# Supplementary

Split	2D Slices	CTs	Findings Tokens	ICD9 Codes	ICD10 Codes	Patients
Train	6,387,231	15,331	6,036,645	577,998	1,261,561	11,010
Validation	2,099,217	5,060	1,985,925	178,194	379,733	3,644
Test	2,142,061	5,137	2,029,001	197,821	399,986	3,667
Total	10,628,509	25,528	10,051,571	954,013	2,041,280	18,321

Table 1: Summary of pretraining dataset splits.

Demographic	Patients (n = 18,321)	Value
Age	-	53.8±19.5
Gender		
Female	10,254	55.97%
Male	8,065	44.02%
Self-Reported Race/Ethnicity		
Non-Hispanic White	8,660	47.27%
Asian	2,673	14.59%
Black	952	5.20%
Hispanic White	515	2.81%
Pacific Islander	294	1.60%
Native American	65	0.35%
Unknown	5,127	27.98%
Patient Class		
Inpatient	6,834	37.31%
Emergency Services	6,388	34.87%
Outpatient	2,959	16.16%
Observation	1,452	7.93%

Table 2: *Internal dataset characteristics*. The age value is provided as mean  $\pm$  standard deviation. All other values are provided as percentages of the total patients (n = 18,321).

Demographic	Patients (n = 5,804)	Value
Age	-	61.4±16.5
Gender		
Female	2,982	51.38%
Male	2,822	48.62%
Self-Reported Race/Ethnicity		
Non-Hispanic White	4,576	78.84%
Black	270	4.65%
Hispanic White	198	2.76%
Asian	87	1.50%
Native American	31	0.53%
Pacific Islander	4	0.07%
Unknown	602	10.37%

Table 3: *External dataset characteristics*. The age value is provided as mean  $\pm$  standard deviation. All other values are provided as percentages of the total patients (n = 5,804).

Encoder	Init	Labels	$\frac{Stem}{\mathrm{KS}_z} /$		<b>All Pho</b> N=691, Pr	ecodes ev=3.1%		Upper ( by Prev	Quartile valence	Lower Quartile by Prevalence	
			$Stride_z$			Phecode	es w/ AUC	N=173; P	rev=8.7%	N=173, P	rev=0.6%
				AUROC	AUPRC	>0.85	>0.9	AUROC	AUPRC	AUROC	AUPRC
SwinUNETR	MAE	EHR	4/4	.736	.088	52	8	.734	.207	.738	.029
ConvNeXt-T	I3D	EHR	7/2	.768	.106	115	20	.765	.239	.764	.035
ConvNeXt-S	I3D	EHR	7/2	.761	.102	105	14	.760	.234	.750	.031
ConvNeXt-B	I3D	EHR	7/2	.773	.110	132	28	.768	.244	.766	.036
ConvNeXt-B*	I3D	EHR	3/2	.789	.131	180	46	.784	.270	.781	.052
ResNet50	I3D	EHR	3/1	.798	.137	226	72	.787	.280	.797	.049
ResNet152	I3D	EHR	7/2	.798	.135	221	65	.785	.272	.796	.050
$\downarrow$	I3D	EHR	3/2	.798	.136	221	74	.788	.275	.792	.049
	I3D	EHR	3/1	.801	.140	235	79	.789	.279	.801	.054
(Merlin)	I3D	MTL	3/1	.812	.142	259	93	.804	.290	.808	.050

Table 4: *Phenotype classification.* We only include phenotypes that have more than 20 positive examples in the test set in order to ensure a meaningful measure of performance. \* in ConvNext-B\* indicates that instead of inflating the z dimension to a size equal to the 2D kernel height and width of 7, the kernel is inflated to a depth of 3.

