Test Report

Siriraj AI

โดย คณะแพทยศาสตร์ศิริราชพยาบาล

รายงานผลการทดสอบ
โดยราชวิทยาลัยรังสีแพทย์แห่งประเทศไทย

ทดสอบใช้กับภาพรังสีทรวงอก ในกรณี

- อ่านผลซ้ำ (double reading) ให้กับรังสีแพทย์ เพื่อเพิ่มคุณภาพการวินิจฉัย
- เพิ่มความแม่นยำในการค้นหาพยาธิสภาพให้กับรังสีแพทย์
- ประมาณความยาก-ง่ายในการแปลผล
- จัดลำดับความเร่งด่วน (triage) ในการแปลผลให้แก่รังสีแพทย์

รายงานผลการทดสอบปัญญาประดิษฐ์ในภาพถ่ายรังสีทรวงอก

หน่วยงานที่ยื่นทดสอบ

หน่วยงาน	ภาควิชารังสีวิทยา มหาวิทยาลัยมหิดล	Mahidol University Faculty of Medicine Siriraj Hospital
ที่อยู่	2 ถนนพรานนก แขวงศิราราช เขตบางกอกน้อย กรุงเทพฯ 10700	
ชื่อผู้ติดต่อ	นส.อำไพ อุไรเวโรจนากร	

ระบบที่ยื่นทดสอบ

ชื่อระบบ	โครงการการพัฒนาโปรแกรมปัญญาประดิษฐ์ในการวินิจฉัยภาพทางรังสี
เวอร์ชั่น	–
ข้อมูลระบบ	–

ชุดข้อมูลที่ใช้ทดสอบ

รหัสชุดข้อมูล	1A
จำนวนภาพถ่าย	300 ภาพ

คุณลักษณะของข้อมูลที่ใช้ทดสอบ

จำนวนภาพถ่าย 300 ภาพได้ถูกสุ่มมาจากชุดข้อมูลทดสอบจำนวน 1,500 ภาพซึ่งได้รับการสนับสนุนจาก โรงพยาบาลสงขลานครินทร์ จังหวัดสงขลา, โรงพยาบาลเชียงรายประชารักษ์ จังหวัดเชียงราย, โรงพยาบาลอุดรธานี จังหวัดอุดรธานี, โรงพยาบาลสุทธาเวช จังหวัดมหาสารคาม, และกองวัณโรค กรมควบคุมโรค กระทรวงสาธารณสุข โดยทุกภาพจะถูกอ่านโดยรังสีแพทย์ B Reader จำนวน 3 ท่าน

คุณลักษณะของภาพถ่ายรังสีในโครงการ:

- ภาพถ่ายรังสีทรวงอกของผู้ป่วยที่มีอายุตั้งแต่ 15 ปี ด้วยเครื่องเอกซเรย์คอมพิวเตอร์
- ไม่มีภาพถ่ายจากผู้ป่วยที่มีสถานะ HIV Serology เป็นบวก
- ไม่มีภาพถ่ายจากผู้ป่วยที่มีการติดเชื้อที่ปอดแบบฉวยโอกาสอื่น ๆ หรือการติดเชื้อร่วม เช่น การติดเชื้อไมโครแบคทีเรีย ฮิสโตพลาสโมซิส ไครปโตคอคโคซิส เมลิออยโดซิส และแอคติโนมัยโคซิส

การวัดความตรงภายในในชุดข้อมูล ใช้ตัววัดดังต่อไปนี้

- Pairwise Agreement ค่าเฉลี่ยของความเหมือนกันกันระหว่างแต่ละคู่ของรังสีแพทย์
- Intraclass Agreement (ICC) ค่าเฉลี่ยของ Pearson's correlation แบบ ICC(2,3) เมื่อมีรังสีแพทย์ 3 ท่าน อ่านภาพถ่ายรังสีแบบสุ่ม
- การวิเคราะห์ความตรงระหว่างผู้ประเมินโดยใช้สถิติ Pairwise Cohen’s Kappa และ Fless’ Kappa

จำนวนรอยโรคในชุดข้อมูล

ตารางที่ 1 จำนวนรอยโรคที่ระบุโดยรังสีแพทย์ ในภาพถ่ายรังสีได้รับการวินิจฉัยไม่เป็นวัณโรคปอด (N_{Non-TB}) และในภาพถ่ายรังสีได้รับยืนยันผลวัณโรคปอดโดยผลตรวจเสมหะย้อมเชื้อหรือผลตรวจเพาะเชื้อ (N_{TB}) รวมถึงจำนวนรอยโรคได้รับการยืนยันว่ามีรอยโรคจากรังสีแพทย์ส่วนใหญ่ ($N_{Concensus}$) ในชุดข้อมูล 1A จำนวน 300 ภาพ แต่ละภาพมีการอ่านโดยรังสีแพทย์ “B” Reader จำนวน 3 รายแบบสุ่ม

