วิธีการดำเนินงานมาตรฐาน

โครงการแบบทดสอบปัญญาประดิษฐ์สำหรับการคัดกรองวัณโรคปอด ในภาพถ่ายรังสีทรวงอกของกลุ่มประชากรไทย

ราชวิทยาลัยรังสีแพทย์แห่งประเทศไทย
ศูนย์ความเป็นเลิศด้านชีววิทยาศาสตร์
คณะแพทยศาสตร์ มหาวิทยาลัยสงขลานครินทร์
มูลนิธิส่งเสริมวิชาการโรคปอดชนิดวินิจฉัยยาก

สารบัญ

บทนำ

ปัจจุบันเริ่มมีการให้บริการระบบปัญญาประดิษฐ์ในการช่วยอ่านภาพรังสีทรวงอกเพื่อการวินิจฉัยโรค ซึ่งรวมถึงโรควัณโรคปอด ซึ่งระบบปัญญาประดิษฐ์นั้นมีทั้งที่พัฒนาจากองค์กรภาครัฐ และ/หรือองค์กรภาคเอกชนทั้งจากในและต่างประเทศ โดยข้อมูลที่ใช้สร้างระบบปัญญาประดิษฐ์มาจากทั้งข้อมูลสาธารณะ และข้อมูลที่ไม่เป็นสาธารณะ โดยผู้ให้บริการบางเจ้าก็เปิดเผยที่มาของแหล่งข้อมูลแต่บางเจ้าก็ไม่เปิดเผยถึงแหล่งข้อมูลที่ใช้สร้างระบบปัญญาประดิษฐ์

ราชวิทยาลัยรังสีแพทย์แห่งประเทศไทยได้ตระหนักถึงความสำคัญและความแพร่หลายของการใช้งานระบบปัญญาประดิษฐ์ในทางรังสีวินิจฉัย ตั้งแต่ปี พุทธศักราช ๒๕๖๔ โดยทางราชวิทยาลัยได้ออก AI User Guideline เพื่อเป็นแนวทางให้กับผู้ซื้อและผู้ใช้งานระบบปัญญาประดิษฐ์ เพื่อที่จะได้มีข้อพิจารณาที่ถี่ถ้วนก่อนที่จะตัดสินใจซื้อและใช้งาน ซึ่งนับว่าเป็นก้าวที่สำคัญของการยอมรับในการใช้งานระบบปัญญาประดิษฐ์ในงานของรังสีแพทย์

โครงการแบบทดสอบปัญญาประดิษฐ์สำหรับการคัดกรองวัณโรคปอดในภาพถ่ายรังสีทรวงอกของกลุ่มประชากรไทยนี้ได้ทำการรวบรวมภาพถ่ายจากหลายสถาบันในทุกภาคของประเทศไทย จำนวน ๑,๕๐๐ ภาพ และภาพถ่ายรังสีนั้นจะถูกอ่านโดยรังสีแพทย์ที่เป็น NIOSH-B Reader ซึ่งเป็นผู้เชี่ยวชาญด้านการอ่านภาพถ่ายรังสีทรวงอกที่เป็นที่ยอมรับในระดับนานาชาติ จำนวนภาพละ ๓ ท่าน แบบทดสอบได้พัฒนามาจากความคิดเห็นที่สอดคล้องกันของรังสีแพทย์ผู้เชี่ยวชาญจำนวน ๓ ท่าน และสอดคล้องกับผลการทดสอบทางพยาธิวิทยา

โครงการนี้ได้จัดทำขึ้นเพื่อทดสอบระบบปัญญาประดิษฐ์ว่าทำงานได้ดีกับบริบทของกลุ่มประชากรไทยมากน้อยเพียงไร ซึ่งนับเป็นอีกก้าวที่สำคัญในการสร้างมาตรฐานรวมถึงบริบทของการใช้งานระบบปัญญาประดิษฐ์ในทางรังสีวินิจฉัย เพื่อเสริมสร้างความมั่นใจให้กับทั้งผู้ซื้อและผู้ใช้งานระบบปัญญาประดิษฐ์ รวมถึงผู้ป่วยที่ได้รับการวินิจฉัยเบื้องต้นด้วยระบบปัญญาประดิษฐ์ อีกทั้งยังเพิ่มความน่าเชื่อถือของบริษัทปัญญาประดิษฐ์ทั้งในและต่างประเทศที่ต้องการที่จะทดสอบระบบปัญญาประดิษฐ์ที่ตนเองพัฒนากับกลุ่มประชากรของไทย

