ประสิทธิภาพการวิเคราะห์ผลแปปสเมียร์ของ ChatGPT เปรียบเทียบการกับอ่านผลโดยพยาธิแพทย์ ของผู้มารับบริการตรวจคัดกรองมะเร็งปากมดลูก ที่โรงพยาบาลมะเร็งลพบุรี

โดย
นางสาววาทินี อยู่ประยงค์



ประสิทธิภาพการวิเคราะห์ผลแปปสเมียร์ของ ChatGPT เปรียบเทียบกับการอ่านผลโดยพยาธิแพทย์ ของผู้มารับบริการตรวจคัดกรองมะเร็งปากมดลูกที่โรงพยาบาลมะเร็งลพบุรี

วาทินี อยู่ประยงค์

พุทธลักษณ์ พุกเจริญ

ณิชาภัทร มะโนแจ่ม

ธรรพ์ชนก อิ่มโอษฐ์

กฤษฎา ศรีสุรักษ์

สุดารัตน์ น้อยเจริญ

กลุ่มงานพยาธิวิทยากายวิภาค

โรงพยาบาลมะเร็งลพบุรีกรมการแพทย์กระทรวงสาธารณสุข

พ.ศ. 2569

ชื่อเรื่อง : ประสิทธิภาพการวิเคราะห์ผลแปปสเมียร์ของ ChatGPT เปรียบเทียบกับการอ่านผลโดยพยาธิแพทย์ ของผู้มารับบริการตรวจคัดกรองมะเร็งปากมดลูกที่โรงพยาบาลมะเร็งลพบุรี

ผู้วิจัย : วาทินี อยู่ประยงค์, พุทธลักษณ์ พุกเจริญ, ณิชาภัทร มะโนแจ่ม, ธรรพ์ชนก อิ่มโอษฐ์,กฤษฎา ศรีสุรักษ์, สุดารัตน์ น้อยเจริญ

บทคัดย่อ

การวิจัยเชิงวิเคราะห์แบบย้อนหลังครั้งนี้ มีวัตถุประสงค์เพื่อประเมินประสิทธิภาพการวิเคราะห์ผลแปปสเมียร์ของ ChatGPT โดยเปรียบเทียบกับการอ่านผลโดยพยาธิแพทย์ กลุ่มตัวอย่างคือ ผู้มารับบริการตรวจคัดกรองมะเร็งปากมดลูกที่โรงพยาบาลมะเร็งลพบุรี ที่มีผลการตรวจวินิจฉัยและรายงานผลโดยพยาธิแพทย์เสร็จสิ้นเรียบร้อยแล้ว จำนวน 197 ราย ประกอบด้วยกลุ่มผลผิดปกติ 126 ราย และกลุ่มผลปกติ 71 ราย เครื่องมือที่ใช้ในการวิจัยได้แก่ แบบบันทึกวิจัย ซึ่งประกอบด้วยข้อมูล 2 ส่วน คือ ข้อมูลทั่วไปและข้อมูลผลการตรวจวินิจฉัยแปปสเมียร์ ผ่านการตรวจสอบความตรงเชิงเนื้อหาจากผู้เชี่ยวชาญ เก็บรวบรวมข้อมูลโดยตรวจซ้ำสไลด์และคัดเลือกภาพจากกล้องจุลทรรศน์อย่างน้อย 8 ภาพต่อราย นำเข้าสู่ระบบ ChatGPT เพื่อวิเคราะห์ผล แล้วเปรียบเทียบกับผลการอ่านของพยาธิแพทย์ ระหว่างวันที่ 1 ตุลาคม 2566 ถึง 28 กุมภาพันธ์ 2569 วิเคราะห์ข้อมูลโดยใช้สถิติพรรณนาได้แก่ ร้อยละ ค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐานและสถิติอนุมานได้แก่ การทดสอบไคสแควร์ คำนวณค่าความไว ความจำเพาะ ค่าพยากรณ์ผลบวก ค่าพยากรณ์ผลลบ และความแม่นยำ/อัตราความสอดคล้อง รายงานเป็นค่าร้อยละ และค่าสัมประสิทธิ์แคปปาของโคเฮน (Cohen's kappa)

ผลการวิจัยในระดับการจำแนกแบบ 2 กลุ่ม (ปกติ/ผิดปกติ) พบว่า ChatGPT มีความไวร้อยละ 78.57 ความจำเพาะร้อยละ 74.65 ค่าพยากรณ์ผลบวกร้อยละ 84.62 ค่าพยากรณ์ผลลบร้อยละ 66.25 และความแม่นยำ/อัตราความสอดคล้องร้อยละ 77.16 โดยมีค่าสัมประสิทธิ์แคปปาของโคเฮนเท่ากับ 0.518 ซึ่งจัดอยู่ในระดับความสอดคล้องปานกลาง ส่วนในระดับการจำแนกแบบหลายกลุ่มตามระบบ TBS 2014 พบความสอดคล้องแบบตรงกันทุกหมวดร้อยละ 53.30 และค่าแคปปาลดลงเหลือ 0.370 ซึ่งจัดอยู่ในระดับพอใช้ โดยหมวดที่มีความสอดคล้องสูงสุดคือ HSIL (ร้อยละ 93.3) รองลงมาคือ NILM (ร้อยละ 66.3) ขณะที่พบแนวโน้มวินิจฉัยเกินจริงในหมวด LSIL สำหรับหมวด ASC-US และ SCC ไม่พบความสอดคล้องและไม่มีการจำแนกผลเป็นหมวด ASC-H หรือ Adenocarcinoma ตลอดการศึกษา

