Using Natural Language Processing on drug indications to predict working sources of infection

Chang Ho YOON[1], Kevin YUAN[1], Qingze GU[1], Henry NP MUNBY[2], A. Sarah WALKER[1], Tingting ZHU[3], David W. EYRE[1,4]

[1]Nuffield Department of Population Health, University of Oxford, UK [2]University Hospitals Bristol & Weston NHS Trust, UK
[3]Institute of Biomedical Engineering, University of Oxford, UK [4]John Radcliffe Hospital, Oxford, UK

Background

Identifying relevant information from large, unstructured clinical notes is challenging even with state-of-the-art natural language processing (NLP) models. Traditional methods to understand working diagnoses at the time of drug prescription include laborious and costly manual extraction, or the use of International Classification of Diseases (ICD) codes that more accurately reflect the ultimate discharge diagnoses. The increasing use of e-prescribing on electronic health platforms, where drug indications are frequently free-text, opens the opportunity to apply NLP to deduce working diagnoses at the time of prescription. Our study specifically focuses on analysing antibiotic prescribing indications to deduce sources of infection.

Methods

We used a source dataset of 940,887 free-text antibiotic indications from 302,568 in-patient episodes at Oxford University Hospitals, UK, and treated with antibiotics. The top 4000 most frequent, unique antibiotic indications (covering 82% of the source dataset) were selected for independent labelling by two clinical researchers (third reviewer resolved any discrepancies). These 4000 samples comprised the *top frequency* dataset, with ten different ground-truth labels of working sources of infection and diagnostic uncertainty. Using 4000 ground truth labels, we classified the free-text infection sources from the original dataset and compared this distribution of sources of infection with that of ICD10 codes extracted from patient episodes. Using this *top frequency* dataset, we trained three different models (n-gram XGBoost, baseline BERT, Bio+Clinical BERT) on the labelled samples and validated their performance on two distinct test sets: (1) *top frequency test set* from the 80/20 train/test split of the *top frequency* dataset; and (2) *source data test set* – 2000 indications randomly sampled from the original source dataset of >940,000 indications.

Results

Of the 302,568 patient episodes, working diagnoses of source of infection based on drug indications were more specific than those based on ICD10 diagnostic codes (45% (135,721/302,568) vs. 71% (213,444/302,568)) [Figure 1]. Bio + Clinical BERT outperforms the other models both in the *top frequency* and *source data* test sets, scoring higher averages with lower spreads per class [Table 1].

Sources of Infection

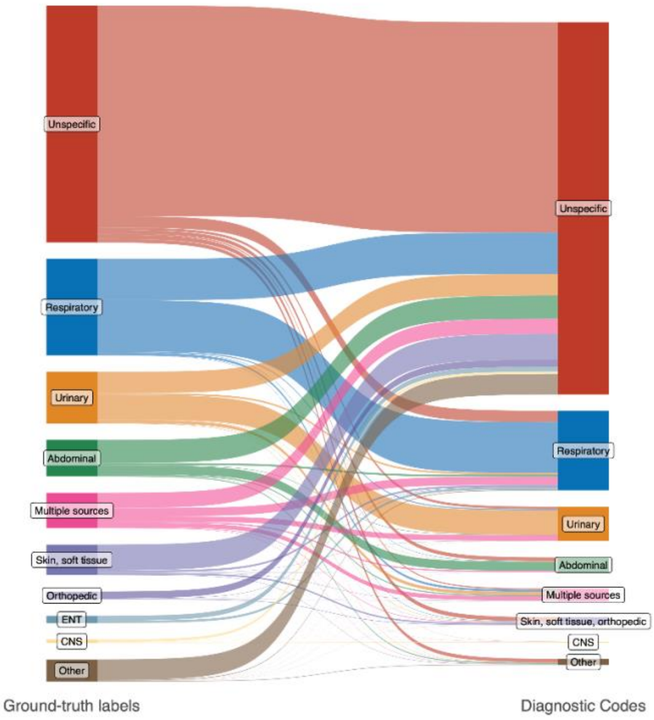

Figure 1*: Comparison of working sources of infection according to clinician-reviewed ground-truth labels vs. ICD10 codes.*

Model Classification Performance

	Top Frequency Test Set *(n= 800)*		
Model	F1	ROC AUC	PR AUC
XGBoost	0.73 (0.29-0.93)	0.91 (0.75-0.96)	0.78 (0.22-0.91)
Baseline BERT	0.89 (0.00-0.98)	0.98 (0.92-0.99)	0.95 (0.54-0.99)
Bio + Clinical BERT	**0.94** **(0.36-0.98)**	**0.99** **(0.96-1.00)**	**0.97** **(0.83-0.99)**
	Source Data Test Set *(n=2000)*		
Model	F1	ROC AUC	PR AUC
XGBoost	0.84 (0.37-0.95)	0.95 (0.77-0.98)	0.86 (0.39-0.97)
Baseline BERT	0.94 (0.57-0.98)	0.98 (0.89-1.00)	0.96 (0.73-0.99)
Bio + Clinical BERT	**0.96** **(0.81-0.98)**	**0.99** **(0.94-1.00)**	**0.97** **(0.84-0.98)**

Table 1: *Comparison of model performances . Metrics are provided as a weighted average across classes, with ranges of classes.*

Conclusion

Fine-tuned NLP models can accurately predict working sources of infection based on unstructured, free-text e-prescriptions, and outperform n-gram + XGBoost. This provides a more real-time and comprehensive distribution of working diagnoses for prescriptions than discharge-related ICD codes, hence offering the possibility of incorporating working diagnoses in, e.g., time-series analyses and time-dynamic AI models. Although significantly less labour-intensive than manual extraction, our case study was limited by the need for ground-truth labels. Further studies will focus on the transferability of locally-trained NLP models to external sites.