แบบทดสอบปัญญาประดิษฐ์

การพัฒนาแบบทดสอบได้รับการสนับสนุนจาก ศูนย์ความเป็นเลิศด้านชีววิทยาศาสตร์ (TCELS), ราชวิทยาลัยรังสีแพทย์แห่งประเทศไทย (RCRT), มูลนิธิส่งเสริมวิชาการโรคปอดชนิดวินิจฉัยยาก (FORLD), คณะแพทยศาสตร์ มหาวิทยาลัยสงขลานครินทร์

ภาพในแบบทดสอบจำนวน 1,500 ภาพนั้น ได้รับการสนับสนุนจาก โรงพยาบาลสงขลานครินทร์ จังหวัดสงขลา, โรงพยาบาลเชียงรายประชารักษ์ จังหวัดเชียงราย, โรงพยาบาลอุดรธานี จังหวัดอุดรธานี, โรงพยาบาลสุทธาเวช จังหวัดมหาสารคาม, และกองวัณโรค กรมควบคุมโรค กระทรวงสาธารณสุข โดยทุกภาพจะถูกอ่านโดยรังสีแพทย์ B Reader จำนวน 3 ท่าน

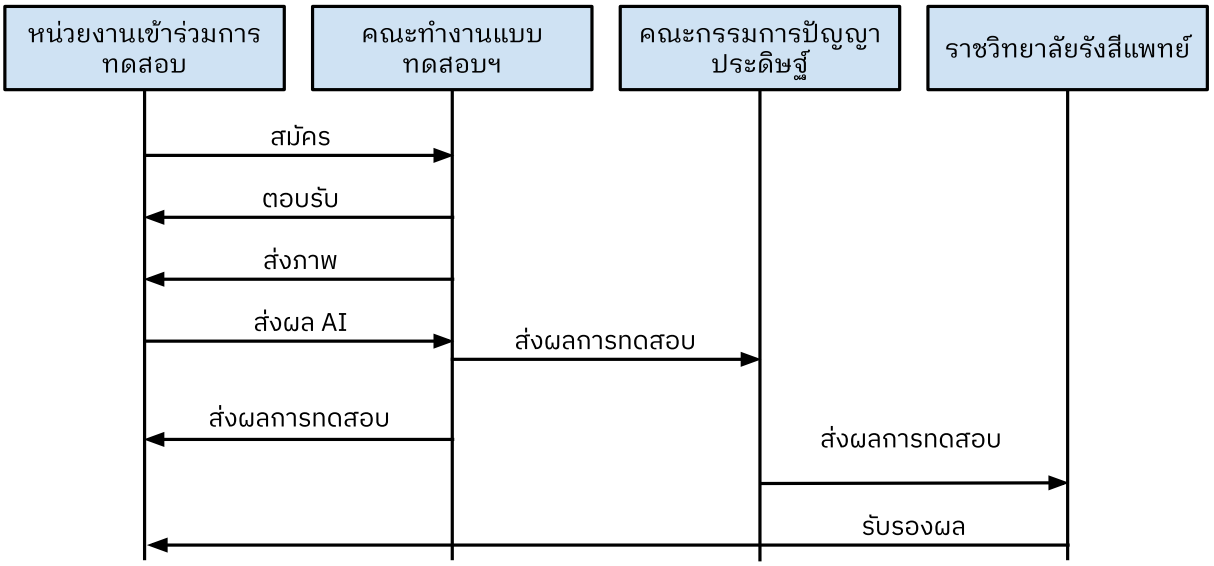
คุณลักษณะของภาพถ่ายรังสีในโครงการ

- ภาพถ่ายรังสีทรวงอกของผู้ป่วยที่มีอายุตั้งแต่ 15 ปี ด้วยเครื่องเอกซเรย์คอมพิวเตอร์
- ภาพถ่ายรังสีทรวงอกได้รับการวินิจฉัยวัณโรคปอด ต้องได้รับการยืนยันโดยผลตรวจเสมหะย้อมเชื้อ Acid-Fast Bacillus (AFB) เป็นบวก หรือ ผลตรวจเพาะเชื้อ Mycobacterium Tuberculosis เป็นบวก หรือ Polymerase Chain Reaction (PCR) for tuberculosis เป็นบวก และ ยืนยันโดยผลอ่านของรังสีแพทย์จากโรงพยาบาลที่ส่งภาพเข้าสู่โครงการฯ และยืนยันอีกครั้งโดยรังสีแพทย์ B Reader 3 ท่าน
- ภาพถ่ายรังสีทรวงอกที่ไม่ได้ได้รับการวินิจฉัยวัณโรคปอด ถูกยืนยันโดยผลอ่านของรังสีแพทย์จากโรงพยาบาลที่ส่งภาพเข้าสู่โครงการฯ และยืนยันอีกครั้งโดยรังสีแพทย์ B Reader
- ไม่มีภาพถ่ายจากผู้ป่วยที่มีสถานะ HIV Serology เป็นบวก
- ไม่มีภาพถ่ายจากผู้ป่วยที่มีการติดเชื้อที่ปอดแบบฉวยโอกาสอื่น ๆ หรือการติดเชื้อร่วม เช่น การติดเชื้อไมโครแบคทีเรีย ฮิสโตพลาสโมซิส ไครปโตคอคโคซิส เมลิออยโดซิส และแอคติโนมัยโคซิส