Init	Labels	Split Text	$\begin{array}{c} Stem \\ \mathrm{KS}_z / \\ \mathrm{Stride}_z \end{array}$	All Phecodes N=691, Prev=3.1 Phece			s w/ AUC	Upper ( by Prev N=173; Pr	<b>)uartile</b> v <b>alence</b> rev=8.7%	Lower Quarti by Prevalenc N=173, Prev=0.6		
				AUROC	AUPRC	>0.85	>0.9	AUROC	AUPRC	AUROC	AUPRC	
I3D	Staged	X	3/1	.804	.145	240	84	.792	.287	.798	.056	
I3D	Staged	1	3/1	.807	.146	249	87	.795	.291	.801	.056	
I3D	MTL	X	3/1	.814	.153	267	103	.806	.302	.807	.058	
Rand	MTL	1	3/1	.786	.124	117	46	.778	.261	.779	.042	
I3D	MTL	1	3/1	.812	.142	259	93	.804	.290	.808	.050	

Table 5: *Phenotype classification ablation study.* We perform ablation studies where we examine the impact of I3D initialization, staged training versus multi-task learning (MTL) with EHR and radiology reports, and splitting the report text with every other batch for finer grain contrastive learning.

Method	Average F1 Score
OpenCLIP (Internal)	.276 [.262, .288]
BioMedCLIP (Internal)	.285 [.274, .295]
Merlin (Internal)	.741 [.727, .755]
Merlin (External)	.647 [.607, .678]
Maglin (VarSa)	.767 [.630, .667]

Table 6: Zero-shot classification. The internal and external numbers (first 4 rows) represent averages over the 30 findings for the internal clinical dataset and the external clinical dataset respectively. The bottom row represents F1 score for zero-shot classification of vertebral fractures on the VerSe dataset.

Init	Labels	Split Text	Average F1 Score
12D	Demont	1	720[714 744]
15D	Report	~	./50[./14, ./44]
I3D	Staged	X	.669 [.653, .683]
I3D	Staged	1	.735 [.719, .748]
I3D	MTL	X	.656 [.640, .671]
Rand	MTL	1	.698 [.681, .711]
I3D	MTL	1	.741 [.727, .755]

Table 7: Zero-shot classification ablation study. We measure zero-shot performance as we vary parameters of I3D versus random initialization, staged training versus multi-task learning (MTL) with EHR and radiology reports, and training with the full findings sections versus using radiology report splitting.

Task	Method		Recall@1			Recall@	8
		N=32	N=64	N=128	N=32	N=64	N=128
Img→F	OpenCLIP	.030	.016	.009	.243	.125	.062
	BioMedCLIP	.040	.021	.010	.298	.156	.083
	Merlin	.780	.696	.608	.989	.968	.927
F→Img	OpenCLIP	.033	.017	.009	.250	.125	.061
	BioMedCLIP	.044	.021	.012	.306	.156	.079
	Merlin	.776	.687	.594	.988	.965	.920
Img→I	OpenCLIP	.030	.016	.010	.256	.128	.061
	BioMedCLIP	.036	.017	.009	.273	.141	.073
	Merlin	.352	.253	.174	.796	.663	.532
I→Img	OpenCLIP	.032	.017	.008	.252	.126	.064
	BioMedCLIP	.046	.024	.012	.322	.169	.081
	Merlin	.384	.277	.194	.854	.706	.564

Table 8: *Cross-modality retrieval*. We compare performance of OpenCLIP, BioMedCLIP, and Merlin across several settings: retrieving the correct findings section given an image (Img  $\rightarrow$  F), retrieving the correct image given a findings section (F  $\rightarrow$  Img), retrieving the correct impressions section given an image (Img  $\rightarrow$  I), and retrieving the correct image given an impressions section (I  $\rightarrow$  Img). We perform retrieval within pools of sizes N=32, N=64, and N=128.