Finding	N_{Non-TB}	N_{TB}	$N_{Concensus}$
Abnormality	53	448	158
Small opacity	33	374	135
Primary nodular	1	274	98
Primary reticular	32	87	21
Secondary nodular	9	224	76
Secondary reticular	24	125	30
Large opacity	2	349	119
Mass/nodule	2	114	29
Cavity	0	242	81
Fibrosis	7	196	58
Calcification	5	61	12
Pleural effusion	0	99	34
Pleural thickening	7	122	39
Pneumothorax	0	1	0
Hilar adenopathy	2	87	18
Mediastinal adenopathy	1	30	7
Consistent with tuberculosis	0	422	144
Active Tuberculosis	0	387	136
Patchy infiltration	0	265	99
Cavity with surrounding consolidation	0	210	73
Unilateral hilar/paratracheal lymph node enlargement	0	39	5
Pleural effusion	0	52	15
Miliary nodules	0	74	20
Indeterminate tuberculosis	0	35	6
Reticulonodular infiltration	0	22	4
Destroyed lung or bronchiectasis	0	3	0
Inconsistent with tuberculosis	450	28	156

ความตรงภายในของชุดข้อมูล

ตารางที่ 2 ความตรงภายในระหว่างผู้ประเมิน (Inter-rater Reliability) ของแต่ละรอยโรคในชุดข้อมูล 1A จำนวน 300 ภาพ ของรังสีแพทย์ “B” Reader จำนวน 3 ราย ซึ่งวัดโดยค่าสถิติ Pairwise Agreement, ICC(2,3), Cohen’s kappa และ Fleiss’ kappa

Finding	Agreement	ICC	Cohen’s	Fleiss’
Abnormality	0.8956	0.9181	0.7884	0.7884
Small opacity	0.8356	0.8584	0.6681	0.6681
Primary nodular	0.8422	0.8357	0.6281	0.6282
Primary reticular	0.7911	0.2301	0.0892	0.0897
Secondary nodular	0.7222	0.5352	0.2751	0.2761
Secondary reticular	0.7578	0.3006	0.1260	0.1233
Large opacity	0.8800	0.8993	0.7477	0.7478
Mass/nodule	0.8378	0.5365	0.2777	0.2776
Cavity	0.8600	0.8449	0.6446	0.6439
Fibrosis	0.7978	0.6866	0.4212	0.4212
Calcification	0.8889	0.4027	0.1829	0.1825
Pleural effusion	0.9578	0.9163	0.7838	0.7844
Pleural thickening	0.8711	0.7319	0.4762	0.4752
Pneumothorax	0.9978	0.0000	0.0000	0.0000
Hilar adenopathy	0.8644	0.4867	0.2395	0.2394
Mediastinal adenopathy	0.9556	0.6008	0.3130	0.3318
Consistent with tuberculosis	0.9644	0.9751	0.9286	0.9286
Active Tuberculosis	0.9178	0.9373	0.8323	0.8323
Patchy infiltration	0.8178	0.7940	0.5613	0.5614
Cavity with surrounding consolidation	0.8333	0.7755	0.5333	0.5342
Unilateral hilar/paratracheal lymph node enlargement	0.9244	0.2289	0.0846	0.0887
Pleural effusion	0.9444	0.7431	0.4941	0.4897
Miliary nodules	0.8978	0.5905	0.3212	0.3227
Indeterminate tuberculosis	0.9444	0.5114	0.2588	0.2568
Reticulonodular infiltration	0.9600	0.3705	0.1947	0.1613
Destroyed lung or bronchiectasis	0.9933	0.0067	0.0015	0.0033
Inconsistent with tuberculosis	0.9644	0.9751	0.9286	0.9286

ตารางที่ 3 การตีความค่า ICC และ Kappa ตาม Landis and Koch (1977)[1]

ICC/Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost Perfect

[1] Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. In Biometrics (Vol. 33, Issue 1, p. 159). JSTOR. https://doi.org/10.2307/2529310