ภาพถ่ายแต่ละภาพจะถูกอ่านโดยรังสีแพทย์จำนวน 3 ท่าน โดยรังสีแพทย์ B Reader จะอ่านภาพและบันทึกข้อมูล Case Record Form ในภาคผนวก แต่ละรอยโรคที่บันทึกจะถูกยืนยันจากรังสีแพทย์ B Reader อย่างน้อยสองท่าน

ในการทดสอบนั้นโครงการจะสุ่มภาพให้กับผู้เข้าทำการทดสอบ โดยจะเป็นภาพถ่ายรังสีทรวงอกได้รับการวินิจฉัยวัณโรคปอดและภาพถ่ายรังสีทรวงอกที่ไม่ได้ได้รับการวินิจฉัยวัณโรคปอด ในอัตราส่วนที่เท่ากัน

ขั้นตอนการทดสอบปัญญาประดิษฐ์



1. หน่วยงานเข้าร่วมการทดสอบกรอกแบบฟอร์มเพื่อขอใช้งานแบบทดสอบ โดยการทดสอบจะแบ่งเป็น 2 ระดับ
    a. <u>ระดับที่ 1</u> หน่วยงานเข้าร่วมการทดสอบจะได้รับรายงานผลการทดสอบสำหรับรอยโรคที่ผ่านเกณฑ์
    b. <u>ระดับที่ 2</u> หน่วยงานเข้าร่วมการทดสอบจะได้รับรายงานผลการทดสอบแบบละเอียด รวมถึงทุกค่าเมตริกซ์ที่ใช้สำหรับรายงานผล และบทวิเคราะห์โดยนักวิทยาศาสตร์ข้อมูลทางการแพทย์และรังสีแพทย์
2. หน่วยงานเข้าร่วมการทดสอบตกลงว่าจะจับคู่รอยโรคใดของปัญญาประดิษฐ์ที่รายงานกับรอยโรคในที่มีในแบบทดสอบของโครงการ
3. หน่วยงานเข้าร่วมการทดสอบเซ็นต์เอกสาร Non-disclosure Agreement (NDA) กับทางโครงการเพื่อทำการยืนยันที่จะไม่ใช้ภาพชุดนี้สำหรับงานอื่น เช่น การเรียนรู้ของ AI และจะไม่เก็บภาพชุดนี้ไว้หลังจากการทดสอบเสร็จสิ้น
4. หน่วยงานเข้าร่วมการทดสอบเลือกแนวทางการทดสอบ ซึ่งแบ่งเป็น 2 แนวทาง คือ
    a. หน่วยงานเข้าร่วมการทดสอบส่ง API มาให้เจ้าหน้าที่ของโครงการ ซึ่งทางเจ้าหน้าที่จะทำการส่งภาพ De-identified DICOM เข้า API นั้นและนำผลลัพธ์ไปวิเคราะห์
    b. โครงการฯ ส่งภาพ De-identified DICOM ให้ Vendor และ Vendor ส่งผลการทดสอบให้กับโครงการฯ ตามแบบรายงานผลการทดสอบในภาคผนวก
5. โครงการส่งผลลัพธ์กลับไปยังหน่วยงานเข้าร่วมการทดสอบ ซึ่งจะใช้เวลาไม่เกิน 2 อาทิตย์หลังจากทางโครงการได้รับผลการทดสอบ
6. หากหน่วยงานเข้าร่วมการทดสอบต้องการใบรับรองปัญญาประดิษฐ์จากราชวิทยาลัยรังสีแพทย์ ทางโครงการจะต้องดำเนินการเพิ่มเติมเพื่อให้คณะกรรรมการฯ พิจารณา ซึ่งจะใช้เวลาเพิ่มเติมประมาณ 1 เดือน

การวิเคราะห์ผลการทดสอบ

ภาพถ่ายผู้ป่วยวัณโรคปอด (Ground Truth)