Task	Init	Labels	Split		Recall@	1		Recall@	8
			Text	N=32	N=64	N=128	N=32	N=64	N=128
Img→F	I3D	Report	1	.778	.692	.598	.989	.967	.921
	I3D	Staged	X	.654	.547	.449	.967	.920	.844
	I3D	Staged	✓	.672	.561	.457	.972	.925	.848
	I3D	MTL	X	.812	.726	.639	.988	.969	.937
	Rand	MTL	1	.690	.583	.468	.978	.937	.867
	I3D	MTL	1	.780	.696	.608	.989	.968	.927
F→Img	I3D	Report	1	.775	.686	.584	.991	.969	.921
	I3D	Staged	X	.646	.539	.434	.965	.913	.841
	I3D	Staged	1	.664	.555	.445	.970	.923	.841
	I3D	MTL	X	.801	.718	.626	.988	.968	.933
	Rand	MTL	1	.683	.571	.455	.980	.940	.869
	I3D	MTL	1	.776	.687	.594	.988	.965	.920
Img→I	I3D	Report	1	.364	.265	.187	.812	.681	.549
	I3D	Staged	X	.307	.220	.159	.737	.590	.449
	I3D	Staged	$\checkmark$	.328	.228	.163	.780	.634	.500
	I3D	MTL	X	.372	.275	.196	.799	.667	.543
	Rand	MTL	$\checkmark$	.288	.202	.131	.740	.592	.453
	I3D	MTL	1	.352	.253	.174	.796	.663	.532
I→Img	I3D	Report	1	.382	.274	.192	.850	.709	.574
	I3D	Staged	X	.324	.234	.161	.770	.616	.490
	I3D	Staged	✓	.348	.246	.168	.811	.672	.523
	I3D	MTL	X	.400	.294	.216	.817	.698	.568
	Rand	MTL	✓	.289	.200	.127	.779	.613	.460
	I3D	MTL	1	.384	.277	.194	.854	.706	.564

Table 9: *Cross-modality retrieval ablation study*. We compare retrieval performance across three axes of weight initializations, methods for incorporating EHR and radiology reports into training, and splitting or using the full findings during training.

Encoder	Init	Labels	% <i>Tr</i> 10% 100%	<b>CKD</b> 14/46 90/513	<b>DM</b> 9/46 80/474	HTN 11/34 111/404	IHD 9/49 69/518	<b>CVD</b> 20/52 136/504	<b>OST</b> 10/47 68/527	Average
Swin Transformer ResNet152 ↓	- - I3D I3D	- - - EHR	10% 10% 10% 10%	.46 [.40, .50] .53 [.48, .58] .72 [.67, .76] .58 [.53, .63]	.70 [.66, .75] .66 [.61, .71] <b>.72 [.67, .76]</b> .64 [.59, .69]	.43 [.39, .48] .60 [.55, .64] .67 [.63, .71] .47 [.42, .51]	.53 [.49, .57] .49 [.45, .54] .67 [.63, .71] .75 [.71, .78]	.53 [.50, .57] .58 [.54, .62] .67 [.64, .71] .69 [.65, .72]	.56 [.51, .61] .50 [.44, .56] .66 [.61, .71] .62 [.57, .67]	.54 [.52, .55] .56 [.54, .58] .68 [.67, .70] .62 [.60, .64]
(Merlin)	I3D	MTL	10%	.74 [.70, .78]	.70 [.66, .75]	.69 [.65, .73]	.70 [.67, .74]	.73 [.69, .76]	.69 [.65, .73]	.71 [.69, .72]
Swin Transformer ResNet152 ↓	- I3D I3D	- - - EHR	100% 100% 100% 100%	.55 [.50, .59] .63 [.58, .67] .74 [.70, .77] .76 [.73, .80]	.73 [.68, .77] .74 [.69, .78] .74 [.70, .78] .72 [.68, .76]	.61 [.57, .65] .65 [.61, .69] .71 [.67, .75] .74 [.70, .77]	.52 [.48, .56] .65 [.61, .69] .68 [.64, .72] .74 [.70, .78]	.54 [.49, .57] .57 [.54, .61] .68 [.64, .71] .73 [.69, .76]	.60 [.55, .66] .60 [.55, .66] .74 [.69, .78] .68 [.64, .73]	.59 [.57, .61] .64 [.62, .66] .71 [.70, .73] .73 [.71, .74]
(Merlin)	I3D	MTL	100%	.77 [.74, .81]	.72 [.68, .76]	.75 [.72, .79]	.76 [.72, .79]	.74 [.71, .77]	.80 [.76, .84]	.76 [.74, .77]

Table 10: *Multi-disease 5-year prediction*. We fine-tune Merlin for 5-year disease prediction. All data used in this evaluation, including train, val, and test splits, are held out from pretraining.

