

การประเมินคุณภาพของคำตอบจาก ChatGPT สำหรับคำถามด้านข้อมูลยา : การศึกษาในรายการยาที่คัดเลือกจากบัญชียาหลักแห่งชาติ

สมหมาย พิมพ์อูบ, จิตติมา พุ่งส่วย

กลุ่มงานเภสัชกรรม โรงพยาบาลอุดรดิตถ์

บทคัดย่อ

วัตถุประสงค์: เพื่อประเมินคุณภาพคำตอบที่ได้จาก ChatGPT เมื่อสอบถามข้อมูลเกี่ยวกับยาในบัญชียาหลักแห่งชาติ โดยประเมินในด้านความถูกต้อง ความเหมาะสม ความครบถ้วน ความสม่ำเสมอ ความเข้าใจง่ายของภาษา และความปลอดภัย รวมทั้งวิเคราะห์ปัจจัยที่สัมพันธ์กับการผ่านเกณฑ์คุณภาพของคำตอบ **วิธีการ:** การวิจัยเชิงวิเคราะห์แบบภาคตัดขวางครั้งนี้ ประเมินคำตอบจาก ChatGPT จำนวน 960 คำตอบเมื่อถามด้วยคำถามที่เกี่ยวกับข้อบ่งใช้ ขนาดยา วิธีใช้ยา และอาการไม่พึงประสงค์ของยาในบัญชียาหลักแห่งชาติ ทั้งยาแผนปัจจุบันและยาสมุนไพร โดยถามด้วยคำถามแบบปลายเปิดและปลายปิดใน 4 ช่วงเวลา คือ วันที่ 0, 7, 14 และ 28 โดยการถามในวันที่ 0 จะสอบถาม 3 ครั้งแบบต่อเนื่อง แต่การสอบถามในวันอื่น ๆ เป็นการสอบถามเพียงครั้งเดียว เภสัชกรผู้เชี่ยวชาญ 2 คนประเมินคำตอบอย่างเป็นอิสระต่อกันโดยใช้มาตรวัดลิเคิร์ต 5 ระดับใน 6 มิติของการประเมิน การวิเคราะห์ความสอดคล้องระหว่างผู้ประเมินใช้ weighted Cohen's kappa และ intraclass correlation coefficient (ICC) การวิเคราะห์ปัจจัยที่สัมพันธ์กับการผ่านเกณฑ์คุณภาพของคำตอบใช้ generalized estimating equations (GEE) **ผลการวิจัย:** คำตอบจาก ChatGPT มีคะแนนเฉลี่ยมากกว่า 4.0 จากคะแนนเต็ม 5 ในทุกมิติการประเมิน โดยด้านความถูกต้องมีคะแนนเฉลี่ยสูงสุด (4.60 ± 0.67) รองลงมาคือความปลอดภัย (4.43 ± 0.69) และความเหมาะสม (4.40 ± 0.69) ในทุกมิติมีสัดส่วนของคำตอบที่ผ่านเกณฑ์การประเมิน (ได้คะแนน ≥ 4) มากกว่าร้อยละ 92 ความสอดคล้องระหว่างผู้ประเมินอยู่ในระดับต่ำในทุกด้านของการประเมิน ผลการวิเคราะห์ด้วย GEE ไม่พบความสัมพันธ์อย่างมีนัยสำคัญทางสถิติระหว่างโอกาสการผ่านเกณฑ์คุณภาพของคำตอบ กับหมวดคำถาม ประเภทยา ประเภทคำถาม หรือช่วงเวลาการประเมิน **สรุป:** ChatGPT มีศักยภาพในการให้ข้อมูลยาในบัญชียาหลักแห่งชาติ โดยคุณภาพของคำตอบอยู่ในระดับสูงในทุกมิติของการประเมิน การศึกษาไม่พบความแตกต่างของคุณภาพคำตอบตามช่วงเวลาการประเมินในการศึกษานี้ อย่างไรก็ตาม ChatGPT ยังอาจมีข้อจำกัดด้านความครบถ้วนและความสม่ำเสมอในบางบริบท และไม่ควรใช้ทดแทนการตัดสินใจทางคลินิกหรือการให้คำแนะนำโดยบุคลากรทางการแพทย์และสาธารณสุข ทั้งนี้ ควรมีการศึกษาเพิ่มเติมในบริบททางคลินิกที่หลากหลายและซับซ้อนมากขึ้นเพื่อสนับสนุนการประยุกต์ใช้ปัญญาประดิษฐ์เพื่อตอบคำถามเกี่ยวกับข้อมูลยาอย่างเหมาะสมและปลอดภัย

คำสำคัญ: แชนจีพีที ปัญญาประดิษฐ์ บัญชียาหลักแห่งชาติ ข้อมูลยา คุณภาพของคำตอบ

รับต้นฉบับ: 3 พ.ค. 2569, **ได้รับบทความฉบับปรับปรุง:** 7 มิ.ย. 2569, **รับลงตีพิมพ์:** 4 ก.ค. 2569

ผู้ประสานงานบทความ: สมหมาย พิมพ์อูบ กลุ่มงานเภสัชกรรม โรงพยาบาลอุดรดิตถ์ อำเภอเมือง จังหวัดอุดรดิตถ์ 53000

E-mail: smptong@hotmail.com

Evaluation of ChatGPT Response Quality for Drug Information Queries: A Study of Medications Selected from National List of Essential Medicines

Sommai Pimoaub, Thitima Tungsuay

Pharmacy Department, Uttaradit Hospital

Abstract

Objective: To evaluate the quality of responses generated by ChatGPT regarding medications listed in the National List of Essential Medicines (NLEM), focusing on accuracy, appropriateness, completeness, consistency, readability, and safety, and to analyze factors associated with meeting the quality threshold. **Methods:** This cross-sectional analytical study evaluated 960 ChatGPT responses concerning indications, dosages, administration, and adverse drug reactions of both modern and herbal medicines in the NLEM. Questions included both open-ended and closed-ended formats, administered across four time points: days 0, 7, 14, and 28. On day 0, three consecutive queries were performed, whereas subsequent days involved a single query. Two expert pharmacists independently assessed the responses using a 5-point Likert scale across six dimensions. Inter-rater reliability was analyzed using weighted Cohen's kappa and intraclass correlation coefficient (ICC), while factors associated with passing the quality threshold were analyzed using generalized estimating equations (GEE). **Results:** Responses generated by ChatGPT achieved mean scores exceeding 4.0 out of 5 across all evaluation dimensions. The highest mean score was observed in accuracy (4.60 ± 0.67), followed by safety (4.43 ± 0.69) and appropriateness (4.40 ± 0.69). More than 92% of responses in all dimensions met the quality threshold (score ≥ 4). Inter-rater agreement was low across all domains. GEE analysis revealed no statistically significant associations between the likelihood of meeting the quality threshold and question category, drug class, question type, or assessment time point. **Conclusion:** ChatGPT demonstrates significant potential in providing information on NLEM medications, with high-quality responses across all evaluated dimensions. This study found no significant variations in response quality over the assessment period. However, ChatGPT may still encounter limitations regarding completeness and consistency in certain contexts. Therefore, it should not replace clinical decision-making or professional counseling by healthcare providers. Further studies in more diverse and complex clinical contexts are warranted to support the appropriate and safe application of artificial intelligence for medication-related inquiries.

Keywords: ChatGPT, artificial intelligence, National List of Essential Medicines, drug information, response quality

บทนำ

จากการสำรวจอนามัยและสวัสดิการ พ.ศ. 2564 และการสำรวจสุขภาพประชาชนไทยโดยการตรวจร่างกาย ครั้งที่ 6 พ.ศ. 2562–2563 รายงานว่า ประชากรไทย โดยเฉพาะผู้สูงอายุมีภาวะโรคไม่ติดต่อเรื้อรังสูงและโรคที่พบบ่อย ได้แก่ ความดันโลหิตสูง เบาหวาน และภาวะไขมันในเลือดสูง (1,2) ผู้ป่วยโรคเรื้อรังเหล่านี้จำเป็นต้องได้รับยาอย่างต่อเนื่องเพื่อควบคุมอาการของโรคและลดความเสี่ยงต่อภาวะแทรกซ้อนระยะยาว นอกจากนี้ ปัญหาสุขภาพที่พบบ่อยในเวชปฏิบัติทั่วไปยังครอบคลุมถึงโรคติดเชื้อ อากาศปวด และโรคหรืออาการที่เกี่ยวข้องกับระบบทางเดินอาหาร ซึ่งมักต้องอาศัยการใช้ยาอย่างเหมาะสมเพื่อการรักษา บรรเทาอาการ และป้องกันภาวะแทรกซ้อน

บัญชียาหลักแห่งชาติเป็นรายการยาสำคัญของประเทศที่ใช้เป็นฐานในการกำหนดการเข้าถึงยาและการใช้ยาที่เหมาะสม โดยครอบคลุมยาที่จำเป็นสำหรับการดูแลปัญหาสุขภาพสำคัญของประชาชนไทย (3) ยาสำหรับควบคุมโรคเรื้อรัง เช่น ยาลดน้ำตาลในเลือด ยาลดความดันโลหิต และยาลดไขมันในเลือด เป็นกลุ่มยาที่มีความสำคัญต่อการดูแลผู้ป่วยโรคเรื้อรังในระบบบริการสุขภาพของประเทศไทย (1,2) นอกจากนี้บัญชียาหลักแห่งชาติสำหรับยาแผนปัจจุบันแล้ว ประเทศไทยยังจัดทำบัญชียาสมุนไพรแห่งชาติด้วย ซึ่งสะท้อนบทบาทของยาสมุนไพรในระบบสุขภาพของประเทศ (3,4)

การใช้ยาอย่างไม่สมเหตุผลโดยเฉพาะการใช้ยาปฏิชีวนะโดยไม่มีข้อบ่งชี้ เป็นปัญหาสำคัญด้านการใช้ยาในประเทศไทย เนื่องจากเพิ่มความเสี่ยงต่อการดื้อยา ภาวะแทรกซ้อน และภาระค่าใช้จ่ายด้านสุขภาพ จึงมีการส่งเสริมการใช้ยาอย่างสมเหตุผลมาอย่างต่อเนื่อง (5) ในบริบทของสังคมดิจิทัลปัจจุบัน การสืบค้นข้อมูลสุขภาพผ่านช่องทางออนไลน์และการใช้แบบจำลองภาษาขนาดใหญ่ (large language models; LLMs) เช่น ChatGPT ได้รับความสนใจเพิ่มขึ้นอย่างต่อเนื่อง (6) อย่างไรก็ตาม การได้รับข้อมูลสุขภาพจากแหล่งดิจิทัลที่ขาดการตรวจสอบหรือมีข้อมูลคลาดเคลื่อนอาจนำไปสู่ความเข้าใจผิดและการตัดสินใจด้านสุขภาพที่ไม่เหมาะสม โดยเฉพาะในสภาพแวดล้อมออนไลน์ที่มีทั้งข้อมูลที่น่าเชื่อถือหรือชวนให้เข้าใจผิดปะปนกันอยู่เป็นจำนวนมาก (7)

ปัจจุบัน ChatGPT ซึ่งเป็นระบบปัญญาประดิษฐ์บนพื้นฐานของ LLMs สามารถรับข้อความป้อนเข้าเพื่อ