เพื่อวัดประสิทธิภาพของระบบปัญญาประดิษฐ์ ภาพถ่ายผู้ป่วยวัณโรคปอด และภาพถ่ายผู้ป่วยที่ไม่มีวัณโรคปอดที่ใช้ในโครงการทุกภาพเป็นภาพถ่ายที่ถูกยืนยันทั้งจากผลตรวจทางพยาธิวิทยา และจากรังสีแพทย์ B reader จำนวน 3 ท่าน โดยภาพนั้นจะถูกส่งมากจากโรงพยาบาลที่เข้าร่วมโครงการพัฒนาแบบทดสอบฯ และเป็นภาพที่ไม่ได้ถูกนำไปใช้พัฒนาระบบปัญญาประดิษฐ์ เพื่อให้ได้ชุดข้อมูล ground truth ที่สมบูรณ์ที่สุด รังสีแพทย์จะต้องกรอกข้อมูลในแบบบันทึกข้อมูลแบบละเอียด ซึ่งแบบบันทึกข้อมูลนั้นได้ถูกออกแบบมาจาก ILO classification guidelines (โปรดดูเอกสารแนบ)

ขนาดตัวอย่าง (Sample size)

ภาพถ่ายรังสีที่จะนำมาทดสอบปัญญาประดิษฐ์นั้นจะต้องมีความหลากหลาย ในโครงการนั้นเรามีภาพถ่ายรังสีจำนวน 1,500 ภาพ จากโรงพยาบาล 5 แห่งในประเทศไทย ซึ่งจะมีความใกล้เคียงสูงกับกลุ่มประชากรไทยซึ่งเป็นเป้าหมายของการทดสอบ ทั้งนี้จำนวนภาพที่จะถูกคัดเลือกมาทดสอบสำหรับปัญญาประดิษฐ์นั้นจะถูกเลือกแบบสุ่มเป็นจำนวนมากพอที่จะเป็นตัวแทนของกลุ่มประชากรเป้าหมาย ในการคำนวณขนาดตัวอย่าง เราได้ใช้หลักการของ Obuchowski *et al.* (2000) ซึ่งขนาดตัวอย่างถูกประมาณภายใต้เงื่อนไข type I error ที่ 5% หรือ $\alpha = 0.05$ (มีความเป็นไปได้ 5% ที่จะปฏิเสธสมมติฐานที่ว่ารังสีแพทย์และปัญญาประดิษฐ์มีความแม่นยำต่างกัน ทั้งที่ทั้งคู่มีความแม่นยำเท่ากัน) และ power ที่ 95% หรือ $\beta = 0.95$ (มีความเป็นไปได้ 95% ที่จะปฏิเสธสมมติฐานที่ว่ารังสีแพทย์และปัญญาประดิษฐ์มีความแม่นยำต่างกัน ทั้งที่ทั้งคู่มีความแม่นยำต่างกัน) ภายใต้เงื่อนไข ค่าพื้นนที่ใต้กราฟตัวรับ มากกว่าหรือเท่ากับ 0.90 (area under a receiver operating characteristic curve ≥ 0.90), ค่าความไวมากกว่าหรือเท่ากับ 0.80 (sensitivity ≥ 0.80) ที่อัตราส่วนผลบวกเทียม 0.10 (false positive rate of 0.10) และ ค่าความจำเพาะมากกว่าหรือเท่ากับ 0.80 (specificity ≥ 0.80) ที่อัตราส่วนผลลบเทียม 0.10 (false negative rate of 0.10) ในแบบทดสอบที่อัตราส่วนของภาพถ่ายรังสีผู้ป่วยวัณโรคปอดและผู้ป่วยไม่เป็นโรควัณโรคปอด 1:1 ขนาดตัวอย่างที่เหมาะสมคือภาพถ่ายรังสีจำนวน 300 ภาพ โดยเป็นภาพถ่ายรังสีของผู้ป่วยวัณโรคปอดจำนวนประมาณ 150 ภาพ และภาพถ่ายรังสีของผู้ป่วยที่ไม่เป็นวัณโรคปอดจำนวนประมาณ 150 ภาพซึ่งจำนวนนี้สามารถระบุได้ว่าปัญญาประดิษฐ์นั้นมีความสามารถมากน้อยเพียงไรในทางปฏิบัติสำหรับกลุ่มประชากรเป้าหมาย

เกณฑ์การวัดความสามารถ (Metrics)

การนำเสนอผลลัพธ์ของปัญญาประดิษฐ์นั้นทำได้หลายวิธี ซึ่งเกณฑ์การวัดความสามารถนั้นจะต้องสามารถบอกถึงความสามารถของปัญญาประดิษฐ์ได้ สำหรับการวัดความสามารถของปัญญาประดิษฐ์เทียบกับรังสีแพทย์ B Reader นั้นจะดำเนินการโดยเทียบความตรงภายในชุดข้อมูลของรังสีแพทย์ (B Rader vs B Reader) และความตรงภายนอกของรังสีแพทย์แต่ละท่าน (B Reader vs AI)