ความเชื่อมโยงของรอยโรคในแบบทดสอบกับพยาธิสภาพของที่ระบุจากปัญญาประดิษฐ์

เนื่องจากต้องมีการเชื่อมโยงรอยโรคในแบบทดสอบปัญญาประดิษฐ์กับพยาธิสภาพที่ระบุโดยปัญญาประดิษฐ์เพื่อวัดประสิทธิภาพของปัญญาประดิษฐ์ ตารางที่ 4 แสดงความเชื่อมโยงที่ระบุร่วมกันระหว่างโครงการฯ และหน่วยงานผู้ยื่นทำการทดสอบ รายงานผลการทดสอบฉบับนี้จะใช้คจะยึดตามรายการรอยโรคของโครงการฯ โดยจะใช้คะแนนที่มากที่สุดในพยาธิสภาพของที่ระบุจากปัญญาประดิษฐ์ หากมีมากกว่า 1

ตารางที่ 4 ความเชื่อมโยงของรอยโรคในแบบทดสอบกับพยาธิสภาพของที่ระบุจากปัญญาประดิษฐ์

Finding	Vendor's Finding
Small Opacity	Edema Infiltration
Large Opacity	Pneumonia Consolidation Atelectasis
Fibrosis	Fibrosis
Pleural Effusion	Effusion
Pleural Thickening	Pleural Thickening
Pneumothorax	Pneumothorax

หมายเหตุ: เนื่องจากหน่วยงานผู้ยื่นทำการทดสอบ (ภาควิชารังสีวิทยา มหาวิทยาลัยมหิดล) แจ้งว่าไม่มีรังสีแพทย์ดำเนินการส่วนนี้ให้ รายการความเชื่อมโยงนี้จึงถูกระบุโดยคณะรังสีแพทย์ของโครงการแบบทดสอบฯ

ผลการทดสอบ

การวัดความตรงภายนอกนั้นวัดโดยใช้ Pairwise Agreement หรือค่าเฉลี่ยของความเหมือนกันกันระหว่างคู่ของแต่ละรังสีแพทย์และปัญญาประดิษฐ์และ Pairwise Cohen's Kappa หรือค่าเฉลี่ยของค่าสถิติ Cohen's Kappa ระหว่างคู่ของแต่ละรังสีแพทย์และปัญญาประดิษฐ์ เพื่อเปรียบเทียบระหว่างรังสีแพทย์กับปัญญาประดิษฐ์ ("B" vs AI) และรังสีแพทย์ด้วยกัน ("B" vs "B")

เนื่องจากหน่วยงานยื่นทำการทดสอบได้ให้ค่า Threshold สำหรับ Abnormality Finding เท่านั้น ซึ่งคือ 0.50 ไม่ได้ให้ค่า Threshold สำหรับ Finding ที่เหลือ แต่ทางหน่วยงานแนะนำให้ใช้ค่า 0.50

ตารางที่ 4 การวิเคราะห์ความตรงกันภายนอกระหว่างรังสีแพทย์และปัญญาประดิษฐ์

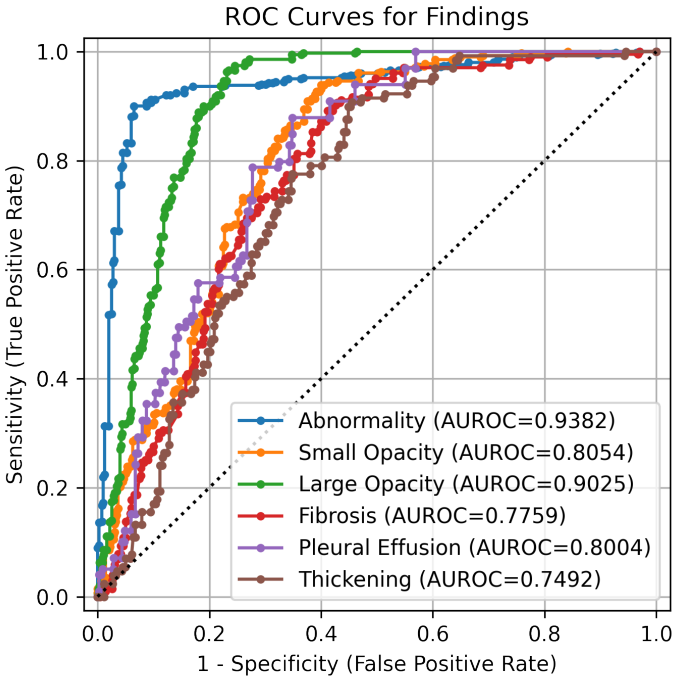
Finding	N	Threshold	Pairwise Agreement		Cohen's Kappa	
			"B" vs "B"	"B" vs AI	"B" vs "B"	"B" vs AI
Abnormality	501	0.50	0.9007	0.8922	0.7993	0.7805
Small Opacity	407	0.50	0.8321	0.6223	0.6714	0.1910
Large Opacity	351	0.50	0.8839	0.8367	0.7557	0.6771
Fibrosis	203	0.50	0.7967	0.7178	0.3927	0.3416
Pleural Effusion	99	0.50	0.9578	0.8811	0.7697	0.0534
Pleural Thickening	129	0.50	0.8711	0.7611	0.4554	0.1916
Pneumothorax	1	0.50	0.9977	0.9089	0.8333	0.000

