Pathformer: Language Models for Automated CPT Billing Code Assignment in Multi-Institutional Pathology Reports

Renner, Elijah (School: Thetford Academy)

Accurate pathology billing is crucial for reimbursement, regulatory compliance, and mitigating the $28.8 billion Medicare improper payment crisis amid a growing shortage of medical coders. CPT codes 88300-88309 denote increasing complexity in examination of surgical specimens, yet reports contain interpretive information subject to ambiguous assignment. Prior automated coding approaches are constrained by limited context windows that truncate key findings, generic vocabularies, and uni-institutional training that generalizes poorly. This study introduces Pathformer, a multi-institutional, long-context transformer-based framework for automated primary CPT code prediction in pathology. Five transformers were pretrained on 59,923 cases from Dartmouth-Hitchcock and fine-tuned on 174,045 cases from Cedars-Sinai, amounting to over 233,000 reports. To date, the two corpora are the largest applied to this task. This pretraining-finetuning strategy was compared against models trained solely on the secondary institution's data and against traditional baselines (Naive Bayes, Random Forest, XGBoost). The top-performing model, a pretrained SciBERT-Longformer, achieved a macro-F1 of 0.8912, surpassing all baselines and prior BERT-based approaches. Pretraining yielded the largest improvements for rare, high-complexity codes (e.g., 88302, 88309) where misbilling risk is greatest. SHAP analyses confirmed that Pathformer predictions are grounded in clinically salient terminology aligned with CPT guidelines. With inference at ~23ms per report, Pathformer demonstrates the accuracy, interpretability, and efficiency necessary for high-volume billing workflows and offers a scalable path toward reducing administrative burden and ensuring proper reimbursement.

Awards Won:

Fourth Award of $600