding*, as determined by the Benjamini-Hochberg Procedure with False Discovery Rate (FDR) of 1%.

	One-sided two-sample t test p-value	Benjamini-Hochberg Procedure critical value with False Discovery Rate (FDR) of 1%	Whether result is significant
BLEU > CheXbert	3.68e-3	2.50e-3	N
BLEU > BERTScore	1.48e-4	1.67e-3	Y
BLEU > RadGraph	6.44e-7	8.33e-4	Y
CheXbert > BLEU	9.96e-1	8.33e-3	N
CheXbert > BERTScore	1.34e-1	5.00e-3	N
CheXbert > RadGraph	9.77e-3	3.33e-3	N
BERTScore > BLEU	1.00e0	9.17e-3	N
BERTScore > CheXbert	8.66e-1	5.83e-3	N
BERTScore > RadGraph	1.33e-1	4.17e-3	N
RadGraph > BLEU	1.00e0	1.00e-2	N
RadGraph > CheXbert	9.90e-1	7.50e-3	N
RadGraph > BERTScore	8.67e-1	6.67e-3	N

Table S4(c). Multiple hypothesis testing outputs in terms of *total number of clinically significant and insignificant errors* in *incorrect location/position of finding*. Significance of BLEU having a *less prominent failure mode* than CheXbert vector similarity in terms of *total number of clinically significant and insignificant errors* in *incorrect location/position of finding*, as determined by the Benjamini-Hochberg Procedure with False Discovery Rate (FDR) of 1%.

	One-sided two-sample t test p-value	Benjamini-Hochberg Procedure critical value with False Discovery Rate (FDR) of 1%	Whether result is significant
BLEU < CheXbert	4.83e-4	8.33e-4	Y
BLEU < BERTScore	1.60e-2	2.50e-3	N
BLEU < RadGraph	1.51e-1	5.00e-3	N
CheXbert < BLEU	1.00e0	1.00e-2	N
CheXbert < BERTScore	8.90e-1	7.50e-3	N
CheXbert < RadGraph	9.89e-1	9.17e-3	N
BERTScore < BLEU	9.84e-1	8.33e-3	N
BERTScore < CheXbert	1.10e-1	3.33e-3	N
BERTScore < RadGraph	8.63e-1	6.67e-3	N
RadGraph < BLEU	8.49e-1	5.83e-3	N
RadGraph < CheXbert	1.14e-2	1.67e-3	N
RadGraph < BERTScore	1.37e-1	4.17e-3	N

Table S4(d). Multiple hypothesis testing outputs in terms of *clinically significant errors* in *incorrect location/position of finding*. Significance of BLEU having a *less prominent failure mode* than CheXbert vector similarity in terms of *clinically significant errors* in *incorrect location/position of finding*, as determined by the Benjamini-Hochberg Procedure with False Discovery Rate (FDR) of 1%.

	One-sided two-sample t test p-value	Benjamini-Hochberg Procedure critical value with False Discovery Rate (FDR) of 1%	Whether result is significant
BLEU < CheXbert	2.74e-4	8.33e-4	Y
BLEU < BERTScore	2.07e-2	1.67e-3	N
BLEU < RadGraph	5.75e-2	3.33e-3	N
CheXbert < BLEU	1.00e0	1.00e-2	N
CheXbert < BERTScore	9.25e-1	6.67e-3	N
CheXbert < RadGraph	9.67e-1	8.33e-3	N
BERTScore < BLEU	9.79e-1	9.17e-3	N
BERTScore < CheXbert	7.53e-2	4.17e-3	N
BERTScore < RadGraph	6.65e-1	5.83e-3	N
RadGraph < BLEU	9.42e-1	7.50e-3	N
RadGraph < CheXbert	3.32e-2	2.50e-3	N
RadGraph < BERTScore	3.35e-1	5.00e-3	N

Table S5. The average, 95% confidence interval and range of metric scores of impression-generating models, including metric-oracle models, CXR-RePaiR and the random retrieval baseline model.