การวัดความตรงภายในชุดข้อมูลนั้น จะดำเนินการโดยใช้ตัววัด 3 ตัว ซึ่งตัววัดทั้งสามตัวนั้นสามารถสะท้อนความยากง่ายในการอ่านภาพรอยโรคแต่ละรอยของรังสีแพทย์ได้ คือ

- Pairwise Agreement คือ ค่าเฉลี่ยของความเหมือนกันกันระหว่างแต่ละคู่ของรังสีแพทย์
- Pairwise Cohen's Kappa คือ ค่าเฉลี่ยของความตรงระหว่างผู้รังสีแพทย์แต่ละคู่โดยใช้สถิติ Pairwise Cohen's Kappa
- Intraclass Agreement (ICC) คือ ค่าเฉลี่ยของ Pearson's correlation แบบ ICC(2,3) เมื่อมีรังสีแพทย์ 3 ท่าน อ่านภาพถ่ายรังสีแบบสุ่ม

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

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

นอกจากการวัดข้างต้นแล้ว ยังมีการใช้ Binary classification metrics เพื่อวัดประสิทธิภาพของปัญญาประดิษฐ์เทียบกับรังสีแพทย์ เนื่องจากแต่ละรอยโรคในแต่ละภาพนั้นจะถูกอ่านโดยรังสีแพทย์ 3 ท่าน รอยโรคที่ถูกยืนยันโดยรังสีแพทย์อย่างน้อย 2 ใน 3 ท่าน จะถูกนับเป็นรอยโรคที่เป็น Ground truth สำหรับวัดประสิทธิภาพการวินิจฉัยของปัญญาประดิษฐ์

การวินิจฉัยผิดพลาดของปัญญาประดิษฐ์นั้นแบ่งออกได้เป็นสองประเภท คือผลบวกลวง และผลลบลวง ผลบวกลวงคือการวินิจฉัยว่าผู้ป่วยเป็นโรคทั้งที่ผู้ป่วยไม่ได้เป็นโรค และผลลบลวงคือการวินิจฉัยว่าผู้ป่วยไม่ได้เป็นโรคทั้งที่ผู้ป่วยเป็นโรค เกณฑ์ทั้งสองข้อนี้เป็นสิ่งที่ใช้บ่อยที่สุดทางการแพทย์ ซึ่งจะสะท้อนผ่านค่าความไวและความจำเพาะ

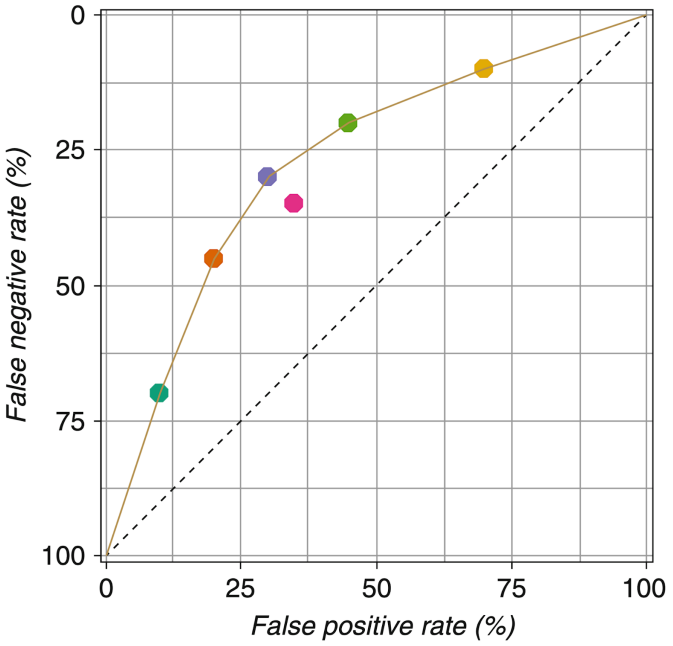
	Case Positive	**Case Negative**	
Predicted Positive	True Positive (TP)	False Positive (FP)	PPV = TP/(TP+FP)
Predicted Negative	False Negative (FN)	True Negative (TN)	NPV = TN/(FN+TN)
	Sensitivity = TP/(TP+FN)	Specificity = TN/(FP+TN)	

ประสิทธิภาพของปัญญาประดิษฐ์สำหรับวินิจฉัยรอยโรคจะได้รับการประเมินโดยการเปรียบเทียบอย่างครอบคลุมด้วยเมตริกการประเมินประสิทธิภาพหลายตัว เช่น ความไว (Sensitivity), ความจำเพาะ (Specificity) และพื้นที่ใต้เส้นโค้ง Receiver Operating Characteristic (ROC)