สร้างคำตอบในลักษณะโต้ตอบคล้ายบทสนทนาจบมนุษย์ (8) แชตบอตที่ขับเคลื่อนด้วย LLMs เหล่านี้ ได้รับความสนใจและถูกนำมาใช้สืบค้นข้อมูลสุขภาพอย่างแพร่หลายเพิ่มมากขึ้น (6) โดย ChatGPT มีขีดความสามารถในการให้ข้อมูลทางการแพทย์ในระดับที่เทียบเคียงกับข้อมูลออนไลน์ทั่วไป และแม้ว่าคุณภาพโดยรวมในปัจจุบันยังมีข้อจำกัด แต่ก็มีแนวโน้มที่จะถูกพัฒนาต่อยอดเป็นแหล่งข้อมูลสำหรับผู้ป่วยและบุคลากรทางการแพทย์ในอนาคต (9) นอกจากนี้ มีรายงานว่า ประชาชนมากกว่าร้อยละ 70 มีความเต็มใจที่จะใช้ระบบปัญญาประดิษฐ์เพื่อสืบค้นข้อมูลสุขภาพและวินิจฉัยตนเอง (6,10) อย่างไรก็ตาม แม้เทคโนโลยีดังกล่าวจะมีศักยภาพในการให้ข้อมูลสุขภาพและข้อมูลยาสูง แต่ข้อมูลที่สร้างขึ้นนั้น อาจยังมีข้อจำกัดด้านความถูกต้องและความน่าเชื่อถือค่อนข้างมาก โดยเฉพาะเมื่อขาดข้อมูลเฉพาะบุคคลของผู้ป่วย (11) นอกจากนี้ ปัญญาประดิษฐ์ประเภท LLMs เช่น ChatGPT นี้ อาจสร้างข้อผิดพลาด (hallucination) ซึ่งเป็นการสร้างคำตอบที่มีความสมเหตุสมผลและน่าเชื่อถือในลักษณะภาษาธรรมชาติของมนุษย์ แต่ไม่สอดคล้องกับข้อเท็จจริงหรือหลักฐานที่ถูกต้อง ส่งผลให้ผู้ใช้งานเชื่อถือในข้อมูลดังกล่าว

ในประเทศไทย เกษัชกรในสถานพยาบาลมีบทบาทหลักในการให้ข้อมูลด้านยา คำแนะนำด้านสุขภาพเบื้องต้น การติดตามการใช้ยา และการจัดการปัญหาที่เกี่ยวข้องกับยาแก่ประชาชน ซึ่งเป็นกลไกสำคัญส่วนหนึ่งของการพัฒนาระบบบริการสุขภาพปฐมภูมิของประเทศ พร้อมทั้งส่งเสริมการใช้ยาอย่างสมเหตุผลเพื่อผลลัพธ์ทางสุขภาพที่ดีขึ้นของบุคคล ขณะเดียวกัน ความก้าวหน้าของเทคโนโลยีปัญญาประดิษฐ์ โดยเฉพาะระบบ LLMs ได้ก่อให้เกิดเครื่องมือรูปแบบใหม่ เช่น ChatGPT ซึ่งอาจช่วยสนับสนุนการเข้าถึงข้อมูลสุขภาพ การให้ข้อมูลยา และการตอบคำถามทางคลินิกในบางบริบท อย่างไรก็ตาม หลักฐานเชิงประจักษ์ในประเทศไทยชี้ให้เห็นว่า ความถูกต้องของคำตอบจากแชตบอตยังมีข้อจำกัดอยู่มาก จึงควรใช้เป็นเครื่องมือเสริมควบคู่กับการดูแลโดยเภสัชกรและผู้เชี่ยวชาญด้านสุขภาพอย่างใกล้ชิด (10)

จากการทบทวนวรรณกรรมอย่างเป็นระบบพบว่า คุณภาพของคำตอบที่สร้างโดยระบบปัญญาประดิษฐ์ในการแพทย์มีความแปรปรวนแตกต่างกันไปตามลักษณะคำถามและบริบททางคลินิก อีกทั้งยังมีข้อจำกัดด้านความถูกต้อง ความครบถ้วน และความเหมาะสมของข้อมูลใน

บางกรณี ส่งผลให้ประเด็นด้านความปลอดภัยและความน่าเชื่อถือยังคงเป็นข้อพิจารณาสำคัญก่อนนำข้อมูลไปใช้ในการจัดบริการสุขภาพ (12) แม้ว่าจะมีการประเมินคุณภาพของคำตอบจาก ChatGPT ในบริบทด้านสุขภาพและการแพทย์อยู่บ้างแล้ว แต่ผลการศึกษาที่มีความหลากหลายแตกต่างกันตามลักษณะคำถามและบริบทที่ประเมิน อีกทั้งหลักฐานส่วนใหญ่ยังมาจากการศึกษาในต่างประเทศและมุ่งประเมินคำตอบทางการแพทย์ในภาพรวม ขณะที่หลักฐานเชิงประจักษ์เกี่ยวกับคุณภาพของคำตอบด้านข้อมูลยาในบริบทของประเทศไทยยังมีจำกัด ถึงแม้จะเริ่มมีการศึกษาเพื่อประเมินประสิทธิภาพของเครื่องมือปัญญาประดิษฐ์ในการตอบคำถามข้อมูลยาในภาษาไทยบ้างแล้วก็ตาม (10) แต่ปัจจุบัน ยังไม่มีข้อมูลเพียงพอเกี่ยวกับคุณภาพของคำตอบสำหรับยาในบัญชียาหลักแห่งชาติในหัวข้อสำคัญ ได้แก่ ข้อบ่งใช้ ขนาดยา วิธีใช้ และอาการไม่พึงประสงค์ ซึ่งเป็นข้อมูลที่เกี่ยวข้องโดยตรงกับการใช้ยาอย่างเหมาะสม และการให้คำปรึกษาทางเภสัชกรรม

การศึกษามีวัตถุประสงค์เพื่อประเมินคุณภาพของคำตอบจาก ChatGPT ต่อคำถามด้านข้อมูลยาที่คัดเลือกจากบัญชียาหลักแห่งชาติของไทย ในหัวข้อข้อบ่งใช้ ขนาดยา วิธีใช้ และอาการไม่พึงประสงค์ โดยประยุกต์กรอบแนวคิดคุณภาพข้อมูลของ Wang และ Strong (13) ร่วมกับเกณฑ์การประเมินในบริบทข้อมูลสุขภาพและข้อมูลยา เพื่อประเมินด้านความถูกต้อง ความเหมาะสม ความครบถ้วน ความสม่ำเสมอ ความเข้าใจง่ายของภาษา และความปลอดภัยของคำตอบ รวมทั้งวิเคราะห์ปัจจัยที่สัมพันธ์กับการผ่านเกณฑ์คุณภาพโดยรวมของคำตอบ ผลการศึกษาที่ได้จะเป็นข้อมูลสำคัญในการทำความเข้าใจคุณภาพของคำตอบจาก LLMs ในบริบทข้อมูลยาของประเทศไทย และนำไปสู่การกำหนดแนวทางการใช้ปัญญาประดิษฐ์อย่างเหมาะสม การส่งเสริมการใช้ยาอย่างสมเหตุผล ตลอดจนการพัฒนาเครื่องมือสุขภาพดิจิทัลที่มีความปลอดภัยและสอดคล้องกับบริบทของประเทศไทยในอนาคต

วิธีการวิจัย

รูปแบบการวิจัย

งานวิจัยนี้เป็นการวิจัยเชิงพรรณนาเชิงสังเกตที่มีการวัดซ้ำ การวิจัยนี้ได้รับการรับรองจากคณะกรรมการจริยธรรมการวิจัยในมนุษย์ โรงพยาบาลอุดรดิตถ์ (REC No. 16/2567) การศึกษาไม่มีการเก็บรวบรวมข้อมูลส่วน

บุคคล ข้อมูลสุขภาพ หรือข้อมูลที่สามารถระบุตัวบุคคลได้ คำตอบที่ได้จาก ChatGPT ถูกใช้เพื่อวัตถุประสงค์ทางการวิจัยและการประเมินคุณภาพของระบบเท่านั้น ทั้งนี้ การดำเนินการวิจัยเป็นไปตามหลักจริยธรรมการวิจัยและแนวปฏิบัติการวิจัยที่ดี

การคัดเลือกตัวอย่างยา

การศึกษานี้คัดเลือกตัวอย่างยาจำนวน 20 รายการโดยวิธีการเลือกแบบเจาะจง ซึ่งครอบคลุมรายการยาที่ใช้สำหรับปัญหาสุขภาพสำคัญของประเทศและสะท้อนความหลากหลายของข้อมูลยา ได้แก่ โรคไม่ติดต่อเรื้อรัง โรคติดเชื้อ และอาการเจ็บป่วยทั่วไป ทั้งนี้ต้องเป็นรายการยาในบัญชียาหลักแห่งชาติและบัญชียาจากสมุนไพรแห่งชาติ (1-4) จำนวนรายการยาที่เลือกศึกษาไม่ได้กำหนดจากการคำนวณขนาดตัวอย่างทางสถิติ แต่กำหนดเพื่อให้มีความหลากหลายของรายการยาที่ศึกษา รวมถึงความเป็นไปได้ในการประเมินคุณภาพคำตอบโดยผู้เชี่ยวชาญหลายคนภายใต้ทรัพยากรและระยะเวลาที่กำหนด

ตัวอย่างยาที่ใช้ศึกษาประกอบด้วยยาแผนปัจจุบันจำนวน 10 รายการ ตามบัญชียาหลักแห่งชาติ พ.ศ. 2565 (3) ได้แก่ metformin, losartan, simvastatin, enalapril, clopidogrel, amlodipine, amoxicillin, paracetamol, salbutamol และ omeprazole และยาสมุนไพรจำนวน 10 รายการตามบัญชียาจากสมุนไพรแห่งชาติ พ.ศ. 2566 (4) ได้แก่ มะระขี้นก ขมิ้นชัน เถาวัลย์เปรียง ยาเพชรสังฆาต พ้าทะลายโจร ขิง ยาอมมะแว้ง มะขามแขก ยาสหัชคารา และยาตรีผลา

การพัฒนาชุดคำถามและชุดคำสั่ง

ชุดคำถามมาตรฐานครอบคลุมข้อมูล 4 ด้าน ได้แก่ ข้อบ่งใช้ ขนาดยา วิธีใช้ และอาการไม่พึงประสงค์ โดยแต่ละด้านประกอบด้วยคำถาม 2 รูปแบบ ได้แก่ คำถามปลายเปิด และคำถามแบบเลือกตอบดังแสดงในตารางที่ 1 สำหรับคำถามแบบเลือกตอบมี 4 ตัวเลือก โดยมีคำตอบที่ถูกต้อง 1 ตัวเลือก คำถามพัฒนาจากแหล่งอ้างอิงมาตรฐานด้านข้อมูลยา การกำหนดให้มีคำถามทั้งสองลักษณะนี้ มีวัตถุประสงค์เพื่อเปรียบเทียบความสามารถของ ChatGPT ในการสร้างคำตอบภายใต้โครงสร้างของคำถามที่แตกต่างกัน โดยคำถามปลายเปิดสะท้อนลักษณะการใช้งานจริงของประชาชนที่มักสอบถามข้อมูลยาในลักษณะอิสระ ขณะที่คำถามแบบเลือกตอบใช้ประเมินความสามารถในการระบุคำตอบที่เหมาะสมภายใต้ตัวเลือกที่กำหนดไว้ ตัวอย่างของ

ตารางที่ 1. ตัวอย่างคำถามมาตรฐานสำหรับการประเมินข้อมูลยาในแต่ละด้าน

หมวด	จำนวน คำถาม	ตัวอย่างคำถามปลายเปิด	ตัวอย่างคำถามแบบเลือกตอบ
ข้อบ่งใช้	1	ยาเมตฟอร์มิน (metformin) มีข้อบ่งใช้ในการรักษาโรคหรือบรรเทาอาการอะไร	ยาเมตฟอร์มิน (metformin) มีข้อบ่งใช้ในการรักษาโรคหรือบรรเทาอาการในข้อใด a. ความดันโลหิตสูง b. เบาหวานชนิดที่ 2 c. ภาวะไขมันในเลือดสูง d. ภาวะกรดยูริกในเลือดสูง
ขนาดยา	1	ขนาดยาที่เหมาะสมที่ใช้รักษาโรคเบาหวานชนิดที่ 2 ในผู้ใหญ่ที่ค่าไตปกติของยาเมตฟอร์มิน (metformin) คือขนาดเท่าใด	ขนาดยาที่เหมาะสมที่ใช้รักษาโรคเบาหวานชนิดที่ 2 ในผู้ใหญ่ที่มีค่าไตปกติของเมตฟอร์มิน (metformin) คือข้อใด a. 100 มิลลิกรัม วันละครั้ง b. 250 มิลลิกรัม วันละ 2 ครั้ง c. 500–1000 มิลลิกรัม วันละ 2 ครั้ง d. 2000 มิลลิกรัม ครั้งเดียวก่อนนอน
วิธีใช้	1	วิธีใช้ยาเมตฟอร์มิน (metformin) ที่ถูกต้องเหมาะสมในการรักษาโรคเบาหวานชนิดที่ 2 คือข้อใด	วิธีใช้ยาเมตฟอร์มิน (metformin) ที่ถูกต้องเหมาะสมในการรักษาโรคเบาหวานชนิดที่ 2 คือข้อใด a. ก่อนอาหาร b. ขณะท้องว่างตอนเช้า c. พร้อมหรือหลังอาหารเพื่อป้องกันอาการไม่สบายท้อง d. ดอนก่อนนอน
อาการไม่พึงประสงค์	1	ผลข้างเคียงที่พบบ่อยของยาเมตฟอร์มิน (metformin) ในการรักษาโรคเบาหวานชนิดที่ 2 คืออะไร	ผลข้างเคียงที่พบบ่อยของยาเมตฟอร์มิน (metformin) ในการรักษาโรคเบาหวานชนิดที่ 2 คือข้อใด a. นอนไม่หลับ b. ท้องเสียหรือคลื่นไส้ c. ความดันโลหิตต่ำ d. เลือดออกง่าย