ในการวัดประสิทธิภาพในการวินิจฉัยของแต่ละรอยโรคนั้นใช้เกณฑ์ ความไว (Sensitivity), ความจำเพาะ (Specificity), อัตราส่วนการทำนายผู้ป่วยที่เป็นโรคที่ถูกต้อง (Positive Prediction Rate, PPR) และอัตราส่วนการทำนายผู้ป่วยที่ไม่เป็นโรคที่ถูกต้อง (Negative Prediction Rate, NPR) ด้วยเกณฑ์การวินิจฉัย (Diagnostic Threshold) ที่ระบุโดยผู้ผลิต รวมถึงพื้นที่ใต้กราฟ ROC เช่นกัน

ตารางที่ 5 ประสิทธิภาพในการวินิจฉัยของแต่ละรอยโรคของปัญญาประดิษฐ์เปรียบเทียบกับรังสีแพทย์

Finding	N	Threshold	Specificity	Sensitivity	PPV	NPV	AUROC
Abnormality	501	0.50	0.8521	0.9242	0.8870	0.8995	0.9382
Small Opacity	407	0.50	0.9412	0.2383	0.7698	0.5995	0.8054
Large Opacity	351	0.50	0.7541	0.9658	0.7152	0.9718	0.9025
Fibrosis	203	0.50	0.7245	0.6946	0.4234	0.8907	0.7759
Pleural Effusion	99	0.50	0.9838	0.0505	0.2778	0.8934	0.8004
Pleural Thickening	129	0.50	0.8197	0.4109	0.2760	0.8927	0.7492

รูปภาพที่ 1 เส้นโค้ง ROC ของแต่ละรอยโรค



บทวิจารณ์

สำหรับการทดสอบโครงการการพัฒนาโปรแกรมปัญญาประดิษฐ์ในการวินิจฉัยภาพทางรังสี มีรอยโรคจำนวน 7 รายการที่ตรงกับรายการรอยโรคในแบบทดสอบ คือ Abnormality, Small Opacity, Large Opacity, Fibrosis, Pleural Effusion, Pleural Thickening, Pneumothorax

ตารางที่ 6 ความแตกต่างระหว่าง Pairwise Agreement และ Cohen's Kappa

Finding	Pairwise Agreement			Cohen's Kappa		
	B vs "B"	"B" vs AI	Diff	"B" vs "B"	"B" vs AI	Diff
Abnormality	0.9007	0.8922	-0.85%	0.7993	0.7805	-1.88%
Small Opacity	0.8321	0.6223	-20.98%	0.6714	0.1910	-48.04%
Large Opacity	0.8839	0.8367	-4.72%	0.7557	0.6771	-7.86%
Fibrosis	0.7967	0.7178	-7.89%	0.3927	0.3416	-5.11%
Pleural Effusion	0.9578	0.8811	-7.67%	0.7697	0.0534	-71.63%
Pleural Thickening	0.8711	0.7611	-11.00%	0.4554	0.1916	-26.38%
Pneumothorax	0.9977	0.9089	-8.88%	0.8333	0.000	-83.33%

จากตารางที่ 6 เมื่อเทียบ Pairwise Agreement และ Cohen's Kappa ระหว่างรังสีแพทย์กับปัญญาประดิษฐ์ ("B" vs AI) และรังสีแพทย์ด้วยกัน ("B" vs "B") แล้ว โครงการการพัฒนาโปรแกรมปัญญาประดิษฐ์ในการวินิจฉัยภาพทางรังสี มีประสิทธิภาพใกล้เคียงกับรังสีแพทย์ (ความแตกต่างน้อยกว่า

5%) ในส่วนของการวินิจฉัยความผิดปกติ (Abnormality) ซึ่งปัญญาประดิษฐ์มีคะแนนน้อยกว่ารังสีแพทย์อยู่ 0.85% สำหรับ Pairwise Agreement และ คะแนนน้อยกว่ารังสีแพทย์อยู่ 1.88% สำหรับ Cohen's Kappa (N = 501)

ในส่วนของการคัดกรองวัณโรคปอดระบบปัญญาประดิษฐ์ที่ทำการทดสอบไม่มีรายงานการคัดกรองวัณโรคปอด จึงไม่มีรายงานในส่วนนี้

/20 มิถุนายน 2566