กำหนดให้ ผลบวกจริง (True Positive) คือ จำนวนผู้ป่วยได้ถูกวินิจฉัยว่าเป็นวัณโรคปอดได้อย่างถูกต้อง ผลบวกลวง (False Positive) คือ จำนวนผู้ป่วยได้ถูกวินิจฉัยว่าเป็นวัณโรคปอดแต่ไม่ได้เป็นโรค ผลลบจริง (True Negative) คือ จำนวนผู้ป่วยได้ถูกวินิจฉัยว่าไม่ได้เป็นเป็นวัณโรคปอดได้อย่างถูกต้อง และ ผลลบลวง (False Negative) คือ จำนวนผู้ป่วยได้ถูกวินิจฉัยว่าไม่ได้เป็นเป็นวัณโรคปอดแต่เป็นโรค

- **ความไว (Sensitivity)** คือ สัดส่วนของผลบวกที่เป็นจริงสำหรับภาวะนั้น ๆ (เช่น สัดส่วนของการตรวจพบวัณโรคปอดในผู้ที่ป่วยจริง)
- **ความจำเพาะ (Specificity)** คือ สัดส่วนของผลลบที่เป็นจริงสำหรับภาวะนั้น ๆ (เช่น สัดส่วนของการตรวจไม่พบวัณโรคปอดในผู้ที่ไม่ป่วย)

ความไว (Sensitivity) มีประโยชน์ในการวินิจฉัยแยกกันผลลบปลอม (False Negative) เพราะว่าการทดสอบยิ่งไวเท่าไร โอกาสการได้ผลลบ (เช่น การพบว่าไม่มีโรค) ที่ไม่เป็นจริง (เช่น ผู้ป่วยมีโรค) ก็น้อยลงเท่านั้น ความจำเพาะ (Specificity) มีประโยชน์ในการยืนยันภาวะที่มี โดยกันผลบวกปลอม (False Positive) เพราะว่าการทดสอบยิ่งจำเพาะเท่าไร โอกาสการได้ผลบวก (เช่น การพบว่ามีโรค) ที่ไม่เป็นจริง (เช่น ผู้ป่วยไม่มีโรค) ก็น้อยลงเท่านั้น ค่าความไวและค่าความจำเพาะนั้นเป็นค่าที่มีการแลกเปลี่ยนข้อดีข้อเสีย (Tradeoff) ระหว่างกัน โดยทั่วไปแล้วค่าทั้งสองค่านั้นจะต้องมีความสมดุลกัน แต่อย่างไรก็ตามค่าความไวและความจำเพาะนั้นสามารถเปลี่ยนแปลงได้ตามเกณฑ์การประเมิน (Diagnostic Threshold) ที่กำหนดไว้ยกตัวอย่างเช่น ในการทดสอบแยกว่าภาพถ่ายรังสีนั้นมีรอยโรคของโรควัณโรคปอดหรือไม่ ปัญญาประดิษฐ์อาจจะตั้งให้เกณฑ์การประเมินนั้นต่ำ ทำให้มีการส่งสัญญาณเตือนแม้ว่าการวินิจฉัยนั้นมีความเสี่ยงน้อยที่จะตรวจพบรอยโรค เพื่อลดโอกาสเสี่ยงที่จะพลาดการตรวจพบผู้ป่วย ในกรณีที่กำลังใช้แนวทางการค้นหาผู้ป่วยเชิงรุก (Active Case Finding) ในพื้นที่ที่มีความเสี่ยงสูง เช่น เรือนจำ เป็นต้น แต่อย่างไรก็ตาม ผลลบลวงนั้นอาจเป็นปัญหาร้ายแรงได้เนื่องจากการสูญเสียผลประโยชน์จากการแทรกแซงในช่วงต้น เช่นเดียวกันผลบวกลวงจำนวนมากอาจนำไปสู่ความเข้าใจที่ผิดเกี่ยวกับการแพร่กระจายของโรคได้ ซึ่งการเปลี่ยนแปลงของเกณฑ์การประเมิน (Diagnostic Threshold) นี้ ทำให้ค่าความไว (Sensitivity) และความจำเพาะ (Specificity) มีการเปลี่ยนแปลงตามไปด้วย ซึ่งสามารถแสดงให้เห็นได้ในกราฟเส้นโค้ง Receiver Operating Characteristic (ROC) กราฟ ROC เป็นการเทียบ 1 - Specificity ในแกนนอนและ Sensitivity ในแกนตั้ง พื้นที่ใต้กราฟยิ่งใกล้เคียงค่า 1 เพียงใดจะบ่งชี้ถึงประสิทธิภาพที่ดีขึ้นในการแยกแยะระหว่างภาพถ่ายรังสีวัณโรคปอดและภาพถ่ายรังสีที่ไม่ได้เป็นวัณโรคปอด