หมายเหตุ: สำหรับคำถามแบบเลือกตอบ ผู้วิจัยสั่งให้ ChatGPT เลือกคำตอบที่เหมาะสมที่สุดพร้อมอธิบายเหตุผลประกอบเพื่อให้สามารถประเมินคุณภาพของคำตอบโดยใช้เกณฑ์เดียวกับคำถามปลายเปิด

คำถามมาตรฐานสำหรับแต่ละด้านของข้อมูลยาได้แสดงไว้ในตารางที่ 1

คำถามสำหรับรายการอื่นทั้งหมดได้รับการพัฒนาตามแม่แบบคำสั่งมาตรฐานเดียวกัน (template-based approach) โดยคงโครงสร้างคำถาม รูปแบบภาษา และระดับรายละเอียดของบริบทให้เหมือนกัน และปรับเฉพาะชื่อยาและข้อมูลทางคลินิกทั่วไปที่เฉพาะเจาะจงของ

ยาแต่ละรายการ เพื่อให้เกิดความสม่ำเสมอของรูปแบบการตั้งคำถามระหว่างรายการยาที่ศึกษา

ความตรงเชิงเนื้อหาของชุดคำถามมาตรฐานและตัวเลือกคำตอบ ได้รับการประเมินโดยผู้เชี่ยวชาญจำนวน 3 คนในด้านความเหมาะสม ความชัดเจน และความครอบคลุมของเนื้อหา พบว่า ดัชนีความตรงเชิงเนื้อหา รายข้อ (item-level content validity index: I-CVI) ของทุกข้อมี

ค่าไม่น้อยกว่า 0.78 และดัชนีความตรงเชิงเนื้อหาในระดับชุดคำถาม (scale-level content validity index/average: S-CVI) เท่ากับ 0.90 ซึ่งสะท้อนว่าชุดคำถามและตัวเลือกคำตอบมีความตรงเชิงเนื้อหาอยู่ในเกณฑ์ดี

ก่อนการเก็บข้อมูลจริง ผู้วิจัยศึกษานำร่องโดยใช้รายการยาจำนวน 5 รายการที่ไม่รวมอยู่ในการศึกษา เพื่อนำสอบถามใน ChatGPT เพื่อประเมินความชัดเจนของคำถาม ความเป็นไปได้ของกระบวนการเก็บข้อมูล และความเหมาะสมของรูปแบบการบันทึกข้อมูล ทั้งนี้ พบว่าคำถาม ขั้นตอนการเก็บข้อมูล และรูปแบบการบันทึกข้อมูลมีความเหมาะสมสำหรับการนำไปใช้ในการศึกษาจริง

แหล่งอ้างอิงสำหรับการสร้างคำถาม

ชุดคำถามมาตรฐาน ตัวเลือกคำตอบสำหรับคำถามแบบเลือกตอบ และเกณฑ์การประเมินคุณภาพคำตอบ ได้รับการพัฒนาจากแหล่งอ้างอิงมาตรฐานด้านเภสัชกรรมที่กำหนดไว้ล่วงหน้า ได้แก่ 1) เอกสารกำกับยาสำหรับบุคลากรทางการแพทย์ (summary of product characteristics: SmPC) และเอกสารกำกับยาสำหรับผู้ป่วย (patient information leaflet: PIL) 2) ฐานข้อมูลสารสนเทศยา ได้แก่ UpToDate Lexidrug (formerly Lexicomp) (14) และ/หรือ Micromedex (DRUGDEX® System) (15) และ 3) Drug Information Handbook (16)

การสอบถามคำตอบจาก ChatGPT

การสอบถามปัญหาประดิษฐ์ด้วยคำถามดำเนินการระหว่างวันที่ 29 ตุลาคม ถึง 26 พฤศจิกายน พ.ศ. 2568 โดยใช้ เวอร์ชัน GPT-5 ผ่านบัญชีผู้ใช้ฟรีบนเว็บไซต์ของ OpenAI เพื่อให้เหมือนกับการใช้งานของผู้ใช้ทั่วไป การสอบถามดำเนินการภายใต้บัญชีผู้ใช้ ChatGPT บัญชีเดียวกันตลอดระยะเวลาการศึกษา และใช้งานผ่านเว็บเบราว์เซอร์เดียวกันภายใต้สภาพแวดล้อมการใช้งานที่สม่ำเสมอ เพื่อลดความแปรปรวนที่อาจเกิดจากความแตกต่างของบัญชีผู้ใช้หรือแพลตฟอร์มที่ใช้เข้าถึงระบบ ในแต่ละรอบการเก็บข้อมูล ผู้วิจัยเริ่มบทสนทนาใหม่ทุกครั้ง และไม่มีการใช้ประวัติการสนทนาจากรอบก่อนหน้า เพื่อลดอิทธิพลของบริบทจากการสนทนาครั้งก่อนหน้า

ผู้วิจัยหลักเป็นผู้ป้อนคำถามทั้งหมดโดยใช้ prompt template เดียวกันในทุกคำถาม เพื่อควบคุมความสม่ำเสมอของการเก็บข้อมูล คำถามแต่ละข้อถูกส่งผ่านการสนทนาใหม่ที่เป็นอิสระจากกัน โดยไม่มีการให้ข้อมูลเพิ่มเติม คำสั่งติดตาม หรือข้อมูลย้อนกลับภายหลังการ

ตอบ เพื่อให้คำตอบที่ได้รับเกิดจากข้อความคำถามเพียงชุดเดียวและปราศจากอิทธิพลจากบริบทของการสนทนาครั้งก่อนหน้า การสอบถามคำตอบจาก ChatGPT ไม่มีการกำหนดแหล่งสืบค้นหรือเอกสารอ้างอิงเฉพาะให้แก่ ChatGPT เพื่อจำลองลักษณะการใช้งานจริงของประชาชนทั่วไป ซึ่งมักสอบถามโดยไม่ได้รับแหล่งข้อมูลอ้างอิง

ตัวอย่าง prompt template ที่เป็นคำถามปลายเปิดคือ "ยาเมตฟอร์มิน (Metformin) มีข้อบ่งใช้ในการรักษาโรคหรือบรรเทาอาการอะไร กรุณาตอบเพียงหนึ่งรอบ โดยไม่อาศัยบริบทจากการสนทนาครั้งก่อนหน้า และแสดง Session ID, Response ID และวันที่เวลาที่ตอบ (รหัสประจำการบันทึกข้อมูลของผู้วิจัยส่งให้ ChatGPT สร้างขึ้น)" ส่วนตัวอย่าง prompt template ที่เป็นคำถามปลายปิดคือ "ยาเมตฟอร์มิน (Metformin) มีข้อบ่งใช้ในการรักษาโรคหรือบรรเทาอาการในข้อใด a. ความดันโลหิตสูง b. เบาหวานชนิดที่ 2 c. ภาวะไขมันในเลือดสูง d. ภาวะกรดยูริกในเลือดสูง กรุณาตอบโดยไม่มีการอาศัยบริบทจากการสนทนาครั้งก่อนหน้า พร้อมอธิบายเหตุผล และแสดง Session ID, Response ID และวันที่เวลาที่ตอบ (รหัสประจำการบันทึกข้อมูลของผู้วิจัยส่งให้ ChatGPT สร้างขึ้น)"

สำหรับยาแต่ละรายการ มีการสอบถามข้อมูลยา 4 ด้าน ได้แก่ ข้อบ่งใช้ ขนาดยา วิธีใช้ และอาการไม่พึงประสงค์ โดยถามทั้งในรูปแบบคำถามปลายเปิดและคำถามแบบเลือกตอบ การสอบถามทำทั้งสิ้นใน 4 ช่วงเวลา คือ วันที่ 0, 7, 14 และ 28 เพื่อประเมินเสถียรภาพของคำตอบในระยะสั้นและระยะกลาง เนื่องจากผลลัพธ์ของแบบจำลองอาจเปลี่ยนแปลงตามเวลา การสอบถามในวันที่ 0 เป็นการถามซ้ำภายในรอบสนทนาเดียวกัน จำนวน 3 ครั้ง ส่วนการสอบถามในวันที่ 7, 14 และ 28 ทำเพียงครั้งเดียว การสอบถามซ้ำทั้งสองแบบทำให้สามารถประเมินความสามารถในการทำซ้ำของผลลัพธ์ (repeatability) และความสม่ำเสมอของคำตอบ ที่มีการกล่าวถึงในการศึกษาครั้งก่อนหน้า (17,18)

ดังนั้น ในการวิจัยมีการถามทั้งหมด 48 ครั้งต่อรายการยา นั่นคือ ในวันที่ 0 มีการสอบถาม 24 ครั้ง (4 ด้านของคำถาม x 2 ประเภทคำถาม (ปลายเปิดและปลายปิด) x 3 ครั้งภายในรอบสนทนา) ส่วนในวันที่ 7, 14 และ 28 มีการสอบถามรวมทั้งหมด 24 ครั้ง (4 ด้านของคำถาม x 2 ประเภทคำถาม (ปลายเปิดและปลายปิด) x 1 ครั้งภายในรอบสนทนา x 3 เวลา) ส่งผลให้มีคำตอบจาก ChatGPT รวมทั้งสิ้น 960 คำตอบสำหรับยาทั้งหมด 20 รายการ

เกณฑ์การประเมินคำตอบที่ได้จาก ChatGPT

เกณฑ์การประเมินคุณภาพของคำตอบจาก ChatGPT พัฒนาจากกรอบแนวคิดคุณภาพข้อมูลของ Wang และ Strong (13) ร่วมกับเกณฑ์การประเมินเกี่ยวกับข้อมูลสุขภาพและยา โดยครอบคลุม 6 มิติ ดังนี้ 1) ความถูกต้องหรือความสอดคล้องของคำตอบที่ได้กับแหล่งอ้างอิงมาตรฐานที่ใช้ในการสร้างคำถามที่กล่าวมาแล้ว หากข้อมูลระหว่างแหล่งอ้างอิงไม่สอดคล้องกัน ผู้ประเมินพิจารณาตามลำดับความน่าเชื่อถือที่กำหนดไว้ล่วงหน้า คือ ให้ SmPC เป็นแหล่งอ้างอิงหลัก และใช้ PIL ประกอบการพิจารณาในประเด็นที่เกี่ยวข้องกับข้อมูลสำหรับผู้ไ้ยา รองลงมาคือฐานข้อมูล UpToDate และ/หรือ Micromedex ตามด้วย Drug Information Handbook 2) ความเหมาะสมหรือความเกี่ยวข้องและความสอดคล้องของคำตอบที่ได้กับคำถามที่ถาม 3) ความครบถ้วนหรือความครอบคลุมของประเด็นสำคัญที่ควรปรากฏในคำตอบ 4) ความสม่ำเสมอหรือความคงที่ของคำตอบเมื่อมีการสอบถามซ้ำ 5) ความชัดเจนของภาษาหรือความเข้าใจง่าย ความเป็นระเบียบและความคลุมเครือของภาษา และ 6) ความปลอดภัยหรือการไม่มีข้อมูลที่อาจก่อให้เกิดอันตรายหรือการใช้ยาอย่างไม่เหมาะสม

