Large language models accurately identify people who inject drugs in Australian electronic health records

David Goodman-Meza[1,2], *Marianne Martinello*[2,3], *Jeffrey Masters*[2,4], *Gail V. Matthews*[1,2], *Gregorey J. Dore*[1,2]

[1] St Vincent's Hospital Sydney, Sydney, Australia
[2] Kirby Institute, University of New South Wales, Sydney, Australia
[3] Prince of Wales Hospital, Sydney, Australia
[4] Royal Prince Alfred Hospital, Sydney, Australia

Presenter's email: dgoodman@kirby.unsw.edu.au

Introduction:
People who inject drugs (PWID) are at risk for severe infections, but International Classifications of Diseases (ICD) coding fails to capture this group accurately. Advances in natural language processing (NLP), particularly large language models (LLMs), offer a scalable alternative by extracting relevant information from clinical text. This study assessed the performance of off-the-shelf LLMs in identifying PWID and related attributes from discharge summaries at an Australian tertiary hospital.

Methods:
This cross-sectional study included patient's first admissions to the Infectious Diseases service at St Vincent's Hospital Sydney between January 2018 and December 2022. Discharge summaries were manually annotated for PWID status, substances used, injection recency (current defined injection within 30 days) and treatment. Six open-source LLMs (*Gemma3, Llama 3.2, Mistral, Phi4, Hippomistral, Llama3-med*) were evaluated without fine-tuning. Diagnostic performance was assessed using weighted macro-F1 scores—which balance sensitivity and positive predictive value across labels—and individual metrics (e.g., sensitivity, specificity, F1) with bootstrapped 95% confidence intervals (CI).

Results:
Of 857 individual admissions, manual review identified 147 (17.1%) as PWID, but only two (0.2%) were classified using relevant ICD codes. The best-performing model (*Phi4*) achieved a macro-F1 score of 0.78 (95% CI: 0.61–0.87). Its performance in identifying PWID was high (F1 = 0.96, 95% CI: 0.93–0.98), with lower accuracy for current PWID (F1 = 0.79, 95% CI: 0.78–0.85) and historical PWID (F1 = 0.52, 95% CI: 0.38–0.65). The model showed strong diagnostic performance for identifying heroin (F1=0.86, 95% CI: 0.77–0.94) and methamphetamine use (F1=0.97, 95% CI: 0.93–0.99), while performance was low for prescription opioids (F1=0.25, 95% CI: 0.13–0.36) and benzodiazepines (F1=0.51, 95% CI: 0.37–0.65).

Conclusion:
LLMs accurately identified PWID from clinical text though extraction of detailed attributes remains challenging.

Implications:
LLM-based NLP tools could improve monitoring of drug-related harms during hospitalizations and guide data-informed health policy, if rigorously validated.

Disclosure of Interest Statement: *No pharmaceutical grants or direct funding were received in the development of this study. The Kirby Institute is funded by the Australian Department of Health. The views expressed in this publication do not necessarily represent the position of the Australian Government. MM has received an NHMRC Investigator Grant (2034418). GVM has received an NHMRC Investigator Grant (2016698). GJD has received an NHMRC Investigator Grant (2008276). DGM has received research grants from Gilead. MM has received compensation for travel and consultancy from Abbvie. JM has received an honorarium for a lecture from ViiV. G. J. D. is a consultant/advisor and has received research grants from AbbVie, Abbot Diagnostics, Gilead Sciences, Bristol Myers Squibb, Cepheid, GlaxoSmithKline, Merck, Janssen, and Roche. G. V. M. has received research*

grants from ViiV, Gilead, and Janssenn. All other authors report no potential conflicts. GM has received research grants from ViiV, Gilead, and Jansenn. GD has received research grants from Gilead and Abbvie.