องค์การอาหารและยาของสหรัฐอเมริกาแนะนำให้ใช้เกณฑ์หลายตัวในการศึกษาปัญญาประดิษฐ์ทางด้านการแพทย์ อย่างน้อยที่สุด ควรมีการรายงานความไว (sensitivity) ความจำเพาะ (specificity) และ อัตราส่วนการทำนายที่ถูกต้อง (positive prediction rate, ppv) สำหรับงานจำแนกประเภทของภาพ เช่น การจำแนกว่าภาพถ่ายรังสีนั้นเป็นของผู้ป่วยวัณโรคปอดหรือผู้ป่วยที่ไม่เป็นวัณโรคปอด เป็นต้น ซึ่งการจำแนกข้อมูลเป็นสองกลุ่มนี้ หากผู้ผลิตปัญญาประดิษฐ์คืนค่าผลลัพธ์เป็นคะแนน 0 - 1 หรือ 0 - 100 จำเป็นที่จะต้องเกณฑ์การวินิจฉัย (diagnostic threshold หรือ cut-off value) มาด้วย อีกทั้งการแสดงผลลัพธ์เหล่านี้บทกราฟ ROC เมื่อมีการเปลี่ยนแปลงเกณฑ์การวินิจฉัย (diagnostic threshold) และคำนวณ AUROC ก็เป็นสิ่งที่องค์การอาหารและยาของสหรัฐอเมริกาแนะนำเพื่อใช้เป็นการเปรียบเทียบปัญญาประดิษฐ์

ดังนั้นการรายงานผลการทดสอบปัญญาประดิษฐ์ของโครงการ จะมีการรายงานทั้ง Pairwise Agreement, Intraclass Agreement และ Pairwise Cohen's Kappa ซึ่งสะท้อนความตรงกันภายในและภายนอกของชุดข้อมูลกับปัญญาประดิษฐ์ รวมถึง ความไว (sensitivity), ความจำเพาะ (specificity), อัตราส่วนการทำนายผู้ป่วยที่เป็นโรคที่ถูกต้อง (positive prediction rate, ppv) และ อัตราส่วนการทำนายผู้ป่วยที่ไม่เป็นโรคที่ถูกต้อง (negative prediction rate, ppv) โดยผู้ผลิตจะต้องระบุเกณฑ์การวินิจฉัย (diagnostic threshold หรือ cut-off value) มาด้วย ซึ่งจะสะท้อนประสิทธิภาพของปัญญาประดิษฐ์สำหรับวินิจฉัยรอยโรค และจะมีการรายงานของเส้นโค้ง ROC และพื้นที่ใต้กราฟ AUROC เช่นกัน

การรายงานผลลัพธ์จะประกอบด้วยตารางจำนวน ภ ตาราง คือ

1. จำนวนรอยโรคที่ระบุโดยรังสีแพทย์
2. ความตรงภายในระหว่างรังสีแพทย์
3. ความตรงภายนอกเทียบรังสีแพทย์กับปัญญาประดิษฐ์
4. ประสิทธิภาพในการวินิจฉัยของแต่ละรอยโรคของปัญญาประดิษฐ์

ตารางที่ 1 จำนวนรอยโรคที่ระบุโดยรังสีแพทย์

Finding	N_{Non-TB}	N_{TB}	$N_{Concensus}$
Abnormality 1	:	:	:
:	:	:	:

ตารางที่ 2 ความตรงภายในระหว่างรังสีแพทย์

Finding	Agreement	ICC	Cohen's	Fleiss'
Abnormality 1	:	:	:	:
:	:	:	:	:

ตารางที่ 3 ความตรงภายนอกเทียบรังสีแพทย์กับปัญญาประดิษฐ์

Finding	N	Threshold	Pairwise Agreement		Cohen's Kappa	
			"B" vs "B"	"B" vs AI	"B" vs "B"	"B" vs AI
Abnormality 1	:	:	:	:	:	:
:	:	:	:	:	:	:

ตารางที่ 4 ประสิทธิภาพในการวินิจฉัยของแต่ละรอยโรคของปัญญาประดิษฐ์

Finding	N	Threshold	Specificity	Sensitivity	PPV	NPV	AUROC
Abnormality 1	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:

สถิติของแบบทดสอบ

หากสุ่มภาพจำนวน 300 ภาพจากโครงการเพื่อใช้ในการทดสอบ โดยมีอัตราส่วนของผู้ป่วยวันโรคปอดเท่ากับผู้ป่วยที่ไม่เป็นวัณโรคปอด จะได้จำนวนรอยโรคดังต่อไปนี้

Abnormalities	N
Pulmonary tuberculosis	150
Small opacities	120 - 132
Large opacity	114 - 126
Mass/Nodule	36 - 54
Cavity	72 - 84
Fibrosis	66 - 78
Calcification	18 - 30
Pleural effusion	18 - 30
Pleural thickening/calcification	66 - 78
Pneumothorax	1 - 3
Hilar adenopathy	66 - 78
Mediastinal adenopathy	0

เอกสารอ้างอิง

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. In Biometrics (Vol. 33, Issue 1, p. 159). JSTOR.