การประเมินคำตอบใช้มาตรวัดแบบลิเคิร์ต 5 ระดับ (1–5 คะแนน) โดยคะแนน 1 สะท้อนการไม่เป็นไปตามเกณฑ์อย่างชัดเจน ขณะที่คะแนน 5 สะท้อนการเป็นไปตามเกณฑ์อย่างครบถ้วน ส่วนคะแนนระดับกลางสะท้อนระดับคุณภาพที่อยู่ระหว่างสองขั้วดังกล่าว ผู้อ่านสามารถเข้าถึงรายละเอียดของเกณฑ์การประเมินและแนวทางการให้คะแนนจากแฟ้มที่อยู่ใน URL ซึ่งแสดงในภาคผนวก

แม้ว่าคำถามปลายเปิดและปลายปิดจะมีรูปแบบการตอบแตกต่างกัน แต่คำตอบทั้งหมดได้รับการประเมินด้วยเกณฑ์คุณภาพเดียวกันโดยเปรียบเทียบกับแหล่งอ้างอิงมาตรฐานที่กำหนดไว้ล่วงหน้า สำหรับคำถามแบบเลือกตอบ การประเมินครอบคลุมทั้งความถูกต้องของตัวเลือกและคุณภาพของเหตุผลประกอบคำตอบ ส่งผลให้คะแนนจากคำถามทั้งสองรูปแบบสามารถเปรียบเทียบกันได้ภายใต้กรอบการประเมินเดียวกัน

กระบวนการประเมิน

การประเมินคุณภาพของคำตอบดำเนินการโดยผู้ประเมินหลัก 2 คน ซึ่งเป็นเภสัชกรผู้เชี่ยวชาญด้านศูนย์ข้อมูลยาและบริหารคลังเวชภัณฑ์ ก่อนเริ่มการประเมิน ผู้

ประเมินทั้งสองคนได้รับคู่มือการให้คะแนน และมีการประชุมร่วมกันเพื่อทำความเข้าใจนิยามเชิงปฏิบัติการและแนวทางการให้คะแนนในแต่ละมิติเพื่อลดความคลาดเคลื่อนที่อาจเกิดจากความแตกต่างในการตีความเกณฑ์การให้คะแนน ในกรณีที่ผู้ประเมินทั้งสองให้คะแนนไม่สอดคล้องกัน ผู้ประเมินคนที่สามซึ่งเป็นเภสัชกรผู้เชี่ยวชาญด้านเภสัชกรรมคลินิกทำหน้าที่เป็นผู้ตัดสิน โดยพิจารณาจากแหล่งอ้างอิงมาตรฐานที่กำหนดไว้ล่วงหน้า แล้วใช้ความเห็นของผู้ประเมินท่านที่สามในการวิเคราะห์ข้อมูล

ผู้ประเมินทั้งสองประเมินคุณภาพของคำตอบจาก ChatGPT อย่างอิสระจากกัน คะแนนก่อนกระบวนการปรับความเห็นร่วม (pre-adjudication scores) ถูกนำมาใช้ประเมินความสอดคล้องระหว่างผู้ประเมินด้วย weighted Cohen's kappa และ intraclass correlation coefficient (ICC; two-way random-effects, absolute agreement) การศึกษาถือว่าผู้ประเมินมีความเห็นสอดคล้องกันหาก 1) ทั้งสองให้คะแนนอยู่ในช่วง 1–2 เหมือนกัน แต่ไม่จำเป็นต้องได้ระดับที่ตรงกัน หรือ 2) ทั้งสองให้คะแนนอยู่ในช่วง 4–5 เหมือนกัน แต่ไม่จำเป็นต้องได้ระดับที่ตรงกัน หรือ 3) ทั้งสองให้คะแนน 3 ตรงกันทั้งคู่ การแปลผลค่า kappa และ ICC อ้างอิงตามเกณฑ์ของ Landis และ Koch (19) และ Koo และ Li (20)

การวิเคราะห์ข้อมูล

ตัวแปรตาม คือ การผ่านเกณฑ์คุณภาพของคำตอบ (ผ่าน/ไม่ผ่าน) การศึกษานี้กำหนดให้คำตอบที่มีคะแนนประเมินเฉลี่ยรวมทั้ง 6 มิติอย่างน้อย 4 (จากคะแนนเต็ม 5) คือ คำตอบที่ไม่ผ่านเกณฑ์คุณภาพ การวิเคราะห์ปัจจัยที่สัมพันธ์กับโอกาสที่จะผ่านเกณฑ์คุณภาพของคำตอบใช้สถิติ generalized estimating equations (GEE) สำหรับข้อมูลแบบทวิภาค โดยกำหนดรายการยาแต่ละชนิดเป็นกลุ่ม (cluster) เนื่องจากมีการสังเกตข้อมูลหลายครั้งภายในรายการยาเดียวกัน การวิเคราะห์ข้อมูลใช้แบบจำลอง binomial family with logit link และ exchangeable correlation structure เพื่อคำนึงถึงความสัมพันธ์ระหว่างข้อมูลที่ได้จากการสังเกตซ้ำภายในกลุ่มเดียวกัน ผลการวิเคราะห์รายงานในรูปของ odds ratio (OR) พร้อมช่วงความเชื่อมั่นร้อยละ 95 โดยกำหนดระดับนัยสำคัญทางสถิติที่ $P < 0.05$ การวิเคราะห์ข้อมูลทั้งหมดใช้โปรแกรม Stata Statistical Software (StataCorp LP, College Station, TX, USA)

ตารางที่ 2. คุณภาพของคำตอบจาก ChatGPT เมื่อสอบถามเกี่ยวกับข้อมูลการใช้ยาในบัญชียาหลักแห่งชาติจำแนกตาม 6 มิติที่ประเมิน (N = 960)¹

ระดับคุณภาพ ²	จำนวนคำตอบ (ร้อยละ)					
	ความถูกต้อง	ความเหมาะสม	ความครบถ้วน	ความสม่ำเสมอ	ความชัดเจนของภาษา	ความปลอดภัย
1	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (0.1)
2	18 (1.9)	19 (2.0)	24 (2.5)	24 (2.5)	24 (2.5)	21 (2.2)
3	47 (4.9)	57 (5.9)	49 (5.1)	32 (3.3)	32 (3.3)	40 (4.2)
4	238 (24.8)	404 (42.1)	623 (64.9)	661 (68.9)	543 (56.6)	401 (41.8)
5	657 (68.4)	480 (50.0)	264 (27.5)	243 (25.3)	361 (37.6)	497 (51.8)
ค่าเฉลี่ย±SD	4.60 ± 0.67	4.40 ± 0.69	4.17 ± 0.63	4.17 ± 0.60	4.29 ± 0.65	4.43 ± 0.69
มัธยฐาน (IQR) ³	5 (4–5)	4.5 (4–5)	4 (4–5)	4 (4–5)	4 (4–5)	5 (4–5)

1: ข้อมูลในตารางนี้คำนวณจากคะแนนสุดท้ายภายหลังการปรับความเห็นร่วมในกรณีที่มีผู้ประเมินทั้งสองเห็นต่างกัน

2: ระดับคุณภาพมี 5 ระดับ คือ 1 (คุณภาพต่ำที่สุดหรือมีข้อผิดพลาดสำคัญ) 2 (คุณภาพต่ำ) 3 (คุณภาพปานกลาง) 4 (คุณภาพดี) และ 5 (คุณภาพดีมากหรือมีความครบถ้วนเหมาะสมที่สุด)

3: IQR: interquartile range หรือพิสัยระหว่างควอไทล์

ผลการวิจัย

คุณภาพของคำตอบจาก ChatGPT

ตารางที่ 2 แสดงผลการประเมินคำตอบจาก ChatGPT โดยผู้เชี่ยวชาญ พบว่า คำตอบมีคะแนนคุณภาพเฉลี่ยมากกว่า 4.0 (จากคะแนนเต็ม 5) ในทุกมิติของการประเมิน ด้านที่มีคะแนนเฉลี่ยสูงสุดคือความถูกต้อง (4.60 ± 0.67) รองลงมาคือความปลอดภัย (4.43 ± 0.69) และความเหมาะสม (4.40 ± 0.69) ขณะที่ความครบถ้วน (4.17 ± 0.63) และความสม่ำเสมอ (4.17 ± 0.60) มีคะแนนเฉลี่ยต่ำกว่ามิติอื่นเล็กน้อย ส่วนด้านความเข้าใจง่ายของภาษามีค่าเฉลี่ย 4.29 ± 0.65

เมื่อพิจารณาการกระจายของระดับคะแนนพบว่า คำตอบส่วนใหญ่ได้รับคะแนนระดับ 4 (คุณภาพดี) และ 5 (คุณภาพดีมาก) ในทุกมิติ โดยเฉพาะด้านความถูกต้องและความปลอดภัยซึ่งมีสัดส่วนคะแนนระดับ 5 สูงที่สุด คิดเป็นร้อยละ 68.4 และ 51.8 ตามลำดับ สัดส่วนคำตอบที่ได้คะแนน ≥ 4 คือ ร้อยละ 92.1–94.2 ในทุกด้านของการประเมิน ด้านความสม่ำเสมอและความชัดเจนของภาษามีสัดส่วนผ่านเกณฑ์สูงสุด (ร้อยละ 94.2) รองลงมาคือความปลอดภัย (ร้อยละ 93.5) และความถูกต้อง (ร้อยละ 93.2) ขณะที่ความเหมาะสมมีสัดส่วนผ่านเกณฑ์ต่ำที่สุด (ร้อยละ 92.1) สะท้อนว่าความแตกต่างของสัดส่วนการผ่านเกณฑ์

ระหว่างมิติมีไม่มาก ค่ามัธยฐานของคะแนนคุณภาพอยู่ที่ 4 หรือ 5 คะแนนในทุกมิติ และมีพิสัยระหว่างควอไทล์อยู่ในช่วง 4–5 คะแนน คำตอบที่ได้รับคะแนนต่ำ (1–2 คะแนน) ในสัดส่วนเพียงเล็กน้อย (ร้อยละ 1.9–2.5)

ความสอดคล้องระหว่างผู้ประเมิน

ตารางที่ 3 แสดงความสอดคล้องของผลการประเมินจากผู้ประเมิน 2 ท่านก่อนกระบวนการปรับความเห็นที่แตกต่าง โดยปรับระดับผลการประเมินจาก 5 ระดับเป็น 3 ระดับ พบว่า ค่า ICC ของการประเมินในมิติต่าง ๆ อยู่ในช่วง 0.08–0.28 ซึ่งเมื่อแปลผลตามเกณฑ์ของ Koo และ Li (20) พบว่าจัดอยู่ในระดับต่ำทั้งหมด (ICC น้อยกว่า 0.50) โดยมีค่าความถูกต้องมีค่าความสอดคล้องสูงสุด (ICC = 0.28, 95% CI 0.06–0.54) รองลงมาคือ ความเหมาะสม (ICC = 0.22, 95% CI 0.05–0.46) และ ความปลอดภัย (ICC = 0.17, 95% CI 0.05–0.38) ขณะที่มิติความสม่ำเสมอมีค่าความสอดคล้องต่ำที่สุด (ICC = 0.08, 95% CI 0.02–0.19) นอกจากนี้ ค่าร้อยละความสอดคล้องระหว่างผู้ประเมินทั้งสองอยู่ในช่วงร้อยละ 31.4–81.1 โดยมีมิติความถูกต้องมีค่าร้อยละความสอดคล้องสูงสุด (ร้อยละ 81.1) และมิติ ความสม่ำเสมอมีค่าต่ำที่สุด (ร้อยละ 31.4) (ตารางที่ 3)

ตารางที่ 3. ความสอดคล้องของผลการประเมินระหว่างผู้ประเมิน 2 ท่านก่อนกระบวนการปรับความเห็นที่แตกต่าง (N = 960)