	BLEU	BERTScore	CheXbert vector similarity	RadGraph F1	RadCliQ
BLEU metric-oracle	0.557 [95% CI 0.547 0.567] Range [0.009, 1.000]	0.661 [95% CI 0.652 0.670] Range [-0.266, 1.000]	0.689 [95% CI 0.678 0.699] Range [-0.088, 1.000]	0.476 [95% CI 0.464 0.489] Range [0.000, 1.000]	0.081 [95% CI 0.044 0.118] Range [-1.441, 2.567]
BERTScore metric-oracle	0.491 [95% CI 0.479 0.503] Range [0.000, 1.000]	0.721 [95% CI 0.714 0.729] Range [0.033, 1.000]	0.738 [95% CI 0.728 0.748] Range [-0.050, 1.000]	0.498 [95% CI 0.486 0.511] Range [0.000, 1.000]	-0.095 [95% CI -0.129 -0.062] Range [-1.441, 2.162]
CheXbert metric-oracle	0.381 [95% CI 0.367 0.395] Range [0.000, 1.000]	0.573 [95% CI 0.563 0.585] Range [-0.225, 1.000]	0.954 [95% CI 0.952 0.957] Range [-0.013, 1.000]	0.403 [95% CI 0.390 0.417] Range [0.000, 1.000]	0.052 [95% CI 0.017 0.086] Range [-1.441, 2.543]
RadGraph metric-oracle	0.366 [95% CI 0.356 0.377] Range [0.000, 1.000]	0.541 [95% CI 0.533 0.549] Range [-0.102, 1.000]	0.739 [95% CI 0.729 0.748] Range [-0.028, 1.000]	0.677 [95% CI 0.668 0.686] Range [0.000, 1.000]	-0.020 [95% CI -0.051 0.009] Range [-1.441, 2.500]
CXR-RePaiR	0.055 [95% CI 0.053 0.057] Range [0.000, 0.383]	0.193 [95% CI 0.188 0.198] Range [-0.402, 0.633]	0.379 [95% CI 0.370 0.387] Range [-0.146, 0.973]	0.090 [95% CI 0.086 0.095] Range [0.000, 1.000]	1.642 [95% CI 1.625 1.659] Range [-0.645, 2.904]
Random retrieval of impression	0.048 [95% CI 0.044 0.053] Range [0.000, 1.000]	0.222 [95% CI 0.216 0.227] Range [-0.326, 1.000]	0.269 [95% CI 0.259 0.279] Range [-0.226, 1.000]	0.050 [95% CI 0.045 0.055] Range [0.000, 1.000]	1.755 [95% CI 1.733 1.776] Range [-1.441, 3.111]

Table S6. The average, 95% confidence interval and range of metric scores of findings-generating models, including M^2 Trans.

	BLEU	BERTScore	CheXbert vector similarity	RadGraph F1	RadCliQ
M^2 Trans	0.220 [95% CI 0.214 0.224] Range [0.001, 0.643]	0.386 [95% CI 0.380 0.391] Range [-0.127, 0.738]	0.452 [95% CI 0.441 0.463] Range [-0.089, 0.987]	0.244 [95% CI 0.238 0.250] Range [0.000, 0.823]	1.059 [95% CI 1.037 1.083] Range [-0.701, 2.279]
Random retrieval of findings	0.123 [95% CI 0.119 0.127] Range [0.000, 0.778]	0.323 [95% CI 0.318 0.328] Range [-0.138, 0.845]	0.235 [95% CI 0.225 0.245] Range [-0.217, 0.920]	0.105 [95% CI 0.101 0.109] Range [0.000, 0.866]	1.553 [95% CI 1.533 1.573] Range [-0.938, 2.537]

Table S7. The average, 95% confidence interval and range of metric scores of models that jointly generate findings and impression sections, including R2Gen, WCL and CvT2DistilGPT2.

	BLEU	BERTScore	CheXbert vector similarity	RadGraph F1	RadCliQ
R2Gen	0.137 [95% CI 0.133 0.141] Range [0.000, 0.826]	0.271 [95% CI 0.266 0.276] Range [-0.318, 0.853]	0.286 [95% CI 0.276 0.295] Range [-0.204, 0.993]	0.134 [95% CI 0.130 0.138] Range [0.000, 0.883]	1.552 [95% CI 1.533 1.571] Range [-1.038, 3.063]
WCL	0.144 [95% CI 0.141 0.148] Range [0.000, 0.612]	0.275 [95% CI 0.271 0.279] Range [-0.332, 0.745]	0.309 [95% CI 0.300 0.318] Range [-0.214, 0.991]	0.143 [95% CI 0.139 0.147] Range [0.000, 0.694]	1.511 [95% CI 1.493 1.529] Range [-0.608, 3.060]
CvT2DistilGPT2	0.143 [95% CI 0.139 0.147] Range [0.000, 0.826]	0.280 [95% CI 0.275 0.285] Range [-0.351, 0.842]	0.335 [95% CI 0.326 0.344] Range [-0.224, 0.993]	0.154 [95% CI 0.149 0.159] Range [0.000, 0.883]	1.463 [95% CI 1.443 1.484] Range [-1.027, 2.909]
Random retrieval of jointly the findings and impression	0.100 [95% CI 0.097 0.103] Range [0.000, 0.498]	0.256 [95% CI 0.252 0.261] Range [-0.383, 0.809]	0.190 [95% CI 0.182 0.198] Range [-0.281, 0.940]	0.090 [95% CI 0.087 0.093] Range [0.000, 0.726]	1.726 [95% CI 1.710 1.741] Range [-0.724, 3.067]