	BLE	U↑	ROUG	GE-2↑	BEF	<b>RT</b> ↑	RadGraph-F1 $\uparrow$		
Section	RadFM	Merlin	RadFM	Merlin	RadFM	Merlin	RadFM	Merlin	
Lower thorax	.001	.019	.070	.332	.406	.615	.020	.319	
Liver and biliary tree	.001	.269	.025	.389	.328	.641	.080	.380	
Gallbladder	.000	.006	.006	.632	.534	.851	.152	.721	
Spleen	.000	.002	.004	.710	.382	.853	.283	.805	
Pancreas	.000	.001	.010	.700	.447	.849	.091	.748	
Adrenal glands	.006	.030	.067	.882	.490	.942	.106	.879	
Kidneys and ureters	.005	.269	.040	.385	.368	.654	.091	.387	
Gastrointestinal tract	.001	.013	.037	.152	.398	.531	.092	.167	
Peritoneal cavity	.000	.206	.005	.390	.387	.702	.050	.335	
Pelvic organs	.000	.233	.009	.358	.328	.656	.036	.432	
Vasculature	.000	.026	.004	.485	.232	.748	.006	.548	
Lymph nodes	.003	.023	.119	.601	.502	.775	.031	.542	
Musculoskeletal	.001	.046	.018	.303	.449	.689	.008	.293	
Full report	.000	.102	.011	.262	.224	.588	.008	.293	

Table 11: *Radiology report generation*. We compare Merlin and RadFM for generating radiology report sections corresponding to various anatomies, as well as the full findings.

# а

b

# Annotation Color Key

Green: correct, exists in both human, model reports
Purple: correct, exists in only human report
Blue: correct, exists in only model report
Orange: mischaracterized positive finding for that anatomy
Pink: false positive
Red: false negative

# Human Report

lower thorax: small bilateral pleural effusions with associated atelectasis consolidation. liver and biliary tree: stable subcentimeter hypodensity in segment 8 of liver , too small to characterize. mild periportal edema likely postsurgical. new linear hypoattenuating area within lateral segment of left lobe, most likely related to retraction, gallbladder; surgically absent. spleen: normal. pancreas: normal. adrenal glands: normal. kidneys and ureters: normal. bowel: status post recent whipple procedure. no intrahepatic ductal dilatation or pancreatic ductal dilatation. stomach does not appear significantly distended. mild wall thickening of gastrojejunostomy due to lack of distention or postoperative edema. postoperative stranding in the resection bed, tracking along sma and smv. peritoneal cavity: 8.5 x 1.6 x 4.9 cm fluid collection in right anterior pararenal space, with attenuation compatible with hematoma. a small amount of nonspecific free fluid is also present in the pelvis. abdominal wall: mild diffuse subcutaneous stranding along bilateral flanks. bladder small amount of air in the bladder likely due to recent catheterization, uterus and ovaries: redemonstration of a fibroid originating from the fundus of the uterus. vasculature: mild narrowing of portosplenic confluence and proximal smv, likely secondary to postop edema. hepatic artery is patent. lymph nodes: a few of left paraaortic subcentimeter lymph nodes are mildly increased in size, most likely reactive. musculoskeletal: normal.

#### Merlin Report

lower thorax: there is a small right pleural effusion with associated compressive atelectasis. liver and biliary tree: normal. gallbladder: surgically absent, spleen: normal, pancreas: postsurgical changes of whipple procedure, the pancreaticojejunostomy is patent with no evidence of obstruction or leakage at this level, there are multiple fluid collections in the surgical bed, including a 3 x 2 cm collection adjacent to the hepaticojejunal anastomosis (4/107), and a 5 x 6 mm collection anterior to the sma (4/98), adrenal glands: normal, kidneys and ureters: normal, gastrointestinal tract: postsurgical changes of whipple procedure. no evidence for bowel obstruction or inflammation. peritoneal cavity: there is a small amount of free fluid in the pelvis. abdominal wall: postsurgical changes are seen along the anterior midline abdomen with mild subcutaneous stranding and foci of gas, likely related to recent surgery. bladder: normal, uterus and ovaries: surgically absent, vasculature: patent. lymph nodes: normal.