มิติการประเมิน	ร้อยละความสอดคล้อง	Kw ¹	ICC(2,1) (95% CI) ²	การแปลผลค่า ICC ³
ความถูกต้อง	81.1	0.28	0.28 (0.06–0.54)	ต่ำ
ความเหมาะสม	73.5	0.22	0.22 (0.05–0.46)	ต่ำ
ความครบถ้วน	48.1	0.15	0.15 (0.04–0.32)	ต่ำ
ความสม่ำเสมอ	31.4	0.08	0.08 (0.02–0.19)	ต่ำ
ความชัดเจนของภาษา	44.3	0.12	0.12 (0.03–0.28)	ต่ำ
ความปลอดภัย	59.1	0.17	0.17 (0.05–0.38)	ต่ำ

1: Kw คือ Weighted Cohen's Kappa (quadratic weights) ค่าสถิติต่าง ๆ ในตารางคำนวณโดยปรับผลการประเมินจาก 5 ระดับ เป็น 3 ระดับ คือ คุณภาพของคำตอบในระดับต่ำ ปานกลาง และสูง

2: ICC(2,1) คือ two-way random-effects, absolute-agreement, single-measure intraclass correlation coefficient

3: การแปลผลค่า ICC อ้างอิงเกณฑ์ของ Koo and Li (20): ต่ำ (< 0.50), ปานกลาง (0.50–0.75), ดี (0.75–0.90) และ ดีเยี่ยม (> 0.90)

ปัจจัยที่สัมพันธ์กับการผ่านเกณฑ์คุณภาพ

จากตารางที่ 4 สัดส่วนของคำตอบที่ผ่านเกณฑ์คุณภาพในแต่ละหมวดคำถามอยู่ระหว่างร้อยละ 84.2–96.7 หมวดอาการไม่พึงประสงค์มีสัดส่วนของคำตอบที่ผ่านเกณฑ์สูงสุด (ร้อยละ 96.7) และหมวดข้อบ่งใช้มีสัดส่วนของคำตอบที่ผ่านเกณฑ์ต่ำสุด (ร้อยละ 84.2) พบว่า การวิเคราะห์ด้วย GEE ในภาพรวมพบว่า ประเภทของคำถามที่แตกต่างกัน (ข้อบ่งใช้ ขนาดยา วิธีใช้ยา หรืออาการไม่พึงประสงค์) ส่งผลต่อโอกาสในการผ่านเกณฑ์คุณภาพของคำตอบอย่างมีนัยสำคัญ (Wald $\chi^2 = 7.99$, df = 3, P=0.046) แต่เมื่อใช้คำถามประเภทข้อบ่งใช้เป็นกลุ่มอ้างอิง พบว่า คำถามประเภทขนาดยา วิธีใช้ยา หรืออาการไม่พึงประสงค์ มีคำตอบซึ่งมีโอกาสการผ่านเกณฑ์ไม่ต่างจากกลุ่มอ้างอิงอย่างมีนัยสำคัญทางสถิติ (P>0.05)

จากตารางที่ 5 สัดส่วนคำตอบที่ผ่านเกณฑ์คุณภาพในคำถามเกี่ยวกับยาแผนปัจจุบันและยาสมุนไพร เท่ากับร้อยละ 91.3 และ 88.5 ตามลำดับ ผลการวิเคราะห์ด้วย GEE ไม่พบความสัมพันธ์อย่างมีนัยสำคัญทางสถิติระหว่างประเภทคำถามกับโอกาสการผ่านเกณฑ์คุณภาพของคำตอบ (Wald $\chi^2 = 0.45$, df = 1, P=0.504) โดยยาสมุนไพร มีค่า OR เท่ากับ 0.741 (95% CI 0.308–1.784, P = 0.504) เมื่อเปรียบเทียบกับยาแผนปัจจุบัน

จากตารางที่ 6 สัดส่วนคำตอบที่ผ่านเกณฑ์คุณภาพของคำถามปลายเปิดและคำถามแบบเลือกตอบ เท่ากับร้อยละ 92.1 และ 87.7 ตามลำดับ ผลการวิเคราะห์ด้วย GEE ไม่พบความสัมพันธ์อย่างมีนัยสำคัญทางสถิติระหว่างประเภทคำถามกับโอกาสการผ่านเกณฑ์คุณภาพของคำตอบ (Wald $\chi^2 = 3.41$, df = 1, P=0.065) โดย

ตารางที่ 4. ผลการวิเคราะห์ด้วย GEE¹ เพื่อหาความสัมพันธ์ระหว่างการผ่านเกณฑ์คุณภาพของคำตอบและหมวดคำถาม

หมวดคำถาม	จำนวนคำตอบ (ร้อยละ)		β coefficient	SE	OR	95% CI	P
	ผ่านเกณฑ์ ²	ไม่ผ่านเกณฑ์					
ข้อบ่งใช้	214 (89.2)	26 (10.8)	กลุ่มอ้างอิง	-	1.00	-	-
ขนาดยา	202 (84.2)	38 (15.8)	-0.437	0.488	0.64	0.24–1.68	0.370
วิธีใช้ยา	215 (89.6)	25 (10.4)	0.044	0.628	1.04	0.30–3.57	0.944
อาการไม่พึงประสงค์	232 (96.7)	8 (3.3)	1.259	0.788	3.52	0.75–16.50	0.110

1: วิเคราะห์ด้วย GEE โดยใช้ binomial family with logit link และ exchangeable correlation structure เพื่อคำนึงถึงความสัมพันธ์ของข้อมูลภายในรายการยาเดียวกัน ผลการวิเคราะห์พบ Wald $\chi^2 = 7.99$, df = 3, P=0.046

2: ผ่านเกณฑ์คุณภาพ หมายถึง คำถามที่มีค่าเฉลี่ยของคะแนนประเมินทั้ง 6 มิติ ≥ 4

ตารางที่ 5. ผลการวิเคราะห์ด้วย GEE¹ เพื่อหาความสัมพันธ์ระหว่างการผ่านเกณฑ์คุณภาพของคำตอบและประเภทของยาในคำถาม

ประเภทของยา	จำนวนคำตอบ (ร้อยละ)		β coefficient	SE	OR	95% CI	P
	ผ่านเกณฑ์ ²	ไม่ผ่านเกณฑ์					
ยาแผนปัจจุบัน	438 (91.3)	42 (9.7)	กลุ่มอ้างอิง	-	1.00	-	-
ยาสมุนไพร	425 (88.5)	55 (11.5)	-0.300	0.448	0.741	0.31–1.78	0.504

1: วิเคราะห์ด้วย GEE โดยใช้ binomial family with logit link และ exchangeable correlation structure เพื่อคำนึงถึงความสัมพันธ์ของข้อมูลภายในรายการยาเดียวกัน ผลการวิเคราะห์พบ Wald $\chi^2 = 0.45$, df = 1, P=0.504

2: ผ่านเกณฑ์คุณภาพ หมายถึง คำถามที่มีค่าเฉลี่ยของคะแนนประเมินทั้ง 6 มิติ ≥ 4

คำถามปลายเปิดมีค่า OR เท่ากับ 1.630 (95% CI 0.970–2.738, P = 0.065)

สัดส่วนของคำตอบจาก ChatGPT ที่ผ่านเกณฑ์จำแนกตามรายการยาแสดงอยู่ในจากตารางที่ 7 โดยสัดส่วนดังกล่าวอยู่ในช่วงร้อยละ 60.4–100.0

จากตารางที่ 8 สัดส่วนของคำตอบที่ผ่านเกณฑ์คุณภาพมีค่าใกล้เคียงกันในทุกช่วงเวลาทดสอบ (ร้อยละ 88.75–90.42) ภายใต้เงื่อนไขการสอบถามและรูปแบบ prompt ที่กำหนดไว้ในการศึกษา ผลการวิเคราะห์เพิ่มเติมด้วย GEE ให้ผลสอดคล้องกับข้อมูลเชิงพรรณนาดังกล่าว (Wald $\chi^2 = 3.73$, df = 5, P=0.589) ค่า adjusted OR ในแต่ละช่วงเวลามีค่าใกล้เคียงกับ 1 และช่วงความเชื่อมั่นร้อยละ 95 คร่อมค่า 1 ทั้งหมด ซึ่งบ่งชี้ว่าไม่พบหลักฐานเชิงสถิติที่สนับสนุนความแตกต่างของโอกาสการผ่านเกณฑ์คุณภาพระหว่างช่วงเวลาประเมินภายใต้ข้อมูลการศึกษา

การอภิปรายผล

การประเมินคุณภาพคำตอบของ ChatGPT

คำตอบจาก ChatGPT มีคะแนนคุณภาพเฉลี่ยมากกว่า 4.0 จากคะแนนเต็ม 5 ในทุกมิติการประเมิน

โดยเฉพาะด้านความถูกต้อง ความเหมาะสม และความปลอดภัย สะท้อนว่า ChatGPT สามารถให้ข้อมูลยาพื้นฐานได้ในระดับที่สอดคล้องกับเกณฑ์คุณภาพที่กำหนด อย่างไรก็ตาม การประเมินครั้งนี้มุ่งเน้นคุณภาพของเนื้อหาคำตอบใน 6 มิติ โดยไม่พิจารณาการพิจารณาของผู้เชี่ยวชาญด้านเภสัชกรรม แต่ไม่ได้ประเมินผลลัพธ์ทางคลินิกโดยตรง ดังนั้น ผลการศึกษาจึงยังไม่สามารถยืนยันผลกระทบต่อผลลัพธ์สุขภาพหรือความปลอดภัยของผู้ป่วยในสถานการณ์จริงได้

ผลการประเมินของผู้ประเมินหลัก 2 คนมีความสอดคล้องในระดับต่ำ โดยค่า ICC และ Weighted Cohen's kappa อยู่ระหว่าง 0.08–0.28 และมีคำตอบจำนวนมากที่ต้องเข้าสู่กระบวนการปรับความเห็นร่วม (ร้อยละ 63.3–92.2) ความไม่สอดคล้องดังกล่าวพบเด่นชัดในมิติความครบถ้วน ความสม่ำเสมอ และความชัดเจนของภาษา ซึ่งเป็นมิติที่การให้คะแนนขึ้นอยู่กับความเห็นส่วนบุคคลของผู้ประเมินมากกว่าความถูกต้องของข้อเท็จจริงในคำตอบเพียงอย่างเดียว ส่งผลให้ผู้ประเมินมีความเห็นแตกต่างกัน แม้จะใช้เกณฑ์การประเมินเดียวกัน การศึกษายุทธศาสตร์คะแนนการประเมินจาก 5 ระดับเป็น 3 ระดับ (คุณภาพปานกลาง และสูง) เพื่อลดความแตกต่างของการให้คะแนน

ตารางที่ 6. ผลการวิเคราะห์ด้วย GEE¹ เพื่อหาความสัมพันธ์ระหว่างการผ่านเกณฑ์คุณภาพของคำตอบและรูปแบบของคำถาม

รูปแบบของคำถาม	จำนวนคำตอบ (ร้อยละ)		β coefficient	SE	OR	95% CI	P
	ผ่านเกณฑ์ ²	ไม่ผ่านเกณฑ์					
คำถามแบบเลือกตอบ	421 (87.7)	59 (12.3)	กลุ่มอ้างอิง	-	1.00	-	-
คำถามปลายเปิด	444 (92.1)	38 (7.9)	0.489	0.265	1.63	0.97–2.74	0.065

1: วิเคราะห์ด้วย GEE โดยใช้ binomial family with logit link และ exchangeable correlation structure เพื่อคำนึงถึงความสัมพันธ์ของข้อมูลภายในรายการยาเดียวกัน ผลการวิเคราะห์พบ Wald $\chi^2 = 3.41$, df = 1, P=0.065