ผลการศึกษาสนับสนุนว่า ChatGPT มีศักยภาพในการคัดกรองเบื้องต้นระดับ 2 กลุ่มในระดับที่ยอมรับได้ แต่ยังมีข้อจำกัดสำคัญในการจำแนกรายละเอียดตามระบบ TBS 2014 โดยเฉพาะค่าพยากรณ์ผลลบที่ยังไม่สูงพอ ผู้วิจัยจึงเสนอแนะว่าการนำ ChatGPT มาใช้สนับสนุนงานคัดกรองแปปสเมียร์ควรอยู่ภายใต้การกำกับดูแลของพยาธิแพทย์หรือนักเซลล์วิทยา โดยผลที่จำแนกเป็นปกติควรได้รับการทบทวนซ้ำก่อนสรุปผลขั้นสุดท้าย และไม่ควรใช้ทดแทนการวินิจฉัยของผู้เชี่ยวชาญโดยตรง นอกจากนี้ ควรมีการศึกษาเพิ่มเติมเพื่อพัฒนารูปแบบคำสั่ง (prompt) และให้ข้อมูลเชิงสัณฐานวิทยาที่ช่วยแยกรอยโรคระดับต่ำและระดับสูงให้ชัดเจนขึ้น ตลอดจนเปรียบเทียบรูปแบบข้อมูลนำเข้าที่แตกต่างกัน เช่น การเพิ่มจำนวนภาพตัวแทนต่อรายหรือการใช้ภาพสแกนทั้งสไลด์ เพื่อประเมินว่าข้อจำกัดด้านประสิทธิภาพที่พบเกิดจากความสามารถของโมเดลหรือข้อจำกัดของข้อมูลนำเข้าเป็นสำคัญ

Title : Diagnostic Performance of ChatGPT in Papanicolaou Smear Interpretation Compared with Pathologist Interpretation among Cervical Cancer Screening Participants at Lopburi Cancer Hospital

Researcher : Miss.Watinee Yooprayong**,** Mrs.Phutthalak Pukcharoen**,** Miss.Nichaphat Manochaem**,** Miss.Thanchanok Imaod, Mr. Kitsada Srisuruk, Miss. Sudarat Noyjarean

Abstract

This retrospective analytical study aimed to evaluate the performance of ChatGPT in interpreting Pap smear results compared with interpretations made by pathologists. The sample comprised 197 individuals who underwent cervical cancer screening at Lopburi Cancer Hospital and whose diagnostic results had been finalized and reported by pathologists. Of these, 126 had abnormal results and 71 had normal results. The research instrument was a data-recording form consisting of two sections: general information and Pap smear diagnostic results. Its content validity was reviewed by experts. Data were collected by rescreening the slides and selecting at least eight representative microscopic images per case. The images were submitted to ChatGPT for analysis, and its results were compared with the pathologists' interpretations. Data collection was conducted from October 1, 2023, to February 28, 2026. Data were analyzed using descriptive statistics, including percentage, mean, and standard deviation, and inferential statistics, including the chi-square test. Sensitivity, specificity, positive predictive value, negative predictive value, and accuracy/concordance rate were calculated and reported as percentages, together with Cohen's kappa coefficient.

For binary classification (normal vs. abnormal), ChatGPT achieved a sensitivity of 78.57%, specificity of 74.65%, positive predictive value of 84.62%, negative predictive value of 66.25%, and accuracy/concordance rate of 77.16%. Cohen's kappa coefficient was 0.518, indicating moderate agreement. For multiclass classification according to the 2014 Bethesda System for Reporting Cervical Cytology (TBS 2014), the exact agreement across all diagnostic categories was 53.30%, while Cohen's kappa decreased to 0.370, indicating fair agreement. HSIL showed the highest category-specific agreement (93.3%), followed by NILM (66.3%). ChatGPT demonstrated a tendency to overdiagnose cases as LSIL. No agreement was observed in the ASC-US and SCC categories, and no cases were classified as ASC-H or adenocarcinoma throughout the study.

The findings suggest that ChatGPT has potential as a preliminary tool for binary Pap smear screening, with performance at an acceptable level. However, it has important limitations in detailed classification according to TBS 2014, particularly its relatively low negative predictive value. Therefore, the use of ChatGPT to support Pap smear screening should remain under the supervision of pathologists or cytotechnologists. Results classified as normal should be independently reviewed before a final interpretation is issued, and ChatGPT should not be used as a direct substitute for expert diagnosis. Further studies should focus on refining prompts and incorporating morphological information that more clearly differentiates low-grade from high-grade lesions. Different input formats should also be compared, such as increasing the number of representative images per case or using whole-slide images, to determine whether the observed performance limitations are primarily attributable to the model's capabilities or to constraints in the input data.