#### Takeaways

- Human incorrectly labels the pancreas as normal (red). Later in the report, this was described correctly, indicating the human overlooked a default field in the template,
- Model identifies only one side of bilateral pleural effusions (orange)
   Model describes diffuse fluid below the anastomosis site as a "collection," which is a slight mischaracterization as the fluid is not encapsulated (orange)

# Human Report

С

lower thorax: normal. liver and biliary tree: normal morphology, contour and size of liver. stable appearance of focal, subcentimeter hypodensity in the posterior right hepatic lobe (series 3 / 50), too small to characterize but statistically favored to represent a cyst. gallbladder: normal. spleen: normal. pancreas: normal. adrenal glands: normal. kidneys and ureters: normal morphology with bilateral kidneys. normal parenchymal enhancement and excretion. focal dilatation of the mid right ureter, without obstructing lesion or calculus seen. no ureteral enhancement. gastrointestinal tract: normal caliber of the small and large bowel . no bowel wall thickening . normal appearance of the appendix ( series 2 / 203 ; 301 / 61 ). peritoneal cavity: no free fluid. abdominal wall: normal. bladder: mild, asymmetric thickening of the bladder wall anteriorly. prostate and seminal vesicles: normal. vasculature: patent. lymph nodes: normal. musculoskeletal: stable appearance of sclerotic focus in the left femoral head and right femoral neck ( series 3/286 ; 3/322 ), likely representing a bone island.

# Merlin Report

lower thorax: no significant pulmonary, pleural, or mediastinal abnormality is seen in the visualized portions of the thorax. liver: normal. gallbladder/bile ducts: the common bile duct is mildly prominent measuring up to 8 mm, likely related to patient's age and postcholecystectomy state. no evidence of choledocholithiasis or intrahepatic biliary dilatation, gallbladder; appears normal, spleen; normal. pancreas: no evidence of masses or calcifications in the pancreas. no pancreatic ductal dilatation. genitourinary: there is a 1 cm cystic lesion within the left kidney, likely representing a simple renal cyst (series 3 image 207). otherwise, the kidneys are unremarkable without evidence of enhancing renal masses, nephrolithiasis, or hydronephrosis. the bladder appears normal, gastrointestinal; the stomach and small bowel appear normal. the colon shows mild diverticulitis. adrenal glands: normal. kidneys and ureters: the kidneys enhance homogeneously. no focal renal lesions. no hydronephrosis or nephroureterolithiasis. bowel: no evidence of obstruction. the appendix is normal in appearance. peritoneal cavity: no ascites, pneumoperitoneum, or focal intraabdominal collections, abdominal wall: unremarkable, pelvis; the liver is normal without evidence of enhancing masses or biliary dilatation. a 1 cm low attenuating lesion is noted in segment ivb (image 97, series 4), which is too small to characterize but likely represents a simple cyst. the hepatic and portal veins are patent. evaluation of the gallbladder demonstrates no evidence of stones, wall thickening, or pericholecystic fluid. the spleen appears normal in size. no evidence of splenic masses. evaluation the pancreas demonstrates. vasculature: the abdominal aorta is normal in caliber, no atherosclerotic calcifications, lymph nodes; no enlarged retroperitoneal, mesenteric, pelvic or inguinal lymph nodes. musculoskeletal: no aggressive osseous lesion is identified in the abdomen or pelvis.

# Takeaways

- The model report (274 words) is considerably longer than the human report (154 words).
- Model correctly describes the gallbladder appearance as normal (green), but this logically contradicts the earlier, false statement of "postcholecystectomy" (pink).
- Model contains two sections for kidneys which logically contradict: the genitourinary section falsely describes a cyst (pink); the kidneys and ureters section describes no finding (green, orange).

Figure 8: *Radiology report generation*. (a) As shown in the annotation color key, we annotate individual phrases to be correct, mischaracterized, false positive, or false negative. (b - c) We provide dense annotations of two sets of human and Merlin generated reports.