2: ผ่านเกณฑ์คุณภาพ หมายถึง คำถามที่มีค่าเฉลี่ยของคะแนนประเมินทั้ง 6 มิติ ≥ 4

ตารางที่ 7. จำนวนและร้อยละของคำตอบที่ผ่านเกณฑ์คุณภาพจำแนกตามรายการยาแต่ละชนิด

ชื่อยา	ผ่านเกณฑ์คุณภาพ ¹	ไม่ผ่านเกณฑ์คุณภาพ	ร้อยละที่ผ่านเกณฑ์	95% CI
enalapril	48	0	100.0	92.6–100.0
มะระขี้นก	48	0	100.0	92.6–100.0
ฟ้าทะลายโจร	47	1	97.9	88.9–99.9
ขมิ้นชัน	47	1	97.9	88.9–99.9
simvastatin	46	2	95.8	85.7–99.5
paracetamol	46	2	95.8	85.7–99.5
amlodipine	46	2	95.8	85.7–99.5
มะขามแขก	46	2	95.8	85.7–99.5
ขิง	46	2	95.8	85.7–99.5
omeprazole	45	3	93.8	82.8–98.7
clopidogrel	45	3	93.8	82.8–98.7
ยาอมมะแว้ง	44	4	91.7	80.0–97.7
amoxicillin	43	5	89.6	77.3–96.5
เลวาล์ยเปรี๊ยะ	42	6	87.5	74.8–95.3
salbutamol	41	7	85.4	72.2–93.9
metformin	40	8	83.3	69.8–92.5
ยาสหัสธารา	40	8	83.3	69.8–92.5
losartan	38	10	79.2	65.0–89.5
ยาตรีผลา	36	12	75.0	60.4–86.4
ยาเพชรสังฆาต	29	19	60.4	45.3–74.2

1: แต่ละรายการยามีจำนวนครั้งของการถาม ChatGPT 48 ครั้ง จึงมี 48 คำตอบที่ต้องประเมิน การผ่านเกณฑ์คุณภาพ หมายถึง ค่าเฉลี่ยของคะแนนประเมินทั้ง 6 มิติ ≥ 4

ระหว่างผู้ประเมิน แต่การคำนวณยังได้ ICC ในระดับต่ำ ดังกล่าวมาแล้ว สะท้อนว่าความแตกต่างระหว่างผู้ประเมิน ไม่ได้เกิดจากการใช้ระดับคะแนนที่ละเอียดเพียงอย่างเดียว

แต่ยังเกิดจากความแตกต่างในการตีความเกณฑ์การประเมินและการใช้ดุลยพินิจในแต่ละมิติ ซึ่งเป็นลักษณะที่พบได้ในการประเมินในเชิงคุณภาพ โดยเฉพาะด้านความ

ตารางที่ 8. ผลการวิเคราะห์ GEE เพื่อหาความสัมพันธ์ระหว่างช่วงเวลาในการประเมินกับโอกาสผ่านเกณฑ์คุณภาพของคำตอบ

ช่วงเวลาการประเมิน	จำนวนคำตอบ (ร้อยละ)		adjusted OR	95% CI	P
	ผ่านเกณฑ์ ²	ไม่ผ่านเกณฑ์			
วันที่ 0 (กลุ่มอ้างอิง)	434 (90.42)	46 (9.58)	1.00	-	-
วันที่ 7	143 (89.38)	17 (10.62)	1.43	0.27–7.68	0.675
วันที่ 14	142 (88.75)	18 (11.25)	1.14	0.48–2.71	0.770
วันที่ 28	144 (90.00)	16 (10.00)	0.94	0.47–1.87	0.866

1: วิเคราะห์ด้วย GEE โดยใช้ binomial family with logit link และ exchangeable correlation structure เพื่อคำนึงถึงความสัมพันธ์ของข้อมูลภายในรายการยาเดียวกัน ผลการวิเคราะห์พบ Wald $\chi^2 = 3.73$, df = 5, P=0.589

2: ผ่านเกณฑ์คุณภาพ หมายถึง คำถามที่มีค่าเฉลี่ยของคะแนนประเมินทั้ง 6 มิติ ≥ 4 ; การสอบถาม ChatGPT ในวันที่ 0 ทำสามครั้งต่อเนื่องกัน แต่การถามในวันที่ 7, 14 และ 28 ทำเพียงครั้งเดียว

เหมาะสม ความครบถ้วน และความชัดเจนของภาษา ดังนั้นการใช้ผู้ประเมินคนที่สามเพื่อชี้ขาดกรณีที่มีความเห็นไม่สอดคล้องกันจึงมีบทบาทสำคัญในการศึกษา

ผลการศึกษาสอดคล้องกับรายงานของ Kung และคณะ (21) ซึ่งพบว่า ChatGPT มีศักยภาพในการตอบคำถามทางการแพทย์และสามารถในการให้เหตุผลประกอบคำตอบในบริบททางการแพทย์ การศึกษาก่อนหน้ารายงานว่า ChatGPT อาจมีศักยภาพในการสนับสนุนงานด้านข้อมูลและงานเภสัชกรรมบางด้าน อย่างไรก็ตาม ยังคงมีข้อกังวลเกี่ยวกับความถูกต้อง ความครบถ้วน และความน่าเชื่อถือของข้อมูลที่ระบบสร้างขึ้น (22,23) ขณะที่สุพรรณิการ์ พรวิวัฒน์ และคณะ (10) รายงานว่า ChatGPT, Google Gemini และ Microsoft Copilot สามารถตอบคำถามข้อมูลยาภาษาไทยได้ในระดับที่ยอมรับได้ แต่ยังมีข้อจำกัดที่ต้องอาศัยการตีความทางคลินิกจากผู้เชี่ยวชาญ

แม้ว่าคะแนนคุณภาพโดยรวมจะอยู่ในระดับสูง แต่ด้านความครบถ้วนและความสม่ำเสมอมีคะแนนเฉลี่ยต่ำกว่ามิติอื่น ซึ่งสะท้อนว่า ChatGPT อาจยังมีข้อจำกัดในการนำเสนอข้อมูลอย่างครอบคลุมและรักษาความคงที่ของคำตอบในบางสถานการณ์ ข้อค้นพบดังกล่าวสอดคล้องกับ Walker และคณะ (9) ที่รายงานว่าการ ChatGPT สามารถให้ข้อมูลทางการแพทย์ได้ในระดับที่เทียบได้กับแหล่งข้อมูลสุขภาพออนไลน์ทั่วไป แต่ยังมีข้อจำกัดด้านคุณภาพและความสอดคล้องกับแนวทางเวชปฏิบัติ นอกจากนี้ Morath และคณะ (22) รายงานว่า ในการตอบคำถามเกี่ยวกับยา ChatGPT อาจให้ข้อมูลที่ไมครบถ้วนหรือให้ข้อมูลที่มีความคลาดเคลื่อนได้ ขณะที่ Rooson และคณะ (23) พบว่า ChatGPT 4.0 มีศักยภาพในการสนับสนุนงานเภสัชกรรมเชิงคลินิกในบางกรณี ดังนั้นความถูกต้องของข้อมูลจึงขึ้นกับบริบทและลักษณะการใช้งาน นอกจากนี้ Huang และคณะ (24) พบว่า ChatGPT มีประสิทธิภาพแตกต่างกันตามประเภทของงาน โดยทำได้ดีในงานให้คำปรึกษาด้านยา แต่ยังมีข้อจำกัดในงานที่ซับซ้อน เช่น การทบทวนใบสั่งยาและการประเมินอาการไม่พึงประสงค์

ปัจจัยที่สัมพันธ์กับการผ่านเกณฑ์คุณภาพ

ผลการวิเคราะห์ด้วย GEE ไม่พบความสัมพันธ์อย่างมีนัยสำคัญทางสถิติระหว่างหมวดคำถาม ประเภทยาประเภทคำถาม หรือช่วงเวลาการประเมิน กับโอกาสการ

ผ่านเกณฑ์คุณภาพของคำตอบ แม้ว่าบางกลุ่มจะมีสัดส่วนการผ่านเกณฑ์แตกต่างกันในการวิเคราะห์เชิงพรรณนา แต่ยังไม่พบหลักฐานเพียงพอที่จะสนับสนุนความสัมพันธ์ดังกล่าว ทั้งนี้ คะแนนที่กระจุกตัวอยู่ในระดับสูงและอัตราการผ่านเกณฑ์ที่สูงในเกือบทุกกลุ่มอาจก่อให้เกิดปรากฏการณ์เพดานคะแนน (ceiling effect) ซึ่งลดความไวในการตรวจจับความแตกต่างระหว่างกลุ่ม

แม้ว่ายาสมุนไพรจะมีแนวโน้มผ่านเกณฑ์คุณภาพต่ำกว่ายาแผนปัจจุบัน (OR = 0.741, 95% CI = 0.308–1.784, P = 0.504) แต่ไม่พบความแตกต่างอย่างมีนัยสำคัญทางสถิติ แนวโน้มดังกล่าวอาจเกี่ยวข้องกับความแตกต่างของแหล่งข้อมูลที่ใช้ในการพัฒนาองค์ความรู้ โดยข้อมูลยาแผนปัจจุบันมักมีหลักฐานเชิงประจักษ์และเอกสารมาตรฐานรองรับอย่างชัดเจน ขณะที่ข้อมูลยาสมุนไพรบางรายการยังมีข้อจำกัดด้านหลักฐานทางวิทยาศาสตร์ ความสม่ำเสมอขององค์ประกอบ หรือมาตรฐานข้อมูลในบางบริบท (26) นอกจากนี้ กลุ่มสมุนไพรที่ใช้ในการศึกษาคครอบคลุมทั้งสมุนไพรเดี่ยวและยาตำรับ ซึ่งมีความซับซ้อนและบริบทของข้อมูลแตกต่างกัน อย่างไรก็ตาม การศึกษานี้ไม่ได้ออกแบบมาเพื่อเปรียบเทียบระหว่างประเภทของสมุนไพรโดยตรง จึงยังไม่สามารถสรุปสาเหตุของความแตกต่างดังกล่าวได้ ทั้งนี้ ผลการศึกษาของ สุพัทธา ลิมสุวรรณโชติ และคณะ (26) รายงานถึงข้อจำกัดด้านความน่าเชื่อถือและความครบถ้วนของคำตอบเกี่ยวกับสมุนไพรไทยที่สร้างโดย ChatGPT ดังนั้น จึงมีความจำเป็นที่ต้องใช้ความระมัดระวังเมื่อนำคำตอบเกี่ยวกับสมุนไพรไปใช้ในทางปฏิบัติ

สำหรับรูปแบบคำถาม แม้ไม่พบความสัมพันธ์อย่างมีนัยสำคัญทางสถิติ แต่ค่าประมาณของความสัมพันธ์บ่งชี้ว่า คำถามปลายเปิดอาจมีโอกาผ่านเกณฑ์คุณภาพสูงกว่าคำถามแบบเลือกตอบ (OR = 1.630, 95% CI = 0.970–2.738, P = 0.065) ทั้งนี้ คำถามปลายเปิดที่ใช้ในการศึกษาส่วนใหญ่เป็นคำถามข้อมูลยาพื้นฐานที่สามารถตอบได้โดยตรงจากองค์ความรู้ด้านยา ขณะที่คำถามแบบเลือกตอบอาจต้องอาศัยการพิจารณาและเปรียบเทียบตัวเลือกหลายทางเลือกก่อนให้คำตอบ อย่างไรก็ตาม ผลดังกล่าวควรตีความด้วยความระมัดระวัง และไม่อาจสรุปได้อย่างชัดเจนว่ารูปแบบคำถามมีผลต่อคุณภาพของคำตอบ ทั้งนี้ LLMs มีความเสี่ยงต่อการสร้างข้อมูลที่คลาดเคลื่อน โดยเฉพาะในสถานการณ์ที่มีความซับซ้อนหรือบริบทที่กำกวม (22,27)

เมื่อวิเคราะห์ตามช่วงเวลา พบว่าคุณภาพของคำตอบมีความคงที่ตลอดระยะเวลา 28 วัน โดยไม่พบความแตกต่างอย่างมีนัยสำคัญทางสถิติ แม้ว่ารายละเอียดของคำตอบอาจแตกต่างกันในแต่ละครั้ง แต่ความแตกต่างดังกล่าวไม่ได้ส่งผลต่อคุณภาพของคำตอบโดยรวม สะท้อนว่า ChatGPT มีความเสถียรของผลลัพธ์ในระดับหนึ่งภายใต้เงื่อนไขการทดสอบเดียวกัน

