Pilot Study: Deep Learning Transformer Predicts Surgical Need in Knee Injury Patients from Imaging Reports Alone

Shragvi Balaji, Alex R Flores, Sean Lau, Elizabeth Ledbetter, Thomas Hamre, Vijay Nitturi, Lorenzo D Deveza

INTRODUCTION:
Only 23% of musculoskeletal referrals lead to surgery, yet imaging is routinely obtained early in evaluation. For providers without immediate orthopedic support, especially in safety-net systems, determining which patients are likely to benefit from subspecialty referral remains a challenge. While radiology reports contain rich clinical detail, no current tools leverage this unstructured text to aid triage. We developed OrthoScreen, a transformer-based natural language processing (NLP) model that aims to identify patients likely to require surgical consideration based solely on their initial radiology reports. Our goal was to explore whether free-text imaging data could inform more timely and equitable access to surgical care.

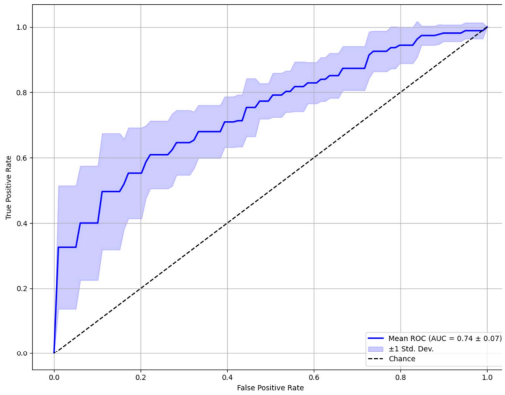
METHODS:
We retrospectively analyzed patients presenting with knee pathology to a single urban safety-net hospital from 2017 to 2020, each with MRI or X-ray and documented follow-up. Text from radiology “Findings” and “Impression” sections were extracted from each initial report. These data were tokenized and used for fine-tuning Bio_ClinicalBERT, to create OrthoScreen, a transformer-based NLP model designed to predict whether surgery was considered within one year. The model was optimized to maximize F1 score and evaluated using stratified 10-fold cross-validation. Performance was measured using AUC, sensitivity, specificity, PPV, NPV, and F1 score.

RESULTS:
A total of 453 patients were included in the final cohort. each with knee imaging (MRI or X-ray) and sufficient follow-up to determine whether surgical consideration occurred within one year. Imaging modality included MRI (n=301) and X-ray (n=152). A total of 272 patients (60.0%) were either recommended for or underwent surgery during the follow-up period. On cross-validation, OrthoScreen achieved averages of AUC: 0.745, sensitivity: 0.933, specificity: 0.352, PPV: 0.689, NPV: 0.837, and F1 score: 0.787.

DISCUSSION AND CONCLUSION:
OrthoScreen demonstrated consistently high sensitivity, suggesting potential utility as a screening or prioritization tool to identify patients who may benefit from orthopedic evaluation. While specificity was limited, this tradeoff reflects a deliberate design choice prioritizing sensitivity in the context of early triage, particularly in resource-constrained or high-volume clinical environments where missed surgical cases may result in delayed care.

This study provides preliminary evidence that transformer-based NLP models can extract clinically relevant signals from unstructured radiology text to support decision-making. Although not intended to replace clinical judgment, OrthoScreen may serve as a scalable adjunct for non-specialist providers navigating musculoskeletal referrals.

Ongoing work aims to enhance model precision by incorporating structured clinical variables, including patient demographics, radiographic features, and diagnosis codes. Given the safety-net setting of the study cohort, OrthoScreen may have particular relevance for improving access and coordination in systems disproportionately affected by delays in specialty care. With further validation and expansion, this tool has the potential to contribute meaningfully to more efficient and equitable musculoskeletal care pathways.


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Mean ROC (AUC = 0.74 ± 0.07)
±1 Std. Dev.
Chance