Qin, Z. Z., Ahmed, S., Sarker, M. S., Paul, K., Adel, A. S. S., Naheyan, T., Barrett, R., Banu, S., & Creswell, J. (2021). Tuberculosis detection from chest x-rays for triaging in a high tuberculosis-burden setting: an evaluation of five artificial intelligence algorithms. In The Lancet Digital Health (Vol. 3, Issue 9, pp. e543–e554). Elsevier BV.

Obuchowski, N. A. (2000). Sample Size Tables For Receiver Operating Characteristic Studies. In American Journal of Roentgenology (Vol. 175, Issue 3, pp. 603–608). American Roentgen Ray Society.

U.S. Food and Drug Administration. Software as a medical device: clinical evaluation. 2017.

(อ.๑) แบบบันทึกข้อมูลหน่วยงานที่เข้าร่วมการทดสอบ

ข้อมูลหน่วยงาน	
หน่วยงาน	
ที่อยู่	
ชื่อผู้ติดต่อ	
เบอร์โทรศัพท์ผู้ติดต่อ	
อีเมล์ผู้ติดต่อ	
ข้อมูลปัญญาประดิษฐ์	
ชื่อปัญญาประดิษฐ์	
เวอร์ชั่น	
บริษัทผู้พัฒนา	
ข้อมูลการทดสอบ	
รูปแบบการทดสอบ	☐ ทดสอบโดยให้โครงการส่งภาพถ่ายเข้าสู่ระบบผ่านทาง API รายละเอียด ☐ ทดสอบโดยใช้ไฟล์ De-identified DICOMs จากทางโครงการ
จำนวนภาพถ่าย	
สำหรับเจ้าหน้าที่	
Hash Set	

(อ.๒) รายการความผิดปกติของปอดในแบบทดสอบ

หน่วยงานที่ทำการทดสอบโปรดกรอกข้อมูลในช่อง Vendor Abnormality

Abnormality	Characteristics	Confidence Level	Distribution	Vendor Abnormality
Lung parenchymal abnormalities				
Small opacities	- Primary Nodular - Primary Reticular - Secondary Nodular - Secondary Reticular	- Low - Medium - High	- Upper Right - Upper Left - Middle Right - Middle Left - Lower Right - Lower Left	
Large opacity		- Low - Medium - High	- Upper Right - Upper Left - Middle Right - Middle Left - Lower Right - Lower Left	
Mass/Nodule		- Low - Medium - High	- Upper Right - Upper Left - Middle Right - Middle Left - Lower Right - Lower Left	
Cavity		- Low - Medium - High	- Upper Right - Upper Left - Middle Right - Middle Left - Lower Right - Lower Left	
Fibrosis		- Low - Medium - High	- Upper Right - Upper Left - Middle Right - Middle Left - Lower Right - Lower Left	

Calcification		- Low - Medium - High	- Upper Right - Upper Left - Middle Right - Middle Left - Lower Right - Lower Left	
Pleural abnormalities				
Pleural effusion		- Low - Medium - High	- Right - Left	
Pleural thickening/ calcification		- Low - Medium - High	- Right - Left	
Pneumothorax		- Low - Medium - High	- Right - Left	
Other abnormalities				
Hilar adenopathy		- Low - Medium - High	- Right - Left	
Mediastinal adenopathy		- Low - Medium - High		
Consistency with post primary pulmonary tuberculosis				
Active pulmonary tuberculosis	- Patchy infiltration - Cavity with surrounding consolidation - Unilateral hilar/paratracheal LN enlargement - Pleural effusion - Military nodules			
Indeterminate pulmonary tuberculosis	- Reticulonodular infiltration - Destroyed lung or bronchiectasis			
Inconsistent with pulmonary tuberculosis				

(อ.๓) ตัวอย่างผลการทดสอบส่งคืนให้โครงการ

รหัสภาพ	คะแนน (0 – 1)			
	Abnormality #1	**Abnormality #2**	**Abnormality #3**	**Abnormality #4**
a75a39c2	0.1021	0.3221	0.2421	0.04
:	:	:	:	:
:	:	:	:	w