การประยุกต์ใช้ผลการวิจัย

ผลการศึกษาสະท้อนว่า ChatGPT มีศักยภาพในการสนับสนุนการเข้าถึงข้อมูลยาพื้นฐานของประชาชน โดยเฉพาะข้อมูลเกี่ยวกับข้อบ่งใช้ ขนาดยา วิธีใช้ และอาการไม่พึงประสงค์ ซึ่งอาจช่วยส่งเสริมความรอบรู้ด้านสุขภาพและเพิ่มความสะดวกในการเข้าถึงข้อมูลยาในบริบทของระบบบริการสุขภาพไทย อย่างไรก็ตาม ผลการศึกษานี้สะท้อนคุณภาพของคำตอบภายใต้เงื่อนไขการวิจัยที่กำหนดไว้เท่านั้น จึงยังไม่สามารถสรุปถึงประสิทธิผลหรือความปลอดภัยของการนำไปใช้ในสถานการณ์จริงได้โดยตรง

แม้ ChatGPT จะสามารถให้ข้อมูลยาได้ในระดับที่ดีโดยรวม แต่ยังพบคำตอบบางส่วนที่ไม่ผ่านเกณฑ์การประเมินในบางมิติของคุณภาพ ดังนั้นผลลัพธ์ของการศึกษานี้จึงสนับสนุนให้พิจารณาใช้ระบบปัญญาประดิษฐ์เป็นเครื่องมือสนับสนุนการเข้าถึงและสืบค้นข้อมูลมากกว่าการใช้เป็นแหล่งข้อมูลเพียงแหล่งเดียวหรือใช้แทนการตัดสินใจของบุคลากรทางการแพทย์ โดยเฉพาะในสถานการณ์ที่ซับซ้อน เช่น การปรับขนาดยา การพิจารณาข้อห้ามใช้ การประเมินอันตรกิริยาระหว่างยา หรือการดูแลผู้ป่วยที่มีโรคร่วมหลายชนิด (28)

ข้อจำกัดของการศึกษา

การศึกษานี้มีข้อจำกัดหลายประการ ประการแรก มีการประเมินยาเพียง 20 รายการจากบัญชียาหลักแห่งชาติ และบัญชียาจากสมุนไพรแห่งชาติ และมุ่งเน้นคำถามข้อมูลยาพื้นฐาน จึงอาจไม่สามารถสะท้อนศักยภาพของ ChatGPT ในสถานการณ์ทางคลินิกที่มีความซับซ้อนสูงได้อย่างครบถ้วน ประการที่สอง แม้การศึกษานี้จะใช้ผู้ประเมินที่มีความเชี่ยวชาญและมีกระบวนการปรับความเห็นร่วมเพื่อกำหนดคะแนนสุดท้าย แต่ความสอดคล้องระหว่างผู้ประเมินก่อนการปรับความเห็นอยู่ในระดับต่ำในทุกมิติของการประเมิน สะท้อนว่ามีคุณภาพบางด้าน เช่น ความ

ครบถ้วน ความสม่ำเสมอ และความชัดเจนของภาษา อาจมีความแปรปรวนในการให้คะแนนระหว่างผู้ประเมินได้ แม้จะใช้เกณฑ์การประเมินเดียวกัน ซึ่งอาจส่งผลต่อความสามารถในการทำซ้ำของผลการประเมินในบริบทอื่น

ประการที่สาม แม้ GEE จะเหมาะสมสำหรับข้อมูลแบบการวัดซ้ำ แต่ผลลัพธ์ที่ได้เป็นการประมาณผลกระทบในระดับประชากร และยังไม่ได้นำถึงปัจจัยบางประการที่อาจมีผลต่อคุณภาพคำตอบ เช่น ความซับซ้อนของคำถาม รูปแบบ prompt และปฏิสัมพันธ์ระหว่างตัวแปรต่าง ๆ นอกจากนี้ จำนวนรายการยาที่ศึกษาอาจยังไม่เพียงพอสำหรับการตรวจพบความแตกต่างขนาดเล็กในบางการเปรียบเทียบ ส่งผลให้ค่าประมาณบางส่วนมีช่วงความเชื่อมั่นค่อนข้างกว้าง นอกจากนี้ ผลการศึกษายังสะท้อนคุณภาพของ ChatGPT ภายใต้เวอร์ชัน ช่วงเวลาการเก็บข้อมูล และเงื่อนไขการใช้งานที่ศึกษาเท่านั้น เนื่องจากระบบอาจมีการปรับปรุงหรือเปลี่ยนแปลงได้อย่างต่อเนื่อง ดังนั้นควรมีการศึกษาติดตามในเวอร์ชันและบริบทการใช้งานที่แตกต่างกัน รวมทั้งเปรียบเทียบกับระบบปัญญาประดิษฐ์แบบรูปร่างอื่น เพื่อประเมินความคงที่และความสามารถในการทำซ้ำของผลลัพธ์

ข้อเสนอแนะ

การศึกษาในอนาคตควรกำหนดสมมติฐานหลักและผลลัพธ์หลักไว้ล่วงหน้า พร้อมเลือกวิธีวิเคราะห์ทางสถิติที่เหมาะสมกับข้อมูลแบบการวัดซ้ำ หรือข้อมูลแบบจัดกลุ่ม รวมถึงพัฒนาแนวทางการประเมินเพื่อเพิ่มความสอดคล้องระหว่างผู้ประเมินและความน่าเชื่อถือของผลการศึกษา นอกจากนี้ ควรขยายขอบเขตการศึกษาให้ครอบคลุมรายการยาในบัญชียาหลักแห่งชาติมากขึ้น รวมถึงยาความเสี่ยงสูงและยาที่มีความซับซ้อนทางคลินิก เพื่อเพิ่มความสามารถในการสรุปผล นอกจากนี้ ควรพิจารณาศึกษาความแตกต่างระหว่างสมุนไพรเดี่ยวและยาสมุนไพรตำรับ เพื่อประเมินผลของลักษณะข้อมูลและความซับซ้อนของยาแต่ละประเภทต่อคุณภาพของคำตอบจากระบบปัญญาประดิษฐ์

การศึกษาในอนาคตควรเปรียบเทียบประสิทธิภาพของ ChatGPT กับ LLMs อื่น หรือระบบปัญญาประดิษฐ์เฉพาะทางด้านการแพทย์และข้อมูลยา เพื่อประเมินจุดแข็งข้อจำกัด และความเหมาะสมในการประยุกต์ใช้ในบริบทของระบบสุขภาพไทย การศึกษาเพิ่มเติมในกลุ่มประชากรและสถานการณ์ทางคลินิกที่มีความซับซ้อนมีความจำเป็น เช่น

เด็ก ผู้สูงอายุ หญิงตั้งครรภ์ ผู้ป่วยที่มีความบกพร่องของตับหรือไต รวมถึงการประเมินอันตรายกิริยาระหว่างยา การปรับขนาดยา และการจัดการอาการไม่พึงประสงค์จากยา เพื่อประเมินศักยภาพและข้อจำกัดของระบบด้านความปลอดภัยเชิงคลินิกได้อย่างครอบคลุมยิ่งขึ้น

การศึกษาที่น่าสนใจ คือ ผลของรูปแบบการตั้งคำถาม (prompt engineering) การใช้ฐานข้อมูลเฉพาะทาง และการผสานแหล่งข้อมูลอ้างอิงมาตรฐานต่อคุณภาพของคำตอบ รวมทั้งประเมินประสิทธิภาพของระบบภายใต้รูปแบบคำถามและสถานการณ์ที่ใกล้เคียงการใช้งานจริง เพื่อสะท้อนการใช้งานของประชาชนและบุคลากรสุขภาพได้ดียิ่งขึ้น นอกจากนี้ ควรมีการศึกษาติดตามเพื่อประเมินความเสถียร ความสม่ำเสมอของผลลัพธ์ และความสามารถในการทำซ้ำ (reproducibility) ภายใต้ช่วงเวลา เวอร์ชัน และสภาพแวดล้อมการใช้งานที่แตกต่างกัน โดยเฉพาะสำหรับระบบปัญญาประดิษฐ์ที่มีการพัฒนาและปรับปรุงอย่างต่อเนื่อง

สรุป

ChatGPT มีศักยภาพในการให้ข้อมูลยาภายใต้เงื่อนไขและกรอบการประเมินที่กำหนด โดยคุณภาพของคำตอบโดยรวมอยู่ในระดับสูงในทั้ง 6 มิติที่ประเมิน แม้ว่ามิติด้านความครบถ้วนและความสม่ำเสมอจะมีคะแนนต่ำกว่ามิติอื่นบ้าง นอกจากนี้ ไม่พบความแตกต่างอย่างมีนัยสำคัญทางสถิติของคุณภาพคำตอบระหว่างช่วงเวลาการประเมิน ประเภทยา ประเภทข้อมูลยา หรือรูปแบบคำถามภายใต้แบบจำลองที่ใช้ ผลการศึกษาสนับสนุนศักยภาพของ ChatGPT ในการเป็นเครื่องมือสนับสนุนการเข้าถึงข้อมูลยา และข้อมูลสุขภาพที่เกี่ยวข้องกับการใช้ยา ทั้งสำหรับประชาชนและบุคลากรด้านสุขภาพ โดยเฉพาะการให้ข้อมูลทั่วไปเกี่ยวกับการใช้ยา อย่างไรก็ตาม ควรใช้เป็นเครื่องมือสนับสนุนมากกว่าการใช้แทนการตัดสินใจของบุคลากรด้านสาธารณสุข โดยเฉพาะในสถานการณ์ที่มีความซับซ้อนหรือมีความเสี่ยงสูง

กิตติกรรมประกาศ

คณะผู้วิจัยขอขอบพระคุณ ญญ.มณฑนา คันทะเรศร์ หัวหน้ากลุ่มงานเภสัชกรรม โรงพยาบาลอุดรดิตถ์ ผู้เปิดโอกาสในการทำวิจัยครั้งนี้ ญญ.เทียมใจ ตั้งเจริญไพศาล ผู้ให้คำปรึกษาตลอดโครงการ ญญ.นิภาพันท์ มานักข้อง

ญญ.ปรียานุช รอดทอง ญญ.พีรดา เกตุวีระพงศ์ ผู้เชี่ยวชาญในการประเมินคำถามและคำตอบ และผู้ช่วยศาสตราจารย์ ดร. อศนี วันชัย ผู้อำนวยการวิทยาลัยพยาบาลบรมราชชนนี พุทธชินราช ผู้ให้คำปรึกษากระบวนการวิจัยแก่ผู้วิจัย

การใช้ปัญญาประดิษฐ์ในบทความ

ปัญญาประดิษฐ์ที่ใช้ในการจัดทำบทความ คือ ChatGPT เวอร์ชัน GPT-5.5 โดยใช้ในช่วงเดือน พฤษภาคม-มิถุนายน พ.ศ. 2569 เพื่อสนับสนุนการตรวจทานภาษา การวินิจฉัยและแก้ไขข้อผิดพลาดทางภาษา การวิเคราะห์ความครอบคลุมและความสอดคล้องของเนื้อหาต้นฉบับ รวมถึงให้ข้อเสนอแนะในการปรับปรุงการนำเสนอผลการศึกษา การอภิปรายผล และองค์ประกอบของบทความในเชิงวิชาการ

ผู้เขียนได้ตรวจสอบ แก้ไข และทวนสอบความถูกต้องของเนื้อหาทั้งหมดด้วยตนเอง รวมทั้งตรวจสอบความถูกต้องของข้อมูล การวิเคราะห์ทางสถิติ การตีความผลการศึกษา และเอกสารอ้างอิง โดยผู้เขียนขอรับผิดชอบต่อความถูกต้อง ความครบถ้วน ความคิดริเริ่ม และความน่าเชื่อถือของเนื้อหาในบทความฉบับสมบูรณ์แต่เพียงผู้เดียว

เอกสารอ้างอิง

1. National Statistical Office. The 2021 health and welfare survey [online]. 2022 [cited May 26, 2026]. Available from: www.nso.go.th/nsoweb/storage/survey_detail/2023/20230505170315_44270.pdf
2. Aekplakorn W. Thailand's national health examination survey VI 2019–2020 [online]. 2021 [cited May 26, 2026]. Available from: www.hiso.or.th/hiso/picture/reportHealth/report/sreport6/sreport6_full.pdf
3. National Drug System Development Committee. National list of essential medicines B.E. 2565 [online]. 2022 [cited May 27, 2026]. Available from: ndi.fda.moph.go.th/uploads/file_news/20220808893215585.PDF
4. National Drug System Development Committee. National list of herbal medicinal products B.E. 2566 [online]. 2023 [cited May 27, 2026]. Available from:

