

กรอบงานการแปลงและบูรณาการข้อมูลที่ขับเคลื่อนด้วยคุณภาพข้อมูล เพื่อสร้างรายงานความไวต่อยาต้านจุลชีพสะสมแบบทำซ้ำได้ในระบบโรงพยาบาล

Data Quality–Driven ETL Framework for Reproducible Cumulative Antibigram Generation in Hospital Systems

จิตติพัฒน์ สนม¹, ฉันทชนก อร่ามรัตน์¹, กิตติกร หาญตระกูล¹, พาสน์ ปราโมกษ์ชน¹, อลงกต กองมณี¹, และ สมนึก สินธุปวน^{1,*}

Jittiphat Sanom¹, Chanchanok Aramrat¹, Kittikorn Hantrakul¹, Part Pramokchon¹, Alongkot Gongmanee¹, and Somnuek Sinthupuan^{1,*}

¹สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยแม่โจ้ 50290 ประเทศไทย

* ผู้นิพนธ์ประสานงาน: สมนึก สินธุปวน อีเมล: somnuek@mju.ac.th

บทคัดย่อ:

รายงานความไวต่อยาต้านจุลชีพสะสม (cumulative antibiogram) มีความสำคัญต่อการเลือกยาปฏิชีวนะแบบเชิงประจักษ์และการกำกับการใช้ยา แต่ข้อมูลจุลชีววิทยาจากระบบโรงพยาบาลมักมีค่าว่าง ข้อมูลซ้ำ และการตั้งชื่อเชื้อ/ยาไม่เป็นมาตรฐาน ทำให้ผลคำนวณคลาดเคลื่อน โครงการนี้มีวัตถุประสงค์เพื่อ (1) ประเมินและยกระดับคุณภาพข้อมูลจุลชีววิทยา (2) ทำมาตรฐานและบูรณาการข้อมูลเพาะเชื้อจากหลายแหล่งด้วยกระบวนการ ETL และ (3) สร้างรายงาน cumulative antibiogram ตามแนวทาง CLSI M39 เพื่อสนับสนุนการตัดสินใจทางคลินิก งานวิจัยพัฒนา pipeline แบบครบวงจรประกอบด้วยกรอบ DQA (ความครบถ้วน/ความสอดคล้องกับตารางมาตรฐาน) กระบวนการ ETL ที่ทำมาตรฐานชื่อเชื้อ สิ่งส่งตรวจ ชื่อยาเป็นรหัส WHONET และผล S/I/R ด้วย dictionary (>400 รายการ) ร่วมกับ regex และ fuzzy matching ตรวจสอบ intrinsic resistance กำหนด REVIEW สำหรับรายการที่ตีความไม่ได้ (ไม่นำไปรวมตัวหาร) และคัดเลือก first-isolate ภายใน 30 วันเพื่อลดอคติจากผลซ้ำ จากนั้นสร้าง antibiogram ตาม CLSI M39 โดยกำหนดขั้นต่ำ 30 isolates ต่อคู่เชื้อ-ยา-สิ่งส่งตรวจ และส่งออกเป็น Excel แบบกำหนดสีและกรองข้อมูลได้ นอกจากนี้มีการทดลองเชิงสำรวจด้วย Machine Learning ให้ความแม่นยำราว 70–82% แต่กลุ่ม Intermediate มี recall ต่ำจาก class imbalance จึงเสนอใช้ Macro-F1 และการแบ่งข้อมูลแบบ time-/patient-based ในอนาคต

คำสำคัญ : การประเมินคุณภาพข้อมูล; ETL; ความไวต่อยาต้านจุลชีพสะสม; เชื้อดื้อยา; CLSI M39; WHONET; Fuzzy Matching; การคัดข้อมูลซ้ำ; การเรียนรู้ของเครื่อง

Abstract:

Cumulative antibiograms guide empirical therapy and antimicrobial stewardship, but their accuracy is limited by poor-quality microbiology data from hospital information systems. This study aimed to (1) assess and improve microbiology data quality, (2) standardize and integrate heterogeneous culture datasets for reproducible analysis, and (3) generate CLSI M39-compliant cumulative antibiograms for clinical decision support. We developed an end-to-end pipeline comprising a Data Quality Assessment framework for completeness and consistency against standardized mapping tables, an ETL process that standardizes organism, specimen, antibiotic (WHONET codes), and S/I/R results using dictionary mapping (>400 entries) with regex and fuzzy matching, applies intrinsic-resistance

rules, flags unresolvable records as REVIEW (excluded from denominators), and deduplicates isolates using a 30-day first-isolate rule, and cumulative antibiogram generation with a minimum of 30 isolates per organism–antibiotic–specimen stratum, exported as a filterable, color-coded Excel report. An exploratory machine-learning analysis achieved ~70–82% accuracy but showed low recall for the Intermediate class due to imbalance, supporting Macro-F1 and time-/patient-based splits in future work. This pipeline improves the reliability and reproducibility of antibiogram reporting for clinical decision support and AMR surveillance.

Keywords: Objectives; Data Quality Assessment; ETL; Cumulative Antibiogram; Antimicrobial Resistance; CLSI M39; WHONET; Fuzzy Matching; Deduplication; Machine Learning