ReferralAI: Retrospective and Prospective Validation of a Machine Learning Model to Support Specialist Referral Triage for the Logan ENT Clinic.

Bernard Whitfield[1], Wil Wentholt[3], Steve Graham[3], Ahmad Abdel-Hafez[3], Nicola Slee[2]
1. ENT Service Metro South Logan QLD Australia
2. ENT Service QCH Brisbane QLD Australia
3. Clinical & Business Intelligence SAIS Branch QHealth Brisbane QLD Australia

Introduction: The ENT service at Logan now receives 700 adult and paediatric referrals monthly. Triaging consumes 12 hours of valuable clinical time per month.

Aim: To develop and prospectively validate ReferralAI, a machine learning (ML) decision support tool for categorising referrals into Clinical Prioritisation Criteria (CPC) urgency categories for ENT outpatient services.

Methodology: A retrospective cohort study was conducted using more than 17,000 historical ENT referrals. Model evaluation included class-level precision and sensitivity metrics, fairness analysis across demographic and referral source, and explainability assessment. Prospective validation was performed via real-time silent deployment generating predictions for 840 new referrals without influencing clinical triage decisions.

Results: In retrospective validation, the Adult XGBoost model demonstrated precision of 86%, 43%, and 79% and sensitivities of 81%, 45%, and 76% for Categories 1, 2, and 3 respectively. The Paediatric LightGBM model demonstrated precision of 79%, 67%, and 86% and sensitivities of 77%, 64%, and 90% respectively. During prospective silent validation, adult performance across 576 referrals demonstrated precision of 80%, 50%, and 44% and sensitivities of 16%, 52%, and 71% across categories. Paediatric validation across 173 referrals demonstrated precision of 100%, 73%, and 71% and sensitivities of 8%, 76%, and 78%.

Conclusion: Comparing retrospective and prospective validation, in adults, macro precision and macro sensitivity declined. In paediatrics, macro precision improved from 71% to 81% while sensitivity decreased from 75% to 54%, indicating stronger classification confidence. The next phase will evaluate large language model integration into ReferralAI to improve models' performance, and explainability.