- herbal.fda.moph.go.th/media.php?id=865141728649289728&name=Main medicine-66.pdf
5. Sumpradit N, Chongtrakul P, Anuwong K, Pumtong S, Kongsomboon K, Butdeemee P, et al. Antibiotics smart use: a workable model for promoting the rational use of medicines in Thailand. *Bull World Health Organ.* 2012; 90: 905-13.
6. Yun HS, Bickmore T. Online health information-seeking in the era of large language models: cross-sectional web-based survey study. *J Med Internet Res.* 2025; 27: e68560.
7. Fridman I, Johnson S, Elston Lafata J. Health information and misinformation: a framework to guide research and practice. *JMIR Med Educ.* 2023; 9: e38687.
8. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel).* 2023; 11: 887.
9. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. *J Med Internet Res.* 2023; 25: e47479.
10. Pornwattanakavee S, Leelakanok N, Todsarot T, Guinto GAT, Takun R, Sumativit A, et al. Effectiveness of ChatGPT, Google Gemini, and Microsoft Copilot in answering Thai drug information queries: cross-sectional study. *JMIR AI* 2025;4: e79751.
11. Aydin S, Karabacak M, Vlachos V, Margetis K. Navigating the potential and pitfalls of large language models in patient-centered medication guidance and self-decision support. *Front Med (Lausanne).* 2025; 12: 1527864.
12. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. *J Biomed Inform* 2024; 151: 104620.
13. Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. *J Manag Inf Syst.* 1996; 12: 5-33.
14. Wolters Kluwer. UpToDate Lexidrug (formerly Lexicomp) [online]. 2026 [cited Jun 13, 2026] . Available from: online.lexi.com
15. Merative. DRUGDEX® System [online]. 2026 [cited Jun 13, 2026] . Available from: www.micromedex solutions.com
16. Lexicomp. Adult drug information handbook 2025-2026. 33rd ed. Hudson (OH): Wolters Kluwer Clinical Drug Information, Inc.; 2024.
17. Meyer A, Schömig E, Streichert T. ChatGPT and reference intervals: a comparative analysis of repeatability in GPT-3.5 Turbo, GPT-4, and GPT-4o. *Front Artif Intell.* 2025; 8: 1681979.
18. Funk PF, Hoch CC, Knoedler S, Knoedler L, Cotofana S, Sofo G, et al. ChatGPT's response consistency: a study on repeated queries of medical examination questions. *Eur J Investig Health Psychol Educ.* 2024; 14: 657-68.
19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977; 33 :159-74.
20. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* 2016; 15: 155-63.
21. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health.* 2023; 2: e0000198.
22. Morath B, Chiriac U, Jaskowski E, Deiß C, Nürnberg H, Hörth K, et al. Performance and risks of ChatGPT used in drug information: an exploratory real-world analysis. *Eur J Hosp Pharm.* 2024; 31: 491-7.
23. Roosan D, Padua P, Khan R, Khan H, Verzosa C, Wu Y. Effectiveness of ChatGPT in clinical pharmacy and the role of artificial intelligence in medication

- therapy management. J Am Pharm Assoc (2003). 2024; 64: 422-28.
24. Huang X, Estau D, Liu X, Yu Y, Qin J, Li Z. Evaluating the performance of ChatGPT in clinical pharmacy: a comparative study of ChatGPT and clinical pharmacists. Br J Clin Pharmacol. 2024; 90: 232-8.
25. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaire. J Clin Epidemiol. 2007; 60: 34-42.
26. Limsuwanchote S, Sakunphueak A, Boonrit N, Hopkins AM, Ruanglertboon W. Analysing credibility of information on Thai herbs generated by ChatGPT from pharmacists' perspectives. Thai Journal of Pharmacy Practice 2024; 16: 1257-76.
27. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst [online]. 2023 [cited May 26, 2026]. doi.org/10.1145/3703155
28. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019; 17: 195.

ภาคผนวก

ผู้อ่านสามารถดาวน์โหลดเอกสารในภาคผนวกจากแฟ้ม Appendix_paper.pdf ได้ที่ <https://osf.io/e2ahd>

ชุดข้อมูลเสริม S1 (Supplementary Dataset S1)

ชุดข้อมูลเสริมนี้จัดทำขึ้นเพื่อเพิ่มความโปร่งใสของกระบวนการศึกษาและทำให้สามารถทำซ้ำงานวิจัยได้ ชุดข้อมูลสามารถดาวน์โหลดไฟล์ได้ที่ <https://osf.io/e2ahd> ชื่อไฟล์ SupplementaryDatasetS1.xlsx ชุดข้อมูลประกอบด้วย

- 1) รายการยา 20 รายการ
- 2) คำถามมาตรฐานทั้งหมด โดยคำถามถูกออกแบบให้อยู่ในรูปแบบเดียวกันสำหรับยาทุกชนิด ทั้งคำถามปลายเปิดและคำถามแบบเลือกตอบ เพื่อควบคุม

อคติจากเครื่องมือวัดและเพิ่มความเที่ยงตรงในการเปรียบเทียบผลลัพธ์ระหว่างประเภทยาและรายตัวยา โดยตัวเลือกของคำถามแบบเลือกตอบได้รับการพัฒนาให้มีความสมเหตุสมผลและสอดคล้องกับบริบททางคลินิก เช่น ยาแผนปัจจุบัน มีคำถามปลายเปิด คือ ยาเมตฟอร์มิน (metformin) มีข้อบ่งใช้ในการรักษาโรคหรือบรรเทาอาการอะไร ส่วนคำถามแบบเลือกตอบมีตัวอย่าง เช่น ข้อใดเป็นข้อบ่งใช้ของยาเมตฟอร์มิน (metformin) A. ความดันโลหิตสูง B. เบาหวานชนิดที่ 2 C. ภาวะไขมันในเลือดสูง D. ภาวะกรดยูริกในเลือดสูง

ยาสมุนไพร มีคำถามปลายเปิด คือ ยาฟ้าทะลายโจร ใช้สำหรับรักษาอาการอะไร ส่วน คำถามแบบเลือกตอบ เช่น ข้อใดเป็นข้อบ่งใช้ของยาฟ้าทะลายโจร A. อาการปวดหัวจากไมเกรน B. ไข้ เจ็บคอ บรรเทาอาการหวัด C. ท้องผูกเรื้อรัง D. ความดันโลหิตสูง

3) prompt template สำหรับ prompt template ในคำถามปลายเปิด คือ คำถามมาตรฐาน + กรุณาตอบเพียงหนึ่งรอบ โดยไม่อาศัยบริบทจากการสนทนาก่อนหน้า และแสดง Session ID, Response ID และวันเวลาที่ตอบ (รหัสประจำการบันทึกข้อมูลที่ผู้วิจัยสร้างขึ้น) ส่วน prompt template ในคำถามแบบเลือกตอบ คือ คำถามมาตรฐาน + กรุณาตอบโดยไม่มีการอาศัยบริบทจากการสนทนาก่อนหน้า พร้อมอธิบายเหตุผล และแสดง Session ID, Response ID และวันเวลาที่ตอบ (รหัสประจำการบันทึกข้อมูลที่ผู้วิจัยสั่งให้ ChatGPT สร้างขึ้น)

การศึกษาใช้รูปแบบ prompt เดียวกันในทุกการทดลองเพื่อควบคุมความแปรปรวนของผลลัพธ์ ทั้งนี้ การศึกษานี้ดำเนินการผ่าน ChatGPT เวอร์ชัน GPT-5 ผ่านบัญชีผู้ใช้ฟรีบนเว็บไซต์ของ OpenAI ซึ่งผู้วิจัยไม่สามารถกำหนดค่าพารามิเตอร์ของโมเดล เช่น temperature หรือ top-p ได้โดยตรง

ผู้วิจัยไม่ได้ใช้ role prompting หรือการกำหนดบทบาทเฉพาะให้แก่ระบบ เช่น การระบุให้ ChatGPT ทำหน้าที่เป็นแพทย์หรือเภสัชกร ทั้งนี้ เพื่อสะท้อนลักษณะการใช้งานทั่วไปของผู้ใช้ในสถานการณ์จริง และลดอิทธิพลจากการชี้นำบริบทของคำถาม

- 4) เกณฑ์การประเมิน 6 มิติ พร้อมคะแนนที่ให้โดยผู้เชี่ยวชาญ 2 คน
- 5) คำตอบจาก ChatGPT

ตัวอย่างคำตอบ (คำถามปลายเปิด)

ยาเมตฟอร์มิน (Metformin)

ข้อบ่งใช้ (Indications):

ยาเมตฟอร์มินเป็นยาในกลุ่ม biguanide ที่ใช้สำหรับรักษาโรคเบาหวานชนิดที่ 2 (Type 2 Diabetes Mellitus) เพื่อควบคุมระดับน้ำตาลในเลือด โดยเฉพาะในผู้ป่วยที่น้ำหนักเกินหรืออ้วน อาจใช้ร่วมกับอินซูลินหรือยาลดน้ำตาลชนิดอื่น เพื่อช่วยเพิ่มประสิทธิภาพในการควบคุมระดับน้ำตาลในเลือด

ข้อบ่งใช้อื่น (off-label use) บางกรณี เช่น ภาวะถุงน้ำรังไข่หลายใบ (Polycystic Ovary Syndrome: PCOS) ภาวะดื้ออินซูลิน (Insulin resistance)

(รหัสประจำการบันทึกข้อมูลที่ผู้วิจัยสร้างขึ้น)

Session ID: SIM-20251029-1015A

Response ID: META-002

วันที่/เวลา: 29 ตุลาคม 2025 — 10:15 น.

ตัวอย่างคำตอบ (คำถามแบบเลือกตอบ)

คำตอบ: B. เบาหวานชนิดที่ 2

เหตุผล: ยา เมตฟอร์มิน (Metformin) เป็นยาในกลุ่ม Biguanide ใช้ในการรักษาโรคเบาหวานชนิดที่ 2 (Type 2 Diabetes Mellitus) โดยช่วยลดระดับน้ำตาลในเลือดผ่านกลไกลดการสร้างน้ำตาลจากตับ เพิ่มความไวของเนื้อเยื่อต่ออินซูลิน ลดการดูดซึมน้ำตาลจากลำไส้

ไม่ใช้ในการรักษาโรคความดันโลหิตสูง, ไขมันในเลือดสูง หรือภาวะกรดยูริกในเลือดสูงโดยตรง
(รหัสประจำการบันทึกข้อมูลที่ผู้วิจัยสั่งให้ ChatGPT สร้างขึ้น)

Session ID: SIM-20251029-1002A

Response ID: META-003

วันที่/เวลา: 29 ตุลาคม 2025 — 10:02 น.

6) ข้อมูลรอบการถาม: 1) การประเมินซ้ำภายในรอบสนทนาเดียวกัน จำนวน 3 ครั้งในวันที่ 0 และ 2) การประเมินซ้ำระหว่างช่วงเวลาจำนวน 3 ครั้งในวันที่ 7, 14 และ 28 หลังการประเมินครั้งแรก (สอบถาม ChatGPT เพียงครั้งเดียวในวันที่ 7, 14 และ 28)

ชุดข้อมูลเสริม S2 (Supplementary Dataset S2)

ผู้อ่านสามารถสามารถดาวน์โหลดไฟล์ชุดข้อมูลเสริมนี้ได้ ที่ <https://osf.io/e2ahd> ในชื่อไฟล์ SupplementaryDatasetS2.xlsx) ชุดข้อมูลประกอบด้วย ชื่อ ยา กลุ่มยาใหญ่ หมวดคำถาม ประเภทคำถาม ประเภท session และการให้คะแนนทั้ง 6 มิติโดยผู้เชี่ยวชาญ 3 คน (คะแนนจากผู้ประเมินหลัก 2 คน และคะแนนสุดท้ายภายหลังการปรับความเห็นตามการประเมินโดยผู้เชี่ยวชาญคนที่ 3 ในกรณีที่ผู้ประเมินหลักทั้งสองท่านมีความเห็นไม่สอดคล้องกัน